Impact of Advances in Computing and Communications Technologies on Chemical Science and Technology: Report of a Workshop

Chemical Sciences Roundtable, National Research Council

ISBN: 0-309-59707-2, 236 pages, 8.5 x 11, (1999)

Impact of Advances in Computing and Communications Technologies on Chemical Science and Technology

Report of a Workshop

Chemical Sciences Roundtable
Board on Chemical Sciences and Technology
Commission on Physical Sciences, Mathematics, and Applications
National Research Council

NATIONAL ACADEMY PRESS
Washington, D.C.

NOTICE: The project that is the subject of this report was approved by the Governing Board of the National Research Council, whose members are drawn from the councils of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine. The members of the workshop organizing committee responsible for the report were chosen for their special competences and with regard for appropriate balance.

The National Academy of Sciences is a private, nonprofit, self-perpetuating society of distinguished scholars engaged in scientific and engineering research, dedicated to the furtherance of science and technology and to their use for the general welfare. Upon the authority of the charter granted to it by the Congress in 1863, the Academy has a mandate that requires it to advise the federal government on scientific and technical matters. Dr. Bruce Alberts is president of the National Academy of Sciences.

The National Academy of Engineering was established in 1964, under the charter of the National Academy of Sciences, as a parallel organization of outstanding engineers. It is autonomous in its administration and in the selection of its members, sharing with the National Academy of Sciences the responsibility for advising the federal government. The National Academy of Engineering also sponsors engineering programs aimed at meeting national needs, encourages education and research, and recognizes the superior achievements of engineers. Dr. William A. Wulf is president of the National Academy of Engineering.

The Institute of Medicine was established in 1970 by the National Academy of Sciences to secure the services of eminent members of appropriate professions in the examination of policy matters pertaining to the health of the public. The Institute acts under the responsibility given to the National Academy of Sciences by its congressional charter to be an adviser to the federal government and, upon its own initiative, to identify issues of medical care, research, and education. Dr. Kenneth I. Shine is president of the Institute of Medicine.

The National Research Council was organized by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of furthering knowledge and advising the federal government. Functioning in accordance with general policies determined by the Academy, the Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in providing services to the government, the public, and the scientific and engineering communities. The Council is administered jointly by both Academies and the Institute of Medicine. Dr. Bruce Alberts and Dr. William A. Wulf are chairman and vice chairman, respectively, of the National Research Council.

Support for this project was provided by the National Science Foundation under Grant No. CHE-9630106, the National Institutes of Health under Contract No. N01-OD-4-2139, and the U.S. Department of Energy under Grant No. DE-FG02-95ER14556. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation, the National Institutes of Health, or the U.S. Department of Energy.

International Standard Book Number: 0-309-06577-1
Additional copies of this report are available from: National Academy Press 2101 Constitution Avenue, NW Box 285 Washington, DC 20055 800-624-6242 202-334-3313 (in the Washington metropolitan area) http://www.nap.edu

Printed in the United States of America

CHEMICAL SCIENCES ROUNDTABLE

Staff

BOARD ON CHEMICAL SCIENCES AND TECHNOLOGY

Staff

COMMISSION ON PHYSICAL SCIENCES, MATHEMATICS, AND APPLICATIONS

Preface

The Chemical Sciences Roundtable (CSR) was established in 1997 by the National Research Council (NRC). It provides a science-oriented, apolitical forum for leaders in the chemical sciences to discuss chemically related issues affecting government, industry, and universities. Organized by the NRC's Board on Chemical Sciences and Technology, the CSR aims to strengthen the chemical sciences by fostering communication among the people and organizations—spanning industry, government, universities, and professional associations—involved with the chemical enterprise. The CSR does this primarily by organizing workshops that address issues in chemical science and technology that require national attention.

At its second meeting in December 1997, the CSR identified the topic of information technology as an issue of increasing importance to all sectors of the chemical enterprise. As we rely increasingly on computers for obtaining, recording, communicating, and publishing the scientific data that enables progress in our discipline, it is correspondingly important that we consider the new and developing ways that this essential technology can be used effectively. At the same time, we must also consider the impact of the evolving technology on all sectors of our discipline and on the ways that these sectors interact.

To provide a forum for exploring this topic, an organizing committee was formed, and a workshop was planned for November 1998. The workshop, "The Impact of Advances in Computing and Communications Technologies on Chemical Science and Technology," brought together research scientists and managers from government, industry, and academia to review and discuss the rapid changes in computer technology that are influencing activities in the chemical sciences.

The papers in this volume are the authors' own versions of their presentations, and the discussion comments were taken from a transcript of the workshop. The workshop did not attempt to establish any conclusions or recommendations about needs and future directions, focusing instead on individual problems and challenges identified by the speakers. By providing an opportunity for leaders in each of the areas to share their experience and vision, we intended that the other workshop participants—as well as readers of this proceedings volume—would be able to identify new and useful ways of using the

tremendous power of computing and information technology in their own endeavors. We believe that the workshop was successful in meeting this goal.

WORKSHOP ORGANIZING COMMITEE

THOM H. DUNNING, JR., CHAIR

ALLEN J. BARD

THOMAS F. EDGAR

JEAN H. FUTRELL

RICHARD M. GROSS

BEVERLY K. HARTLINE

ROBERT L. LICHTER

THOMAS A. MANUEL

ROBERT S. MARIANELLI

JANET G. OSTERYOUNG

MICHAEL E. ROGERS

Acknowledgment of Reviewers

This report has been reviewed by individuals chosen for their diverse perspectives and technical expertise, in accordance with procedures approved by the National Research Council's (NRC's) Report Review Committee. The purpose of this independent review is to provide candid and critical comments that will assist the authors and the NRC in making the published report as sound as possible and to ensure that the report meets institutional standards for objectivity, evidence, and responsiveness to the study charge. The contents of the review comments and draft manuscript remain confidential to protect the integrity of the deliberative process. We wish to thank the following individuals for their participation in the review of this report:

Judith Hempel, University of California, San Francisco,
Margaret G. Kivelson, University of California, Los Angeles,
David McLaughlin, Kodak Research Laboratories,
L. Eugene McNeese, Oak Ridge National Laboratory, and
Stanley I. Sandler, University of Delaware.

Although the individuals listed above have provided many constructive comments and suggestions, responsibility for the final content of this report rests solely with the authoring group and the NRC.

Contents

Impact of Advances in Computing and Communications Technologies on Chemical Science and Technology

Summary

The second workshop of the Chemical Sciences Roundtable (CSR), "Impact of Advances in Computing and Communications Technologies on Chemical Science and Technology," was held in Washington, D.C., on November 1-2, 1998. The presentations and discussion at the workshop considered benefits and opportunities for chemical science and technology stemming from ongoing dramatic advances in computing and communications, as well as challenges to be met in using these technologies effectively for research and in applications addressing pressing national problems. This volume presents the results of that workshop.

WHITHER COMPUTING AND COMMUNICATIONS TECHNOLOGIES IN CHEMICAL SCIENCE AND TECHNOLOGY?

Paul Messina of the California Institute of Technology emphasized the importance of ensuring the continuing evolution of computing power, unprecedented levels of which are required to meet such daunting needs as ensuring the safety and reliability of the nation's nuclear arsenal. In particular, he described the Department of Energy's Accelerated Strategic Computing Initiative (ASCI), whose goal is to simulate the effects of aging on the U.S. nuclear stockpile and to assess new weapon designs without further underground nuclear testing. To achieve a computational simulation-based approach to testing will require computer systems capable of trillions to quadrillions of arithmetic operations per second. ASCI is working closely with the computing industry to ensure that this high-end computing capability is fielded in the next 5 to 10 years.

In addition to investing in accelerated development of a new generation of massively parallel computer systems, ASCI is making major investments in computer systems software and scientific simulation software. Rapid progress in both computing and simulation capability is required to make these systems usable for addressing the targeted problems. This massive undertaking—which involves the academic community and the U.S. computer industry in addition to applications scientists and engineers at Los Alamos, Lawrence Livermore, and Sandia national laboratories—is attempting to build

a balanced system in which computing speed and memory, archival storage and capacity, and network speed and throughput are combined to dramatically increase the performance simulations. The approach of using commercially available components to the extent possible will facilitate transfer of the new technologies for use in a number of scientific and engineering pursuits without duplicating ASCI's costs.

Messina asserted that the advances in computing power that will become available in the next 5 to 10 years will be so great that they will change the very manner in which we pursue advances in science and technology.

Peter R. Taylor, San Diego Supercomputer Center and University of California, San Diego, spoke on the state of the art in computational chemistry and the extent to which it moots the requirements of the chemical science community. Computational chemistry—whose major activities can be classified as molecular electronic structure (often referred to as quantum chemistry), reaction and molecular dynamics, and statistical mechanics—is one of the great scientific success stories of the past three decades, as evidenced by the award of the 1998 Nobel Prize in chemistry to John Pople and Walter Kohn. Taylor described computational chemistry as a mature and very successful field that nevertheless requires continuing effort to improve theories, methods, algorithms, and implementation. He also pointed to a need for training students in these areas.

More powerful computers will allow current methods to be extended to over larger molecules, but new methodologies will be needed to address many of the problems of interest to chemists. Taylor stated that the chemical sciences community needs to encourage the implementation of existing methods on now hardware, as well as the development and implementation of new methods. As new methods are developed, possible advantages offered by new computer architectures can be considered; e.g., approaches previously precluded because of requirements for enormous memory might be perfectly feasible on ASCI-class machines. Use of modern software engineering practices and modern computer languages in implementations can increase ease of maintenance. New methods and implementations can also take advantage of modern storage, retrieval, and data management technologies as well as interactive environments in which users can steer simulations and visualize their data.

Susan L. Graham, University of California, Berkeley, started by noting that high-performance computing is difficult. She elaborated on the technical issues that must be addressed if we are to take advantage of the exciting opportunities offered by the ongoing revolutionary increases in computing power. She indicated that one way to get more out of computing is by using parallelism—it reduces the elapsed time required for the most demanding computations, keeps the calculation moving along when delays arise in sequential computation, and overcomes fundamental limitations bounding the speed of sequential computation, such as the speed of light. However, advances from parallelism won't come for free. Issues that must be addressed in improving end-to-end performance of a calculation include identifying the work that can be done in parallel, correctly partitioning that work across the processors, and arranging the data so that it resides close to where it is needed (because of communication delays). Even then, at a lower level in the system, the system software (or a programmer) has to describe the details of how the work is actually done. Graham also mentioned issues in addition to performance that are going to become increasingly problematic, such as security and fault tolerance.

Among the nontechnical issues mentioned were concerns about having enough people with the deep knowledge of both chemistry and information technology required for developing workable problem-solving strategies. In addition, Graham pointed out that the scientific community will have to become

involved in developing software for which vendors do not see a large market and in dealing with issues of access to very high performance systems.

Graham closed by mentioning how implementation of the recommendations from the President's Information Technology Advisory Committee (PITAC), on which she serves, could help address such important issues as the need for investment in long-term information technology R&D.

COMPUTATIONAL MODELING AND SIMULATION IN CHEMICAL SCIENCE AND TECHNOLOGY

John A. Pople of Northwestern University discussed the importance of having reliable data on the thermochemistry of molecules—knowledge of which is vital in the chemical sciences and essential to many technologies. Because experimental measurements yielding thermochemical data are difficult and time-consuming, it is highly desirable to have computational methods that can make reliable predictions. Since the early 1970s when ab initio molecular orbital calculations became routine, one of the major goals of modern quantum chemistry has been the prediction of molecular thermochemical data to chemical accuracy (1 kcal/mol). The Gaussian-n series, with its latest version, Gaussian-3 (G3) theory, achieves that accuracy on average and is computationally feasible for molecules containing up to about eight non-hydrogen atoms; Pople asserted that it represents a great success of quantum chemistry.

Ideally, a method for computation of thermochemical data should be applicable to any molecular system in an unambiguous manner. The method needs to be computationally efficient so that it can be widely applied, should reproduce known experimental data to a prescribed accuracy, and should be similarly accurate when applied to species having larger uncertainty or for which data are not available. The Gaussian-n methods were developed with these objectives in mind. Despite the successes, Pople argued that much remains to be done. Among the challenges will be extension of the methods to larger molecules, increased accuracy in predictions, and extension to heavier elements. The increased computing power obtainable from new generations of computers, such as those with massively parallel architectures, will play an important role in meeting these challenges.

Jeffrey Skolnick of the Scripps Research Institute discussed the role of computational molecular biology in the genomics revolution. Various genome-sequencing projects are providing a plethora of protein sequence information, but with no information about protein structure or function. Making the results of the genome revolution applicable to understanding biological processes requires knowledge of protein structure and function as encoded in the genome. One means of sifting useful proteins out of the genomic databases is the computer prediction of protein function. To extend the level of molecular function annotation to a broader class of protein sequences, a novel method for identification of protein function based directly on the sequence-to-structure-to-function paradigm has been developed. The idea is to predict the native structure first and then to identify the molecular or biochemical function by matching the active site in the predicted structure to that in a protein of known function. Skolnick believes that the next 5 to 10 years are likely to see the development of improved computational tools for genomic screening.

W. David Smith, Jr. of E.I. DuPont described needs and new directions in computing for the chemical process industries. Changing needs for process and enterprise modeling and the capability of computers and software have reached a critical point where vendors, academia, and industry must cooperate to develop the next generation of tools for the process engineer. The potential for the European CAPE OPEN project to bring the technologies and the players together to provide this set of

tools was discussed. Smith also explored how this technology may change the ways in which companies like DuPont perform engineering and process development in the future.

It is not enough to develop computational methods for modeling chemical processes; one must also facilitate the use of these techniques by scientists. This was the underlying theme of the presentation by **Gregory J. McRae**[1] of the Massachusetts Institute of Technology, who provided an overview of a problem-solving environment called the Chemical Engineering Workbench. The Workbench is currently being developed by a research team associated with the National Computational Science Alliance. When completed, it will provide an integrated software environment supporting a broad range of computational tools for modeling chemical and engineering processes, extending from the molecular level to that of full chemical plants. Quantum chemistry and other tools at the molecular level are being coupled with higher-level chemical process modeling, chemical process reaction modeling, plant process design, and process control. The team's initial effort is to develop an advanced reactor design model that can be incorporated into the Workbench. Reactors are the focal points of chemical plants, and mathematically modeling these complex systems places great demands on high-performance computing. Design considerations from the plant level will feed down to the quantum level. Treating chemistry as a design parameter may allow development of innovative reaction systems that minimize environmental problems.

Thomas F. Edgar, University of Texas, **David A. Dixon,** Pacific Northwest National Laboratory, and **Gintaris V. Reklaitis,** Purdue University, gave a multi-author perspective on the computational needs of the chemical industry. The current forces driving the U.S. chemical industry, such as globalization, requirements for minimization of environmental impact, and the need for improved return on investment, require the expanded use and application of new computational technologies. Forecasted future improvements in process modeling, control, instrumentation, and operations are a major component in the recently completed report *Technology Vision 2020: Report of the U.S. Chemical Industry*,[2] which presents a road map for the next 25 years for the chemical and allied industries. Detailed R&D road maps on specific areas of chemical technology summarized in this presentation were prepared in 1997 and 1998. The areas covered were instrumentation, control, operations, and computational chemistry.

REMOTE COLLABORATION AND INSTRUMENTS ONLINE

Raymond A. Bair, Pacific Northwest National Laboratory, emphasized that new technologies for computing and communications offer opportunities to revolutionize not only the scope but also the process of scientific investigation. Bair predicted that a significant contribution of collaboratories will be support for creating and sustaining scientific communities that can interact and share information rapidly to address challenging research problems. Moreover, as the chemical applications and capabilities provided by collaboratories become more familiar, researchers will move significantly beyond current practice to exciting new paradigms for scientific work.

[1] A written contribution for this presentation was not available for inclusion in the workshop proceedings.

[2] American Chemical Society, American Institute of Chemical Engineers, Chemical Manufacturers Association, Council for Chemical Research, and Synthetic Organic Chemical Manufacturers Association. 1996. *Technology Vision 2020: Report of the U.S. Chemical Industry*. Washington, D.C.: American Chemical Society.

Bair detailed some of the requirements for future success, including development of interdisciplinary partnerships of chemists and computer scientists; flexible and extensible frameworks for collaboratories; means to deploy, support, and evaluate collaboratories in the field; input from the nation's research community to development of the next generation of Interact standards; and higher-performance networks, more scalable and capable network standards, and better network management capabilities.

In concluding, he pointed out the opportunity for competitive advantage provided by collaboratories that give access to expertise, data, experiments, or computations that would not otherwise be available to explore a research question or solve a problem. He also noted the positive impact that collaboratories can have on the complexity and scale of chemical problems considered, as well as the crucial part collaboratories can play in projects that depend on having access to large user facilities and/or multidisciplinary research teams. By removing many barriers of time and distance, collaboratories can enhance the exchange of information in the sciences and can also contribute to the management of costs for travel and equipment use.

Bridget Carragher and **Clinton S. Potter,** Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, discussed their experience with the development of remote and automated access to imaging instrumentation in the World Wide Laboratory (WWL) project. They proposed several ways of using remote-access technology in practice, including for service, collaboration, education and training, remote research, and automated and intelligent control of functions usually performed manually by a local operator. Among the advantages they described for remote-access technology—which is one component of a collaboratory—were opportunities for consultation with experts located anywhere, access to a network of distributed expertise, and unprecedented opportunities for education, training, and access for users at institutions lacking the means to support expensive and unique instruments.

Specific examples they reported on involved remote work with a transmission electron microscope, nuclear magnetic resonance imaging spectrometers, and a video light microscope—all of which are accessible in the WWL through Web-browser-based user interfaces. One K-12 education project, Chickscope, showed that very complex remote-access technology could be used effectively by students at all grade levels and also demonstrated all of the components defined for a working collaboratory.

They concluded by noting that wider acceptance of collaboratories in the general scientific community would require demonstration of their impact in the scientific research environment as well as a systematic evaluation of their contribution to productivity.

Thomas A. Finholt, University of Michigan, started by noting that despite the tremendous growth in the knowledge and practical application of chemical principles, the practice of chemistry research and teaching has remained relatively unchanged. The use of the Internet as a worldwide mechanism for scientific communication challenges this status quo. Innovations such as collaboratories that remove constraints of distance and time on scientific collaboration and increase access to scarce instruments will accelerate the flow of information and place new demands on senior scientists as mentors. Finholt pointed out the need to anticipate and influence the development of emerging Internet technologies that will affect how research is done. He discussed the challenges posed by new ways of conducting research in chemistry in terms of our transition from the past, through the present, and into an uncertain future.

In reviewing "where we come from," Finholt discussed three particularly important innovations: the creation and elaboration of the research laboratory, the use of laboratory classes in chemical education, and the use of lecture demonstrations to illuminate and clarify chemical principles. Briefly describing "what we are," he considered opportunities that are now available to chemists through the expansion of

computer and Internet technologies and that can be thought of in terms of the raw performance of computer processors, the capacity of communication networks, the scope of networks, and the evolution of software. Finally, in exploring "where we are going," he used examples from the present to project possible new modes of teaching and doing research, including the collaboratory, that would enable researchers to perform their work without regard to geographical location—interacting with colleagues, using instruments remotely, sharing data and computational resources, and accessing information in digital libraries. He concluded by cautioning that while Web-based tools may alter the landscape of practice and pedagogy, simply "surfing" for information will not replace learning. To master and understand key concepts, students and researchers must absorb and reflect on ideas and avoid the temptation to browse endlessly among an ever-widening array of online resources.

David R. McLaughlin, Eastman Kodak Company, described how Kodak has used computers and information technology to enhance operations in its research laboratories. The effort has focused on creating an electronic or computerized laboratory and delivering information to the scientist's desktop. To illustrate the impact of Kodak's "wired laboratory," he described four ways that advances in computing technology have helped to increase the efficiency of the analytical chemistry laboratory: the first through automation and simplification of some of the tasks associated with analysis and synthesis, the second in management of information and knowledge, the third in the generation and maintenance of data in electronic (digital) form, and the fourth through data analysis and chemometrics. Examples given of components of the wired laboratory included QUANTUM, an integrated spectroscopy information system; a walk-up spectroscopy laboratory with instruments online; increased capabilities for electronic access to information and analytical data; and WIMS, a Web-based information management system. He also provided information on an electronic laboratory notebook that has been developed to assist with the management of experiments, projects, and programs.

McLaughlin characterized the wired laboratory of the future as one in which all scientists would use an intelligent electronic laboratory notebook linked to all of the data-generating equipment, and in which evolving analytical technology in combination with data analysis techniques would reduce the time required for sample preparation and data interpretation. All of these capabilities would be provided through a common Web interface. He indicated that a challenge to the analytical community is to devise real-time measurements that, when displayed in virtual reality systems, will enable researchers to "see" results and thus better understand them. He also noted that high-quality, reliable software available at reasonable cost is one of the most critical needs for the future.

CHEMICAL INFORMATION ONLINE

Gary Mallard, National Institute of Standards and Technology, discussed a variety of issues in the management of chemical data. He suggested that the large growth in publication of information on the Internet is being driven by a reduction in traditional data resources, demand for faster access to data, and increased needs for data for modeling and simulation. The last factor has a particularly strong influence because it has also changed the nature of the data needed. He stressed the importance of quality assurance in data generation and management, including basic data storage and archiving. So that these data will be useful and reliable, critical evaluation must aim to detect errors and inconsistencies in the information that can originate from incomplete data sets, uncertainty in the data, and errors introduced during the compilation process. Mallard stressed the importance of preventing and correcting errors as the scientific community's demands for data continue to increase.

Richard E. Lucier, University of California, described work directed at developing a "digital library" to serve the entire university system in California. He argued that the conventional library—with its archives of traditional research journals—is evolving toward new forms of scholarly communication. In his view traditional libraries, with the services we have come to expect from them, will not continue to be sustainable. Given the information explosion, they simply cost too much. Costs will be controlled by forming large consortia of libraries to leverage buying power. He proposed that comprehensive access to information will replace comprehensive ownership of information, and that world-class libraries will consist of complementary paper and digital holdings. In addressing possible new approaches to scholarly communication, he pointed out the conflicting goals of the current system, which uses publications both as a means to disseminate knowledge and as a mechanism for evaluating the performance of research scholars. New approaches to electronic publication may allow these to be decoupled. Pursuing new approaches will require examination of current copyright policies and practices such as the assignment of rights to publishers.

Lorrin R. Garson, American Chemical Society, presented an analysis of issues related to the emergence of electronic publishing, using the perspective of a scientific society that is already a major publisher in the print medium. He characterized scientific publishing as a field with high costs, diminishing resources for those purchasing publications, competition among publishers, and increasing pressure to publish more material. He noted that both commercial and not-for-profit publishers must strive to operate on a "not-for-loss" basis. He presented arguments that "first-copy" costs account for approximately 80 percent of all publishing costs, regardless of whether paper or electronic distribution is the final result. Consequently, the financial challenges are unlikely to disappear with a move toward electronic publishing.

Garson identified a range of important problems and challenges associated with electronic publishing, including the need for improvements in technology and funding of that investment, assumption of responsibility and costs for archiving of electronic information, terms for and constraints on use of electronic information, and costs of individual subscriptions. He was optimistic about progress in overcoming the technical barriers but indicated that the financial and sociological obstacles are formidable.

1

The Accelerated Strategic Computing Initiative

Paul Messina
California Institute of Technology and Department of Energy

While the increase in computing power over the past 50 years has been staggering, the scientific community will require unprecedented computer speeds as well as massive memory and disk storage to address the pressing problems that the nation will face in the 21st century. One such problem is ensuring the safety and reliability of the nation's nuclear arsenal while fully adhering to the Comprehensive Test Ban Treaty. To address this problem, the U.S. Department of Energy (DOE) established the Accelerated Strategic Computing Initiative (ASCI) in 1996.[1] The goal of ASCI is to simulate the results of new weapons designs as well as the effects of aging on existing and new designs, all in the absence of additional data from underground nuclear tests. This is a daunting challenge and requires simulation capabilities that far surpass those available today.

The goal of ASCI, however, is not a pipe dream. With funding from ASCI, the computer industry has already installed three computer systems, one at Sandia National Laboratories (built by Intel), one at Los Alamos National Laboratory (LANL) (an SGI-Cray computer), and another at Lawrence Livermore National Laboratory (LLNL) (an IBM computer), that can sustain more than 1 teraflops on *real* applications. At the time they were installed, each of these computers was as much as 20 times more powerful than those at the National Science Foundation (NSF) Supercomputer Centers (the Partnerships for Advanced Computational Infrastructure), the National Energy Research Supercomputing Center, and other laboratories. And this is only the beginning. By 2002, the computer industry will deliver a system 10 times more powerful than these two systems and, in between, another computer will be delivered that has three times the power of the LANL/LLNL computers. By the year 2004—only 5 years from now—computers capable of 100 trillion operations per second will be available.

Who needs this much computing power? ASCI clearly does, but the other presentations at this workshop indicate that many chemical applications could also make use of this capability. Similar

[1] Based on the new capabilities being developed by ASCI, the Department of Energy, in its FY2000 budget submission, proposed to extend this concept to its civilian research programs. It requested $70 million for the Scientific Simulation Initiative, DOE's contribution to the President's Initiative on Information Technology for the Twenty-first Century.

computational needs can be put forward by climate and weather modelers, computational materials scientists, and biologists, for example. In the summer of 1998, NSF and DOE sponsored the "National Workshop on Advanced Scientific Computing," which described the opportunities that this level of computing power would create for the research programs funded by NSF and DOE.[2] The good news is that, as a result of the enormous investment that ASCI is making in these machines, it is likely that computers of this power will become available to the general scientific community, and at substantially reduced cost.

ASCI'S NEED FOR ADVANCED SIMULATION CAPABILITY

The ASCI program is a direct result of President Clinton's vision, a vision shared by Congress as well, that the United States can ensure the safety and reliability of its nuclear stockpile without additional nuclear testing. DOE's ASCI program has been designed to create the leading-edge computational modeling and simulation capabilities that are needed to shift from a nuclear test-based approach to a computational simulation-based approach. There is some urgency to putting this simulation capability in place as the nuclear arsenal is getting older day by day—the last nuclear test was carried out in 1992, so by the year 2004, 12 years will have passed since the last test. In addition, nuclear weapons are designed for a given lifetime. Over 50 percent of the weapons in the U.S. arsenal will be beyond their design lifetime by the year 2004, and there is very little experience in aging beyond the expected design life of nuclear weapons. Finally, the individuals who have expertise in nuclear testing are getting older, and, by the year 2004, about 50 percent of the personnel with first-hand test experience will have left the laboratories. The year 2004 is a watershed year for DOE's defense programs.

The Challenges

The Accelerated Strategic Computing Initiative is an applications-driven effort with a goal to develop reliable computational models of the physical and chemical processes involved in the design, manufacture, and degradation of nuclear weapons. Based on detailed discussions with scientists and engineers with expertise in weapons design, manufacturing, and aging and in computational physics and chemistry, a goal of simulating full-system, three-dimensional nuclear burn and safety simulation processes by the year 2004 was established. A number of intermediate, applications-based milestones were identified to mark the progress from our current simulation capabilities to full-system simulation capabilities. Before developing the three-dimensional burn code, codes must be developed to simulate casting, microstructures, aging of materials, crash fire safety, forging, welding of microstructures, and so on.

We cannot meet the above simulation needs unless computing capability progresses along with the simulation capability. To this end, the first ASCI computing system, an Intel system ("Option Red"), was installed at Sandia National Laboratories in Albuquerque in 1997. Sandia and the University of New Mexico wrote the operating system for this machine and, by 1996, it had achieved more than 1 trillion arithmetic operations per second (teraflops) while still at the factory. It also had over one-half terabyte of memory, the largest of any computing system to date. Next, "Option Blue" resulted in the acquisition of two machines: an IBM system at LLNL ("Blue Pacific") and an SGI/Cray system at LANL ("Blue Mountain"). The IBM system achieved its milestone of over 1 teraflops on an application

[2] Department of Energy and National Science Foundation. Report from the "National Workshop on Advanced Scientific Computing" held July 30-31, 1998, in Washington, D.C., J.S. Langer, ed.

on September 22, 1998, an event that was announced by Vice President Gore just one week before this workshop. The "Blue Mountain" machine was delivered in October 1998 and became operational in mid-November.

This is the current status of the ASCI computing systems, but the story does not stop here. As follow-on to the Livermore acquisition, the same IBM contract that led to the 3-teraflops system will be used to procure a 10-teraflops/5-terabyte (memory) system in mid-2000. A 30+ teraflops machine is scheduled for delivery in June of the year 2001, and a 100-teraflops machine is to be delivered in 2004. The above computer schedule is tied directly to the scientific and engineering application needs through detailed estimates of the computing power, memory, and disk storage that the applications will need at any given time.

To follow the pace outlined above, it is necessary to substantially accelerate what the U.S. computer industry would otherwise do. This requires a partnership between ASCI, the applications scientists and engineers, and the U.S. computer industry. In order to produce a computer capable of 100 teraflops by the year 2004, ASCI settled on the strategy of using the products that industry was already concentrating on—relatively small, commodity-priced building blocks—and assembling those components into big systems with the needed computing power. This process seems more efficient because smaller building blocks present fewer problems from an innovation standpoint, there is a much larger commercial market for them (which results in cheaper unit costs), and other applications would be able to benefit from the knowledge created in combining smaller elements. Designing and building a single computer from the ground up might benefit the ASCI applications, but would not serve the rest of the scientific community except by replication of the entire system.

The disadvantage of building a computer system from many smaller units is that the interconnections between these units can become overwhelmingly complicated. Scientists nowadays use 32 or 64 processors, but a different method is needed to connect 6,400 processors. Not only does the network have to have different characteristics, but the software does also. The software must have fast (low-latency) access to memory, even on systems such as these, which have a very complicated memory structure. This is necessary not only for large-scale simulations, but also for the transfer of data and the visualization of results. There is another problem associated with the use of computer systems of this power and complexity: the burden on the user. It will be a challenge to keep the overhead on the user at a reasonable level, especially if the user is at a remote site.

Where Are We Now?

There has been an Intel-designed and Intel-built ASCI computer at Sandia National Laboratories in Albuquerque for 2 years. It has 4,500 nodes, 9,000 processors (two per node), and a peak performance of 1.8 teraflops. During that 2-year period, it was the fastest computer on the planet for scientific simulation. It has a fair amount of memory—half a terabyte—and a very high speed network interconnecting the 4,500 nodes. It is physically a very big system, built from the highest-volume manufactured building blocks available—the Intel Pentium Pro microprocessor. This machine has been used for a number of breakthrough simulations and has been invaluable for the scientific and engineering application teams in their efforts to develop simulation software that scales to large numbers of processors.

The nodes on the two so-called Blue systems, Blue Mountain at Los Alamos (SGI/Cray) and Blue Pacific at Livermore (IBM), are symmetric multi-processors (SMPs), which are interconnected by a scalable, high-speed communications network. The IBM terascale computer system, which was delivered to LLNL in January 1999, is a three-machine aggregate system where each machine comprises 512 four-processor nodes, 488 of which are used for computing. The remaining nodes connect to a router

that links the three separate machines. All together, the Blue Pacific system contains 5,856 processors and has a theoretical peak performance of 3.9 teraflops. The SGI/Cray computer at Los Alamos has only 48 nodes, but each node has 128 processors (SGI Origin 2000s). The peak performance of the SGI/Cray Blue Mountain system is around 3.1 teraflops. These two computer systems, although they use the same fundamental architecture, clearly represent two extremes of that architecture. There is n trade-off between a more complex interconnection of fairly simple nodes in the one case and a simpler connection of more complex building blocks, SMPs with 128 processors sharing memory and providing cache coherence, in the other case. Although the two architectures are the same, tuning a scientific application to be equally efficient on both computers is difficult because the interconnect network is different, as is the amount of shared memory for each processor.

TABLE 1.1 Characteristics of ASCI Computer Systems

	Machine				
Characteristics	Intel (Sandia)	Blue Pacific (Livermore)	Blue Mountain (Los Alamos)	Projected 10 Teraflops	Projected 30 Teraflops
Performance (peak teraflops)	1.8	3.9	3	10	30
Dates operational	May 1997	September 1998	November 1998	June 2000	June 2001
Processor	Pentium Pro	Power PC	MIPS R10000	Power PC	?
Number of nodes	4,568	1,464	48	512	?
Processors per node	2	4	128	16	?
Number of processors	9,000	5,856	6,144	8,192	>8,000 (?)

The architecture of the ASCI Blue Pacific and Blue Mountain computers—interconnected SMPs—will be the dominant architecture over the next few years. Moore's law will lead to faster processors,[3] of course, and there will be different sizes of shared memory processor building blocks and different flavors of the interconnect, but this will be the fundamental architecture by which we will attain the 30-and 100-teraflops performance levels. It is just not feasible to create a super-fast processor in the next 5 years that could attain such performance levels with only 100 or so processors. So, we have to deal with the level of complexity that SMP-based architectures imply. The specifications for these computer systems are summarized in Table 1.1. Figure 1.1 shows the ASCI platforms road map.

It is worth noting, however, that these levels of complexity are not overly different from what we have been dealing with in some laboratories and universities for over 15 years. In 1986 there were systems containing 512 processors (as separate nodes) that presented many of the challenges that ASCI is facing. Soon after those efforts, a few thousand processors were successfully connected. However, this does not diminish the challenge that ASCI is undertaking. It is one thing to get a few computational kernels running on a parallel computer and another thing to get the massive codes used to simulate physical and chemical systems running efficiently on the complex ASCI machines.

As an example of the applications now possible on the ASCI machines, a three-dimensional calculation used the ARES code to simulate Rayleigh-Taylor turbulent mixing and hydrodynamic instabilities found in supernovae and contained over 35 million zones and used 1,920 processors on the

[3] Moore's law, actually a "rule of thumb," states that the computing power of microprocessors doubles every 18 months or so.

Blue Pacific (Figure 1.2). The ARDRA code was used to simulate the neutron flux emitted by a Nova laser experiment in a calculation using 160 million zones in 33 hours using 3,840 processors. Efficient and accurate neutron and photon transport calculations have been conducted on the Blue Mountain machine, demonstrating a l-ram grid resolution using 100 million particles. On a single-processor machine like those typical of most research environments, these calculations would take 10 to 100 years to run. Perhaps more important, some ASCI simulations have already led to insights about what caused some previously not-understood historical nuclear test results.

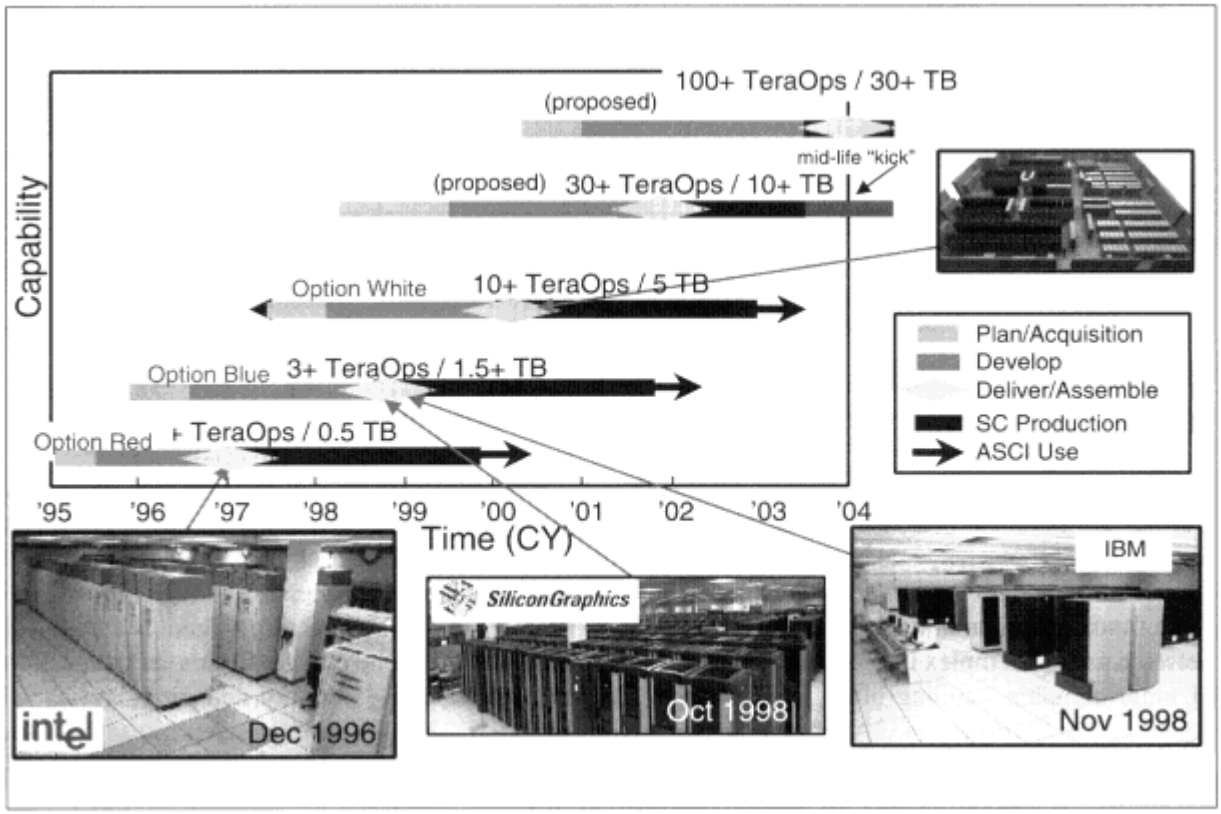
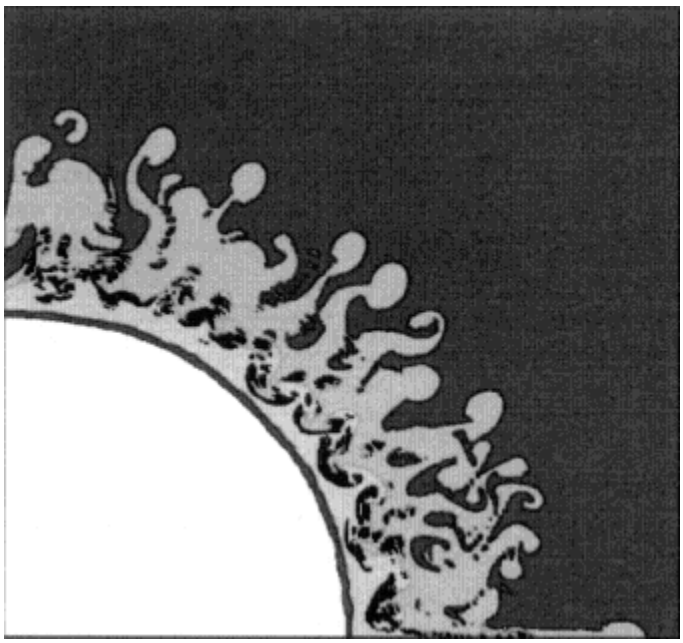


Figure 1.1
ASCI computing systems.

Accelerating Technology Developments

ASCI has an effort it calls "PathForward" whose goal is to enable U.S. companies to develop technologies needed to produce the next-generation ultra-scale computing systems for ASCI. PathForward draws on the capabilities, availability, expertise, and products currently being produced by leading computer companies, focusing on interconnect technologies, data storage technologies, systems software, and tools for large-scale computing systems. These technologies, while critical to ASCI's platform needs, are areas in which private sector development would not otherwise take place, at least not in the time frame required by the stockpile stewardship program. At the same time, they are

investments in which industry sees value for future products and markets: essential scaling and integrating technologies that enable ultra-scale computing systems to be engineered and developed out of commodity computing building blocks.



Turbulent mixing, as it occurs in supernovae, represents the combined effects of nuclear reactions and hydrodynamic mixing that are relevant to stockpile stewardship
The combination of an efficient hydrodynamic algorithm and the performance capability of the Blue Pacific makes such a highly resolved three-dimensional simulation possible

Figure 1.2
AREAS supernova simulation. This calculation included a modest 1098 × 180 × 180 spatial mesh with more than 35 million zones. It was completed over a 72-hour period on 1,920 processors.

Simulation Development Environment

As we all know, software and algorithms are at least as important as hardware capabilities. Here we have some promising indicators. New methodologies for developing portable and modular application programs are starting to prove themselves. Some ASCI codes run on all three ASCI teraflops systems, and enhancements of those codes can be implemented in a fraction of the time that was required with traditional methods.

Algorithms that scale to 6,000 processors have been developed and implemented for a number of ASCI applications. Future ASCI machines are expected to have no more than 10,000 processors.

An objective is to provide a usable and scalable application-development environment for ASCI computing systems, enabling code developers to quickly meet the computational needs of scientists. The development staff works with code developers to identify user requirements and evaluate tool effectiveness, and it develops direct ties with platform partners and other software suppliers to ensure an effective development environment.

An essential part of the ASCI program is the challenge of solving three-dimensional computational problems that are far larger than we have ever solved before. This challenge translates to more data, more computational cycles, and a succession of more powerful computing platforms. From the compu

tational viewpoint, the issue is not simply bigger and faster, but rather a fundamental shift in the way problems are solved. Codes, tools, algorithms, and systems that fail to scale properly as the data and hardware resources grow are useless to ASCI developers.

An evolving effort is a simulation development environment that promotes as much commonality across ASCI platforms as possible. Scalability of tool and library interfaces to address hundreds and thousands of central processing units is under investigation. Infrastructure frameworks and math software needed by the ASCI code teams are in progress, and tools to support code verification and validation are being evaluated. This includes current and emerging standards for languages and programming models, such as MPI and Open MP; common application programming interfaces for parallel input/output (I/O) and linear solvers; common parallel tools; and common ways of accessing documentation.

THE CHALLENGES AHEAD

There are many challenges in addition to those associated with processing power and memory. A simplified way to look at the ASCI program is as a combination of large-scale computing resources, massive scientific databases, and tele-laboratories, all of which must be put into place to support scientific collaborations, perform large-scale simulations, explore complex data sets, and achieve scientific understanding. ASCI is attempting to build a simulation capability that provides a balanced system—one in which computing speed and memory, archival storage speed and capacity, and network speed and throughput are combined to provide a dramatic increase in the performance of scientific simulations. One really does not have a new simulation tool if all one does is collect a bunch of computers in a warehouse—it would be no more than the collections of PCs found in most universities or laboratories today. One must design and build a well-integrated computing system that eliminates the many possible bottlenecks to achieving the needed simulation performance levels (a concept very familiar to chemists).

Archival Storage Systems and Data Management

As computer systems such as those described in Table 1.1 become operational, they will produce a staggering amount of output data—hundreds of terabytes to tens of petabytes—per simulation run. We must be able to store that data and, later, recall it for analysis. This requires the development of efficient data-management techniques as well as fast input/output (I/O) mechanisms. Data management is becoming a very complex issue. There will be so much data that arbitrary file names will no longer be a reasonable format. Also, with a system consisting of thousands of processors and a complex of online disks and archival storage systems, these resources will have to be managed as a single system to provide a rational distributed computing environment.

Input/Output

Chemists are familiar with the importance of efficiently performing I/O—programs like GAUSSIAN do a tremendous amount of I/O. I/O is critical for many scientific simulations in addition to quantum chemistry applications. I/O has been a problem for massively parallel computers, for few systems in the past had truly scalable I/O subsystems. The new machines described in Table 1.1 provide much-enhanced I/O capabilities and, for the first time, we are beginning to find truly scalable I/O subsystems in massively parallel computer systems.

Networks

There will be a need for high-speed networks to connect users to the archival storage systems so that they can access and analyze the results of past simulations. Much of the analysis of the output of the simulations will be done using visualization techniques, but these will not be the simple visualization techniques of today. Because of the tremendous amount of data that will be produced, the user must be able to switch easily from coarse resolution, providing a broad overview of the results, to fine resolution, providing a detailed view of a small region of the data very efficiently. Otherwise even the highest-speed networks available will be overwhelmed.

Data Analysis and Visualization

We will have to develop new techniques to allow users to analyze the data. If the needs of the applications are to be met on schedule, we must develop a Comparable schedule for the development of the infrastructure for "seeing and understanding" the results. To do this will require not just developing bigger display screens, but also innovation in how we display and interact with graphical data. To develop plans to address this problem, ASCI has collaborated with the National Science Foundation on a series of workshops. A report describing the output of these workshops was recently published that lays out a 10-year research and development agenda for visualization and data manipulation.[4] ASCI will be tackling this problem collaboratively with the academic community as well as other scientific communities.

The scientists and engineers involved in ASCI will be linked together by "data and visualization" corridors, so that users can access and interact with the needed resources. The capabilities provided by the "data and visualization corridors" are crucial for impedance matching between the computing systems, the large-scale simulations, and the users because the associated time scales are very different. People work in minutes and hours, and computer results in some cases can take days, and in some cases milliseconds, for generation. In the end, the system must operate at human time scales, for it is humans who are seeking knowledge. The development of the "data and visualization" corridors is also an effort that will be performed in collaboration with the research community in general.

Other challenges abound. For example, how do you represent the data? Isosurfaces and streamlines are possible methods; movies or images can be used to analyze many billions of numbers. The data sources may be real-time or the results of simulation, while other sources may be historical. The scientists and engineers involved in these activities are separated geographically, and there is often temporal separation because they need to refer to previous results. To guide this effort, a road map has been developed for the data-visualization efforts in the ASCI program. The road map links applications, computers, data management, and visualization corridors to ensure that all of the needed capabilities are developed in a timely fashion.

I/O is also important in analyzing the results of the simulations, to "see and understand" the information buried in the data. We estimate that scientists and engineers will need to be able to interact in real time with about 75 billion bytes of data by early 2000. Along with this is a requirement to be able to perform queries on a data archive containing roughly 30 terabytes with transfer rates of perhaps 6 gigabytes a second. This is much faster than anyone can do right now, and yet it is critical to achieving the goals of ASCI. By the time we reach the end of the year 2000, we estimate that ASCI's scientists and

[4] National Science Foundation, Information and Data Management Program. Proceedings from 1998 Information and Data Management Workshop: Research Agenda for the 21st Century. NSF, Washington, D.C.

engineers will need to interact with a quarter of a terabyte extracted from an archive of at least 100 terabytes with a 20-gigabytes-per second transfer rate. And, the growth will not stop there; the root cause of these high requirements is the detail with which ASCI must be able to simulate weapons systems.

Scientific Simulation Software

Developing scientific simulation software for SMP-based parallel computer systems presents a number of challenges because it entails use of a "hybrid programming model." The hybrid-programming model incorporates both of the standard programming models for parallel computers: the shared-memory and message-passing programming models. The shared-memory programming model relies on the ability of the programmer to access memory by its location or name. This approach provides reasonable performance with as many as 128 nodes, all sharing memory. However, although the memory within an SMP is shared, to attain reasonable performance the user must explicitly make the program parallel. Current compiler technology is not yet at the point that it can be relied on to produce efficient parallel code on more than 8 to 16 processors. To pass data between the SMP nodes, the programmer must send an explicit message that communicates the need for a piece of data. This message is passed to the node on which the data resides, which then transmits the requested data to the original node. This is the message-passing programming model. Both the shared-memory and the message-passing models have been in use for some years but with very, very few exceptions, not together. Although it is possible to use the message-passing model within the SMP node as well as between the SMP nodes, the hybrid programming model will provide the best performance for the overall system since it can take full advantage of the fast shared-memory access possible within a node. So, we have some new things to learn!

Computer Systems Software

Because the systems being acquired by ASCI are all first-of-a-kind, perhaps even one-of-a-kind systems, the computer manufacturers typically cannot provide all the software and tools that are needed to efficiently operate and develop software for the computer. ASCI must provide the missing pieces. One task that ASCI has undertaken to accelerate the development of application software for these machines is the construction of a "problem-solving environment." The PSE provides an integrated set of tools, like compilers and debuggers, for producing code for these massively parallel computers. Problem generation and verification will need another set of such tools, as will high-speed storage and I/O. Finally, the distributed computing environment as well as the associated networks will have to be enhanced as well.

THE ASCI ALLIANCES

Los Alamos, Livermore, and Sandia are not alone in this process. The academic community is contributing to understanding how to create the advanced simulations and how to run these simulations on terascale platforms. ASCI is funding five large university teams, each of which is attacking a very complex, multi-disciplinary simulation problem. These teams are in the first year of their initial 5-year funding and have already contributed a number of ideas both in computer science and in science and engineering. There are also a number of other university groups, about 30 in all, that are working on more focused aspects of the problem.

Caltech is one of the five university groups receiving ASCI funding. The Caltech team is tackling the reaction of various materials to shock waves. In the Caltech program a detonation is used to create a shock wave in a shock tube, which then impacts the test material. The goal is to develop techniques to simulate the effect of the shock wave on the test material and to understand how that effect is related to the composition of the test material. To simulate this very complex phenomenon, we must be able to simulate fluid flows, explosions, chemical reactions, shock waves, and materials properties. The effort requires simulations at the atomic level all of the way up to the macroscopic level. This type of academic research will contribute to a general understanding of how to perform complex, multi-disciplinary simulations and so will help general science as well as ASCI. There are a large number of scientific and engineering applications that need to be able to simulate such complex processes.

ASCI'S CONTRIBUTIONS TO THE GREATER SCIENCE COMMUNITY

ASCI will contribute to science in several ways. It will provide the foundations for building a "science of simulation." Through ASCI we will gain a detailed understanding of all the topics and methods that are needed to create multi-physics simulation, multi-system engineering-design codes, and use them to carry out predictive simulations of complex phenomena. Science and industry alike will benefit from ASCI's development of vastly more powerful computational methods, algorithms, software tools, methodologies for program development, validation, and verification, code frameworks, and computing systems with ever greater computing power and massive memories.

ASCI projects are devoting considerable attention and resources to the development of physics models that are grounded on theory and experiment and lead to practical numerical computations. These models are then incorporated into the simulation programs and validated against experimental data and predictions based on accepted theory. While this approach is not unique, the difference here is the computational scale that is possible and use of multiple, intertwined physics models in a single simulation code.

ASCI is investing in better algorithms and computational methods. Even ASCI's extremely powerful platforms will not be able to tackle the weapons stockpile computations by applying brute-force algorithms.

Software design and development for such applications on such machines are particularly challenging and controversial aspects of multi-physics scientific simulations. The experience base is currently limited to a few isolated projects from which it is difficult to draw general conclusions. Instead, ASCI will carry out multiple substantial simulation projects that will provide proofs of principle, at scale, of software development methodologies and frameworks that provide portability, extensibility, modularity, and maintainability while delivering acceptable performance.

ASCI will have shown, through actual accomplishments, the potential of terascale simulations for a broad range of scientific and engineering applications. By doing so, and by demonstrating the feasibility of putting together and using computers of much greater scale than would otherwise be available, ASCI is also re-energizing both the scientific computing research community and the U.S. commercial computer industry. Several federal agencies are already planning to invest in very high-end computing, and others are sure to follow. Computer manufacturers are seeing renewed customer interest in much higher performance, and there is growing realization that there may be a financially viable approach to getting into the highest-end markets. The strategy consists of designing the standard mid-to-large-scale products that sell in large quantities so that they are also suited to serve as building blocks for truly large systems. In addition to stimulating the high-end computer industry, industry will inherit computational tools that will enable it to design better, more competitive products. Computer simulation is already

used extensively in product design and optimization. The much greater simulation capabilities developed by ASCI, hardware and software, will enable industries of all kinds to produce even better products.

Perhaps the most important product of ASCI will be a cadre of thousands of people with solid experience and expertise in predictive, multi-physics simulation. This alone will have a major impact on the future of scientific and engineering computing.

SUMMARY

Over the next few years, we will see remarkable increases in the power of high-end computing capabilities, from 1 teraflops today to 100 teraflops in 2004. DOE's Accelerated Strategic Computing Initiative is driving this increase and is using the resulting computer systems immediately—as soon as they are manufactured—to address a broad range of scientific and engineering simulations. These simulations are critical to ensuring the safety and reliability of the nation's nuclear arsenal. But because the approach is applications-driven, the computer systems being developed will lead to computers that are suitable for most, if not all, high-end technical computing applications.

To the extent possible, ASCI is using commercial building blocks to leverage the cost efficiency of the high-volume computing industry. This approach will make the resulting computer systems readily replicable for others—and, undoubtedly, at substantially reduced cost, because ASCI is funding much of the needed development efforts. The recent report from the NSF/DOE-sponsored "National Workshop on Advanced Scientific Computing" outlined a number of scientific and engineering areas that would be dramatically affected by access to this level of computing capability.[5]

Taking advantage of this new level of computing capability presents a large number of challenges. We have to consider the computing infrastructure as a whole—processing speeds and memory, data storage and management, data analysis and visualization, and networking speed and capacity—if we are to realize the promised dramatic increase in computing capability. We will also have to discover new computational algorithms, programming methods, and even problem formulations in some cases. Despite these challenges, the payoff will be substantial. The advances in computing power that will become available in the next 5 to 10 years will be so great that they will change the very manner in which we conduct the pursuit of science and technology.

DISCUSSION

David Rothman, Dow Chemical Company: Let me say you have succeeded in convincing me how puny my computing project is right now. But one thing I think we have in common is that in any computing project you can identify a number of potential rate-limiting steps in getting to the goal, and you can generally identify one or two things that are very high risk items. Can you tell me what you see as the one or two of the highest-risk parts of this overall program you have to reach your goal?

Paul Messina: I would consider that the highest risk is the operating system and then, software, and tools. Also risky are general issues such as whether the compilers will reach the point of having a respectable fraction of the potential speed of the system. The next highest risk would be getting the algorithms in the time scales where they use fairly efficiently the thousands of processors.

[5] Department of Energy and National Science Foundation. Report from the "National Workshop on Advanced Scientific Computing" held July 30-31, 1998, in Washington, D.C., J.S. Langer, ed.

I have no doubt whatsoever about the ability to do that for scientific algorithms given enough time, so it is the accelerated schedule that is a source of high risk; people are not used to having to program that way. So, it is a good combination of the system's software, which tends to lag pretty badly in terms of its ability to exploit these very complicated processors, and then the applications.

The one thing I did not emphasize at all is that the way we have obtained this wonderful Moore's curve for the processor increase in speed is not so much by having faster clocks on our cycle speeds, but by making the guts of the processor much more complicated. And so, theoretically, if everything works you get this wonderful speed. But it means keeping several arithmetic units busy, loading and storing data simultaneously, and that is a hard thing to do. So, to actually get the benefits of these faster processors, even one processor, is beyond what many compilers can do.

Andrew White, Los Alamos National Laboratory: Thom mentioned the Science Simulation Plan and the PITAC response as a new program for FY2000. How would this new initiative and ASCI/Stockpile Stewardship play together?

Paul Messina: One of the things that I hope to do in part during my 2 years in ASCI is to work out a way that the two efforts can get mutual benefit. So, specific things that I can imagine are that perhaps the SSX SSStar program will select machines that are similar to the ones that ASCI has selected. Then, these two things that I just identified as the highest-risk items, the system software and tools and the algorithms, could be developed jointly so that we would be able to actually use the machines. I would say that is a real target of opportunity, to have very similar machines instead of deciding to diverge here and therefore have to develop all new and different algorithms.

So, at the level of helping to mature the system, I would first think of software, tools, and the algorithms for the applications. One could imagine sharing facilities. That is practical, but often politically unpalatable, so I do not know that there is much hope of doing that. But I think that in figuring out how to use these new systems, making them robust earlier and sharing the results, and maybe even having joint teams doing applications, would help both tremendously.

Thom Dunning: I think one of the real benefits of ASCI is that it has made computational scientists start to ask such questions as, What would we do if we had 100 times more power on the desktop or 10,000 times more power on the very high-end systems? These are very good questions to be thinking about because that is the direction that the computer industry is taking us, whether it succeeds on the accelerated time scale as outlined with ASCI or whether it comes on a little more slowly.

2

Software Development for Computational Chemistry: Does Anything Remain to Be Done?

Peter R. Taylor
San Diego Supercomputer Center and University of California, San Diego

We consider the state of the art of computational chemistry and then discuss to what extent this state of the art meets the requirements of the chemistry community. Our overview is necessarily broad and somewhat superficial, but it supports the view that while computational chemistry is a mature and very successful field, considerable effort is still needed at the level of fundamental research into methods, algorithms, and implementation, and in training students in these areas.

For the purposes of this paper, we consider computational chemistry to comprise the study of the structure, properties, and dynamics of chemical systems. We recognize that this somewhat narrow definition does not properly incorporate areas such as process modeling that are also of importance to the chemical industry, but the discussions at the workshop strongly suggest that the successes and, not “failures," but let us say "unmet expectations," in these other areas are not different in cause or in possible remedies from those in our narrower definition of computational chemistry. With this definition, then, we can argue with some justification that computational chemistry is one of the great scientific success stories of the past decades. Twenty-five years ago quantum-chemical calculations were performed by quantum chemistry specialists, and the results of such calculations were rarely and with difficult acceptance published in general chemistry journals such as the *Journal of the American Chemical Society*. At that time calculations on molecules of more than a dozen atoms were a rarity, and the accuracy of the predictions was often overstated and seldom convincing to experimentalists. We need not labor the point here: suffice it to say that the situation today is exactly the reverse. Indeed, the field of computational chemistry has just been recognized by the award of the 1998 Nobel Prize in chemistry to Pople and Kohn. Might it not, therefore, be time to declare success in this endeavor, and that the field is essentially complete? In this essay, we argue not only that this would be a grievous mistake, but also that considerable additional effort is required in a number of areas for computational chemistry to reach its full potential. Our field is mature, not complete; ripe with opportunities, not sterile.

The major activities in computational chemistry can be classified as molecular electronic structure (quantum chemistry), reaction dynamics, and molecular dynamics. These are used to calculate, respectively, the properties of individual molecules or groups of molecules, kinetic information and reaction

pathways for chemical reactions, and the properties and dynamics of large molecules or large assemblies of molecules. All of these activities have been extraordinarily successful over the past decades and have firmly established computational chemistry as a third methodology alongside experiment and theory. Computational chemists have also been among the foremost users of computer hardware, with substantial requirements for computer time, memory, and disk space, and with a history of exploiting both early access to new architectures and novel solutions to reduce the cost of performing their research. They have successfully exploited new vector and parallel computer architectures as they have become available, and at the same time have developed new algorithms to efficiently use first minicomputers and later RISC workstations, and most recently clusters of commodity PCs. The advent of ASCI-class computing resources presents a new opportunity—a vast increase in computational capability to be exploited. However, it is not obvious that the community is ready to use such massively parallel machines, not the least because even our scalable algorithms have generally been tested in situations only up to a few hundred processors. To use the full power of an ASCI-class machine will require successfully harnessing at least an order of magnitude more processors. Will our current algorithms, and at an even higher level, the computational chemistry methods they implement, be suitable for such architectures? It is these questions we concentrate on here, not the detail issues of whether message-passing implementations are more or less appropriate than shared-memory implementations of our methods. We demonstrate, for example, that much of our current methodology is incapable of extending the accuracy of our description of molecular systems beyond what we can currently achieve, and that new methods must be sought.

We take quantum chemistry, as defined above, as an example. The information for nonempirical dynamics calculations comes from quantum-chemical calculations, so in this sense quantum chemistry is fundamental to nonempirical computational chemistry. Typical quantum chemistry calculations, in 1999, treat a single molecule or perhaps a group of molecules, in vacuo, at a temperature of 0 K. The accuracy achievable varies with the size of the molecule (see Figure 2.1), but the highest-accuracy work is comparable to experimental accuracy for many properties. A simple way to represent the relationship between molecular size and accuracy of results is a graph generally referred to as a Pople diagram.

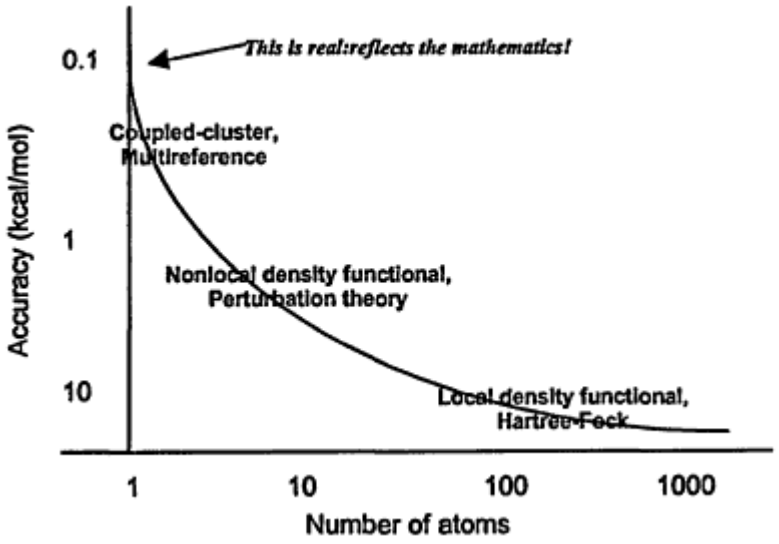


Figure 2.1
Size/accuracy relationship in ab initio calculations.

Figure 2.1 shows the accuracy achievable for different sizes of molecule (or assembly of molecules) and the most common quantum-chemical methods used to achieve that accuracy. One important aspect of Figure 2.1 is the intersection of the curve with the ordinate axis at around 0.1 kcal/mol. This is not a resolution artifact or a mistake in the figure—it reflects a fundamental limitation in our ability to describe the quantum-mechanical motion of the electrons in a molecule, specifically, the cusp behavior as two electrons approach one another. The methods listed in the figure are all based on expansions in one-electron functions, and such expansions are inherently incapable of describing the cusp in a finite number of terms. Increasing the number of terms cannot proceed indefinitely, not only because the calculations become impossibly large, but also because the resulting expansion function sets become linearly dependent. Here is one example of a fundamental problem that must be addressed in order to increase our computational chemistry capabilities—we must develop new methods that better approximate the molecular wave function.

An immediate question might be, Is such increased accuracy really necessary? A flippant response would be that everyone would like more accuracy: like money, or network bandwidth, it is something one cannot have too much of. A more realistic answer would be that an organic chemist planning a synthesis might find herself in a situation where a difference in barrier heights of a few tenths of a kcal/ mol determines whether the ratio of desired product to useless alternative is 9:1 or 1:9. Barrier heights known to 1 kcal/mol or so would be quite useless in this situation. And, as noted above, the accuracy of computed quantum-chemical energies determines the reliability of nonempirical dynamics calculations, so if more accurate dynamics calculations are desired, the accuracy of the quantum-chemical calculations must be improved.

The other axis in Figure 2.1 is molecular size. Again, the crucial question is how to increase the size of the system we can treat. This is not merely in order to treat larger molecules per se, although this is one consideration. Another important issue is the realistic treatment of environment. Most chemistry takes place in the condensed phase or in a gas phase of some density, not in a vacuum. Although some progress has been made in describing solvent effects by embedding the system in a dielectric medium, there is good evidence that detailed solute-solvent interactions are so important in many situations that some solvent molecules must be treated explicitly. We must therefore increase the size of our systems to include perhaps tens of solvent molecules, and thus must develop methods to handle (much) larger systems.

We reiterate that we are not suggesting here that access to much more powerful computers alone is an adequate solution to the accuracy and size problems (indeed, in the case of increased accuracy, it is manifestly not adequate). What is needed is to improve or to develop new methodologies so that they can handle the problems of interest in the way we wish to treat them, taking advantage of the increased computing power of ASCI-class machines to achieve that end. The first steps must be in methodology, not in porting existing approaches to new architectures.

What of other components of computational chemistry, such as reaction dynamics or molecular dynamics simulations? A chat with practitioners in these areas is likely to indicate a need to perform more accurate dynamics studies, on larger systems, including the effects of bulk environment, etc., plus an additional concern about studying phenomena at longer time scales (that is, how best to simulate phenomena that take place in nature on a time scale of microseconds to milliseconds when normal simulation time steps are at the femtosecond level). In the case of increased accuracy and larger systems, the issues are of course similar in spirit to those discussed above in the context of quantum chemistry. Indeed, increased accuracy in reaction dynamics and simulations would probably mandate increased accuracy (for larger systems) in the quantum-chemical calculations used to provide potentials

for the dynamics. Our attention is again focused on the need to develop new methods rather than to simply look at ways to implement existing methods or to port existing programs to new architectures.

Development of new computational methods will lead to new implementations for those methods, which brings additional advantages. First, there is the opportunity to consider explicitly the nature of new computer architectures when developing the methods. For example, methods that might have seemed inappropriate in the past because their implementation would have enormous memory requirements may be perfectly feasible in the large memory environment provided by ASCI-class machines. Second, new implementations can take advantage of modern software engineering practices and modern computer languages, leading to increased ease of maintenance compared to the traditional "dusty decks." Third, new methods and implementations can take advantage of modern technologies, such as the ability to store, retrieve, and manipulate large data sets, or interaction environments via which users can not only visualize their data but also steer simulations.

As we noted at the beginning of this essay, computational chemistry has been a remarkably successful subdiscipline. Nevertheless, we repeat also that success and maturity should not be confused with completion; our efforts have been very successful, but much remains to be done. A primary need is to encourage the development of new methods, and their implementation, rather than just seeking to reimplement existing methods on new hardware. Efforts are needed to increase the accuracy of our results and to achieve that accuracy for larger and larger systems, to model experiments more realistically, and to develop software appropriate to new hardware so as to fulfill the opportunities offered by ASCI-class machines. These efforts, in turn, require strong support from agencies and institutions, not only to support research and development activities, but also to provide a cadre of trained computational scientists through support of education and training.

Acknowledgments: The author was supported by the National Science Foundation through Cooperative Agreement DACI-9619020 and by Grant No. CHE-9700627.

DISCUSSION

Sam Kounaves, Tufts University: Considering the amount of investment that has been put into these types of supercomputing centers and initiatives, one of the final outcomes is, of course, the ability to transfer some of this technology to other areas. In my own particular specialty, for example, we would love to have more in terms of modeling electrode position nucleation for the fusion phenomena. Microsoft last year spent $3 billion just to develop software, and that may not be significant. But are there any plans for transferring? What type of systems do you foresee, for example, to make it more generalized, to make it accessible to other communities in the chemistry world?

Peter Taylor: There are a couple of efforts, I think. Some of the work done at EMSL (Environmental Molecular Sciences Laboratory) at PNNL, in Northwest Chemistry Codes, and some of the activities done under our own support in San Diego and our larger partnership, National Partnership for Advanced Computational Infrastructure, attempt to do just this. And I think there has been some success with some of these things already. Both of these projects really are just getting rolling in terms of dealing with the outside community. But we have plans to do this and I would expect to see the fruits of this over the next couple of years.

That, I think, is something we can do in cases where we already have the methods that we need. If new methods have to be developed, then, of course, one has to be less specific about the lead time. We simply do not know how long it will take to develop some of the new methodology, and so forth. But in

a number of communities, certainly those supported by the Department of Energy and in the NSF program, there is a desire to harden these things and move them out so that use can be made of them in a much larger community.

Peter Cummings, University of Tennessee and Oak Ridge National Laboratory: I just wanted to comment, Peter, about the things that you said were coming from the molecular dynamics community, which I will just generalize as the classical simulation community. I think it is not simply an issue of larger systems. It is also longer simulation times; more complex systems, like proteins and polymers, all have very much longer relaxation times. And I think that one of the challenges facing parallel computing is how to get beyond these much longer time scales, because parallelization does not solve that problem. And, in fact, I think there may be a lot of potential in non-dynamical—that is, Monte Carlo-type—techniques that subvert the time-scale problem. Rethinking those on parallel architectures, I think, can lead to a lot of progress.

Peter Taylor: I think that is a very important point. I certainly would not want anybody to go away with the assumption that I felt I had listed all of the issues here. The one of time scale is important enough that I should have said something about it. It is also another useful illustration of one of the points I was making, because that is an area, clearly, where effort needs to be invested in developing new methods. Getting beyond longer time scales is not something we know how to do today. We need to try to find novel ways, as you say, to subvert, in essence, the time-scale problem faced otherwise in traditional dynamics.

Judith Hempel, University of California, San Francisco: I would like to raise a question that I cannot answer. What is the impact, really, on computational chemistry of the increase in the computing power of the ASCI initiative, and what if that were only the beginning? So, in a sort of visionary sense, if computing power were to increase at this same rate or even at an accelerated rate, what would be the impact of computational chemistry on the world as we know it?

Peter Taylor: Well, I think the issue is once again related to some of these questions about accuracy and about the size of the calculations. If you look at, for example, small-molecule chemistry, then typically the accurate methods we have scale roughly on the order of the fifth or sixth power of the size of the molecule. So with a computer 100,000 times faster than what you have today, you would still only be able to study a system 10 times larger. Now, we would all like to be able to do systems 10 times larger, but that capability does not necessarily seem that much of a reward for a five-orders-of-magnitude gain in computing power. If we really want to see dramatic qualitative changes in the type of work we can do with computational chemistry, we need to put a lot more effort into figuring out how to make the calculations scale better with the size of the system.

Judith Hempel: Do you see a time in the future when quantum mechanics will take over for the empirical methods, like the force-field methods that now use quantum mechanics to extract parameters?

Peter Taylor: What I see is an increasing use of, in essence, hybrid methods. At some level, even with something very large like an enzyme, I do not think it should be necessary—I would like to think we are more ingenious than this—to treat the entire enzyme and the medium in which it is found quantum mechanically. But I suspect that in order to be able to do a good job in predicting the catalysis, the actual

details of what happens at the bottom of the active site in the enzyme, we will need to use quantum mechanics.

The question is, If you start using quantum mechanics there, where do you switch over as you move out in space to a point where you can go over to a lower-level treatment, a semi-empirical or empirical method? There has been a lot of effort in that recently. Some efforts have been more successful than others. I think this is something we will work out over the next few years, and I think this is the real future. Ideally there should be no need to treat a system like that entirely quantum mechanically. If we could do it cheaply enough, well, then we could do it that way. But in practice I do not think that should be necessary.

But I think we do not completely understand yet how to splice these different levels of treatment together. How do we go, for example, from a very accurate quantum chemistry calculation for an area where we are really interested in the details to a lower-level quantum chemistry calculation, say a density functional calculation or something in a larger region, and then eventually to some completely empirical type of molecular mechanics approach in the next layer out?

Thom Dunning: Let me make one addition to that statement. One of the surprises that came out of the ASCI program was that the increased computing power makes it possible—even though, as Peter pointed out, the techniques that we use right now do not scale very well with the size of the system—to actually compute all of the thermodynamics for all of the species and the reactions involved in combusting both gasoline and diesel fuel. And, in gasoline there are some 1,000 species that are thought to be involved, and some 2,000 reactions that the modelers say they need to have kinetic data for.

For diesel fuel you are talking about 2,000 species and 4,000 reactions. And some of those species have not even been seen or characterized in the laboratory yet, so even with the techniques as they are right now, one can, without increasing computing power, actually make substantial contributions to very practical problems that the country faces. But, I would agree entirely with Peter in that we need techniques that scale better with the size of the system.

3

Recent Advances in Computational Thermochemistry and Challenges for the Future

Larry A. Curtiss,
Argonne National Laboratory
and
John A. Pople,
Northwestern University

INTRODUCTION

Knowledge of the thermochemistry of molecules is of major importance in the chemical sciences and is essential to many technologies. Thermochemical data provide information on stabilities and reactivities of molecules that are used, for example, in modeling reactions occurring in combustion, the atmosphere, and chemical vapor deposition. Thermochemical data are a key factor in the safe and successful scale-up of chemical processes in the chemical industry. Despite compilations of experimental thermochemical data for many molecules, there are numerous species for which there are no data. In addition, the data in the compilations are sometimes incorrect. Experimental measurements of thermochemical processes are often expensive and difficult, so it is highly desirable to have computational methods that can make reliable predictions.

Since the early 1970s when ab initio molecular orbital calculations became routine, one of the major goals of modern quantum chemistry has been the prediction of molecular thermochemical data to chemical accuracy (±1 kcal/mol). One of the problems was that the Hartree-Fock calculations done in the 1970s gave large errors (up to 100 kcal/mol) in bond energies. Prediction of accurate thermochemical data required going beyond Hartree-Fock theory to include a sophisticated treatment of electron correlation, which made the calculations very difficult. After several decades of work, considerable progress has been made in attaining the goal of a ±1 kcal/mol accuracy through advances in theoretical methodology, development of computer algorithms, and increases in computer power. It is now possible to calculate reliable thermochemical properties for a fairly wide variety of molecules.

At the ab initio molecular orbital level, the methods currently used for computing thermochemical data range from very high levels of theory to those that combine moderate levels of theory with some form of empirical input. The former is limited to smaller molecules and can attain accuracies of ±0.5 kcal/ mol, while the latter methods can be applied to larger molecules with somewhat less accuracy. The Gaussian-n methods[1] that we have developed fall in the latter category and have been widely used for

[1] For recent reviews see: L.A. Curtiss and K. Raghavachari, in *Computational Thermochemistry*, K.K. Irikura and D.J. Frurip, eds., ACS Symposium Series 677, American Chemical Society, Washington D.C. (1998), pp. 176-197; L.A. Cuss and K. Raghavachari, in *Encyclopedia of Computational Chemistry*, P.v.R. Schleyer, ed., John Wiley, New York (1998).

predicting thermochemical data of molecules. In the second section of this paper we summarize current state-of-the-art methods in computational thermochemistry. In the third section we describe in more detail the Gaussian-n approach and then discuss a recent development in this area, Gaussian-3 (G3) theory, which achieves a new level of accuracy. Finally, in the fourth section we assess prospects for the future of computational thermochemistry.

CURRENT STATE OF THE ART IN COMPUTATIONAL THERMOCHEMISTRY

The available methods for computing thermochemical data range from empirical schemes to ab initio molecular orbital theory.[2] In this paper we restrict our discussion to the ab initio based methods. They can be roughly divided into three types: (1) very high level quantum chemical calculations with no experimental input; (2) composite techniques that combine moderate level ab initio quantum chemical calculations with some form of molecule-independent empirical parameters; and (3) techniques that use molecule-dependent empirical parameters obtained from accurate experimental data in combination with moderate level ab initio quantum chemical calculations. In this section we discuss some examples of these different approaches.

In principle it is known how to compute the thermochemical properties of most molecules to very high accuracy (0.5 kcal/mol). This can be achieved by using very high levels of correlation, such as are obtained with coupled clustered [CCSD(T)] or quadratic configuration [QCISD(T)] methods, and very large basis sets. The results of these calculations are then extrapolated to the complete basis set limit and corrected for some smaller effects such as core-valence effects and atomic spin-orbit effects. Unfortunately, this approach is limited to small molecules because of the ~n^7 scaling (with respect to the number of basis functions) of the correlation methods and the large basis sets used. This methodology has been used by Dunning, Feller, and coworkers[3,4] at Pacific Northwest National Laboratory to systematically study a large number of small molecules having one and two non-hydrogen atoms. They have applied these very high level calculations to a diverse enough set of molecules to show that the methodology does perform to a very high level of accuracy.

Other groups have also used this type of approach to computational thermochemistry. Grey, Janssen, and Schaefer[5] have used CCSD(T) with large basis sets to study the thermochemistry of CH_n and SiH_n hydrides, and some of their cations. They achieved bond energies accurate to 0.5 kcal/mol without any empirical corrections for these small molecules. Petersson and coworkers[6] have used QCISD(T) with very large basis sets and have obtained a mean absolute deviation of 0.53 for a subset of the G2 test set of reaction energies. Bauschlicher, Langhoff, Taylor, and coworkers[7] have used an approach based on converging to the one-particle limit through the use of atomic natural orbitals at a moderate level of correlation treatment. The correlation treatment is calibrated against full configuration interaction calculations on smaller systems or against accurate experimental data, in some cases. They have achieved accuracies of 1 kcal/mol or better using these methods.

Since the very high level calculations are difficult to extend to larger molecules, an alternative

[2] For a recent review see: *Computational Thermochemistry*, K.K. Irikura and D.J. Frurip, eds., ACS Symposium Series 677, American Chemical Society, Washington D.C. (1998).
[3] K.A. Peterson and T.H. Dunning, Jr., *J. Chem. Phys*. 106, 4119 (1997).
[4] D. Feller and K.A. Peterson, *J. Chem. Phys*. 108, 154 (1998).
[5] R.S. Grev, C.L. Janssen, and H.F. Schaefer III, *J. Chem. Phys*. 97, 8389 (1992).
[6] J.A. Montgomery, Jr., J.W. Ochterski, and G.A. Petersson, *J. Chem. Phys*. 101, 5900 (1994).
[7] C.W. Bauschlicher, Jr., and S.R. Langhoff, *Science* 254, 394 (1991).

approach is to use a series of high-level correlation calculations [e.g., QCISD(T), MP4, CCSD(T)] with moderate sized basis sets to estimate the result of a more expensive calculation. The Gaussian-n[8] series described in the next section exploits this idea to predict thermochemical data. In addition, molecule-independent empirical parameters are used in these methods to estimate the remaining deficiencies in the calculation. This will work if the remaining deficiencies are systematic and scale as the number of pairs of electrons. Such an approach has been quite successful in the Gaussian-n series and the latest version, Gaussian-3 theory, achieves an overall accuracy of 1 kcal/mol and is computationally feasible for molecules containing up to about eight non-hydrogen atoms.

Petersson et al.[9] have developed a related series of methods, referred to as complete basis set (CBS) methods, for calculating energies of molecular systems. The central idea in the CBS methods is an extrapolation procedure to determine the projected second-order (MP2) energy in the limit of a complete basis set. This extrapolation is performed pair by pair for all the valence electrons, and is based on the asymptotic convergence properties of pair correlation energies for two-electron systems in a natural orbital expansion. As in the Gaussian-n methods, the higher-order correlation contributions are evaluated by a sequence of calculations with a variety of basis sets. Several empirical corrections, similar in spirit to the higher-level correction used in G2 theory, are added to the resulting energies in the CBS methods to remove systematic errors in the calculations

The third approach is the use of molecule-dependent empirical parameters in combination with a moderate level ab initio molecular orbital method. The use of isodesmic reactions is a primary example of this approach, which can be quite accurate for molecules having no unusual bonding. In the isodesmic approach, a reaction is chosen with the same number of chemical bonds of each formal type (e.g., C-C, C=C, C-N, C-H) on both sides of the reaction, and all of the species have accurate experimental thermochemical data available except the species of interest.[10] A moderate level of theory is used to calculate the reaction energy, and the enthalpy of formation of the unknown species is then extracted. Use of specially chosen reactions can give quite accurate enthalpies of formation because of cancellation of correlation effects and also because they make use of accurate experimental data. However, suitable reference molecules are often not available, making this method inapplicable in many cases.

An extension of the isodesmic reaction scheme is the use of bond additivity corrections. An example of this method is the BAC-MP4 method of Melius and coworkers[11] that uses a computationally inexpensive molecular orbital method and combines it with a bond additivity correction. This procedure uses a set of accurate experimental data to obtain a correction for different types of bonds that is then used to adjust calculated thermochemical data such as enthalpies of formation. Quite accurate results can be obtained if suitable reference molecules are available and if the errors in the calculation are systematic. The bond additivity approach can also be used in combination with computationally more demanding methods such as G2 or G3 theory. In these cases an accuracy of 0.5 kcal/mol can be achieved for large molecules.

[8] For recent reviews see: L.A. Curtiss and K. Raghavachari, in *Computational Thermochemistry*, K.K. Irikura and D.J. Frurip, eds., ACS Symposium Series 677, American Chemical Society, Washington D.C. (1998), pp. 176-197; L.A. Curtiss and K. Raghavachari, in *Encyclopedia of Computational Chemistry*, P.v.R. Schleyer, ed., John Wiley, New York (1998).

[9] J.W. Ochterski, G.A. Petersson, K. Wiberg, *J. Am. Chem. Soc*. **117** , 11299 (1995); J.W. Ochterski, G.A. Petersson, and J.A. Montgomery, Jr., *J. Chem. Phys*. **104**, 2598 (1996).

[10] K. Raghavachari, B.B. Stefanov, and L.A. Curtiss, *J. Chem. Phys* . **106**, 6764-6767 (1997).

[11] P. Ho and C.F. Melius, *J. Phys. Chem*. 94, 5120 (1990); M.D. Allenedorf, C.F. Melius, P. Ho, and M.R. Zachariah, *J. Phys. Chem*. **99**, 15285 (1995).

GAUSSIAN-N THEORY

Ideally, a method for computation of thermochemical data has several requirements to be successful. (1) It should be applicable to any molecular system in an unambiguous manner. (2) The method needs to be computationally efficient so that it can be widely applied. (3) It should be able to reproduce known experimental data to a prescribed accuracy and be applied with similar accuracy to species having larger uncertainty or for which data are not available. The Gaussian-n methods[12] were developed with these objectives in mind.

Gaussian-2 (G2) theory[13] was the second in this series of methods. G2 theory is a composite technique in which a sequence of well-defined calculations is performed to arrive at a total energy of a given molecular species. It was developed for predicting molecular systems containing the elements H and C1, and extended subsequently to third-row non-transition metal elements. The goal was an accuracy of ±2 kcal/mol. There are several steps to G2 theory:

1. Geometries are determined using second-order Møller-Plesset perturbation theory (MP2) with the 6-31G(d) basis set.
2. Zero-point energies are determined at the Hartree-Fock level and scaled to account for known differences between experiment and theory.
3. A series of correlation-level calculations are done using perturbation theory up to fourth-order and quadratic configuration interaction. Large basis sets including polarization and diffuse functions with 6-311G(d,p) as the starting point are used in the correlation calculations. The energies are added together to obtain a total energy.
4. A higher-level correction is added to the total energy in step 3 to account for systematic errors in the energy calculation that scale as the number of pairs of electrons. This is a *single molecule-independent* empirical parameter that is chosen by fitting to a set of accurate experimental data.
5. The zero-point energy is added to the energy in step 4 to obtain a final total energy that is used to calculate thermochemical properties.

A test set of 125 reaction energies having well-established experimental values, referred to as the G2 test set, was used to assess the reliability of G2 theory. The test set was limited to molecules containing one or two non-hydrogen atoms. Since its publication in 1991, G2 theory has been widely used for the prediction of thermochemical data such as enthalpies of formation, bond energies, ionization potentials, electron affinities, and proton affinities.

Evaluation of G2 theory indicates that it meets the first requirement listed at the beginning of this section, but only partially meets the second and third ones. Since G2 theory needs no empirical input that depends on the type of molecule, it can be unambiguously applied to any molecular system. Due to its computationally intensive correlation treatment, G2 theory can handle systems with up to five or six non-hydrogen atoms, but becomes prohibitive beyond about benzene because of the ~n^7 scaling of these methods. On the G2 test set of small molecules, G2 theory is able to reproduce adequately (average absolute deviation of 1.2 kcal/mol) the experimental data; however, on certain larger molecules it fails.

[12] For recent reviews see: L.A. Curtiss and K. Raghavachari, in *Computational Thermochemistry*, K.K. Irikura and D.J. Frurip, eds., ACS Symposium Series 677, American Chemical Society, Washington D.C. (1998), pp. 176-197; L.A. Curtiss and K. Raghavachari, in *Encyclopedia of Computational Chemistry*, P.v.R. Schleyer, ed., John Wiley, New York (1998).

[13] L.A. Curtiss, K. Raghavachari, G.W. Trucks, and J.A. Pople, *J. Chem. Phys*. **94**, 7221 (1991).

The G2/97 test set of larger molecules (302 reaction energies) was recently developed[14] to provide a more stringent test of G2 theory and other methods. An assessment of G2 theory on this test set indicated that it failed for certain types of molecules such as halogen-containing species and systems with unsaturated rings. For example, the G2 enthalpy of formation for SiF_4 is off by 7.1 kcal/mol, and for benzene it is off by 3.9 kcal/mol. Hence, although the average errors in G2 theory are less than 2 kcal/mol, the maximum errors are too large to allow for full confidence in the method.

We recently developed Gaussian-3 (G3) theory,[15] which is significantly more accurate than G2 theory and eliminates most of its deficiencies. G3 theory is a composite technique similar in spirit to G2 theory, but with some new features. Steps 1 and 2 in G2 theory remain the same. Step 3 is modified with different basis sets that are more uniform, yet are smaller so that the calculations require fewer computational resources (although they still scale as ~n^7). The higher-level correction in step 4 is reformulated in terms of four parameters, but remains molecule-independent and is obtained by fitting to a larger set of accurate experimental data. Two new steps are added: a correction for core correlation at the second-order perturbation level using a new basis set, G3 large, and a correction for spin-orbit effects in atoms.

G3 theory was assessed on a total of 299 energies (enthalpies of formation, ionization energies, electron affinities, and proton affinities) in the G2/97 test set.[16] The average absolute deviation from experiment of G3 theory for the 299 energies is 1.01 kcal/mol. For the subset of 148 neutral enthalpies of formation the average absolute deviation is 0.93 kcal/mol. The corresponding deviations for G2 theory are 1.49 and 1.56 kcal/mol, respectively. The improvement over G2 theory is shown in Figure 3.1 for different types of molecules in the G2/97 test set. Many of the deficiencies in G2 theory for the G2/97 test set have been eliminated. Of particular importance is the improvement for 35 non-hydrogen systems, such as SiF_4 and CF_4, for which the average absolute deviation decreases from 2.54 kcal/mol (G2 theory) to 1.72 kcal/mol (G3 theory). Another significant improvement is found for the 47 substituted hydrocarbons in the test set, for which the average absolute deviation decreases from 1.48 to 0.56 kcal/mol. The deviations from experiment for a series of hydrocarbons are given in Figure 3.2 and show excellent agreement with experiment.

In summary, the Gn approach uses a moderate level of ab initio molecular orbital calculation, combined with molecule-independent parameters to account for systematic deficiencies in the calculations. The latest method in this series, G3 theory, gives thermochemical data accurate to 1 kcal/mol with small maximum deviations for molecules containing up to about eight non-hydrogen atoms. The applicability to larger molecules remains to be investigated.

FUTURE OUTLOOK

All three approaches to computational thermochemistry discussed above will play important roles in the future. It should be stressed that, despite the successes, much remains to be clone in the development of computational thermochemistry methods.

One of the major challenges for these methods is the ~n^7 scaling and how it limits the size of molecules for which thermochemical data can be computed. The increase in computer time with size for some typical hydrocarbons is shown in Figure 3.2. Current limitations and capabilities of the various

[14] L.A. Curtiss, K. Raghavachari, P.C. Redfern, and J.A. Pople, *J. Chem. Phys*. **106**, 1063 (1997); L.A. Curtiss, P.C. Redfern, K. Raghavachari, and J.A. Pople, *J. Chem. Phys*. **109**, 42 (1998).

[15] L.A. Curtiss, K. Raghavachari, P.C. Redfern, V. Rassolov, and J.A. Pople, *J. Chem. Phys*. **109**, 7764 (1998).

[16] L.A. Curtiss, K. Raghavachari, P.C. Redfern, and J.A. Pople, *J. Chem. Phys*. **106**, 1063 (1997), L.A. Curtiss, P.C. Redfern, K. Raghavachari, and J.A. Pople, *J. Chem. Phys*. 109, 42 (1998).

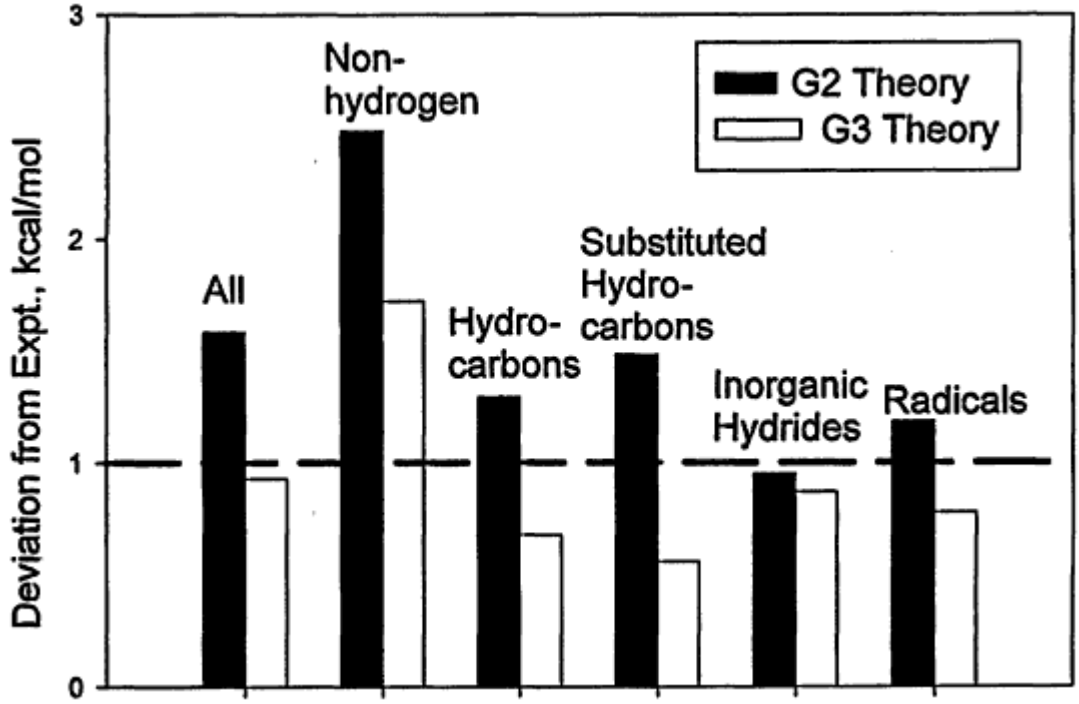


Figure 3.1
Deviation of calculated enthalpies of formation from experiment for the G2/97 test set.

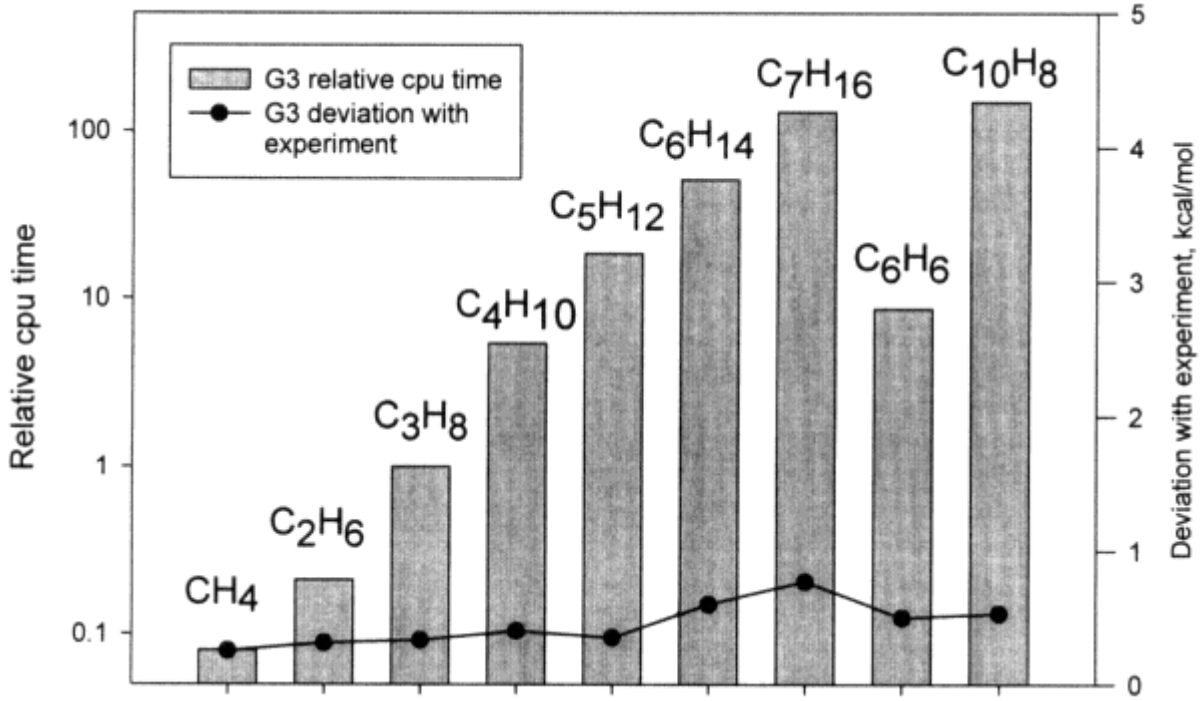


Figure 3.2
Relative cpu times (for a single processor workstation) and accuracy for G3 energy calculations on test molecules containing up to ten carbons.

approaches are summarized in Table 3.1. In the future, thermochemical data for very large molecules will be needed for modeling studies. An example is the simulation of combustion in diesel engines that involves hydrocarbons containing up to 24 carbon atoms. Thus, the scaling problem in computational thermochemistry will have to be addressed in order to make predictions of thermochemical data needed in this area of research. The development of massively parallel computers in the teraflops-speed range, and the software to run on them, will provide computational chemists the opportunity to greatly extend the size of molecules to which the current methods of computational thermochemistry can be applied. However, even this new generation of computers will not be able to handle the very large molecules, so methods to reduce the n^7 scaling will be needed. There are a number of ways by which this might be accomplished: (1) application of scale reduction techniques to coupled cluster and other high-order correlation methods; (2) development of correlation methods that exploit the locality of molecular interactions; (3) modifications to composite methods such as G3 to reduce computational time; and (4) development of density functional methods with improved accuracy.

TABLE 3.1 Accuracy and Applicability of Different Approaches to Computational Thermochemistry[a]

Approach	Example	Accuracy, kcal/mol	Types of Molecules	Size Limit[b]
Direct calculation	CCSD(T) with very large basis sets	0.5	All	2
Composite techniques with molecule-independent empirical parameters	G3 theory	1	All	8
Composite techniques with molecule-dependent empirical parameters	G3 theory combined with isodesmic reactions or bond additivity parameters	0.5	Limited	8

[a] For molecules containing first- and second-row elements. The accuracy and size limits are approximate and are based on current workstation capabilities.
[b] The size limits refer to the number of non-hydrogen atoms and are somewhat dependent on the symmetry of the molecule and number of hydrogen atoms.

Another challenge for the current computational thermochemical methods is remaining problem areas, i.e., failures for certain molecules and lack of the required accuracy for others. Work in the future will focus on determining causes for the failures and how to correct the problems. An important aspect of this work will be the development of expanded test sets of accurate experimental data such as the G2/97 test set. The existence of these test sets will be very important to finding weaknesses of current methodologies as well as developing new and more accurate methods. In addition, contributions to molecular energies that have been neglected as small in the past will be considered in the future in order to attain the required accuracy in larger molecules. These include relativistic effects, nuclear motion, etc. Methods will be developed to account for these effects.

Most of the computational thermochemistry methods that have been developed so far have focused on molecules containing first- and second-row elements. Thermochemical data on compounds containing elements beyond the second row are important but are more difficult to predict accurately because of the increased importance of relativistic and correlation effects; lack of adequate basis sets; and lack of accurate experimental data for assessments. In the future, new developments should provide methods that can handle these more difficult species.

SUMMARY

The thermochemistry of molecules is of major importance in the chemical sciences and is essential to many technologies. It is now possible to compute reliable thermochemical data for a fairly wide variety of molecules through recent advances in theoretical methodology, computer algorithms, and computer power. This paper reviews the current state of the art in computational thermochemistry and describes the Gaussian-n series of methods that have been widely used for thermochemical predictions. The latest in this series, Gaussian-3 theory, gives thermochemical data accurate to 1 kcal/mol with small maximum deviations. Despite the successes, much remains to be done in the future to further develop capabilities for accurate prediction of thermochemical data. Among the challenges will be extension of the methods to larger molecules, increased accuracy in predictions, and extension to heavier elements. The increase in computing power obtainable from new generations of computers, such as those with massively parallel architectures, will play an important role in meeting these challenges.

DISCUSSION

Jack Kay, Drexel University: Could you indicate for which types of molecules the Gaussian methods do the poorest job, and for which types it would do the best job?

John Pople: Well, the poorest ones I showed. They were molecules such as SO_2 and PF_3, generally second-row molecules where we are probably not doing as well as for the first-row ones. The best job is certainly done on the simplest hydrocarbons, which are well known from an empirical point of view, to have great regularities. And if you get a theory that does well on ethane and propane it is going to do well on butane and so forth. That is not surprising. So those are the better examples. Generally, moving toward the right of the periodic table and down in the periodic table makes things more and more difficult.

Jack Kay: How much luck have you had applying that to the metallic elements?

John Pople: Well, the theory is tested against all the known facts, and that includes the molecules like Li_2 and Na_2. It includes completely ionic molecules like lithium chloride and sodium chloride, and completely homoprotic ones like the hydrocarbons. We have, without bias, taken everything that we could find in the literature.

Jack Kay: What about transition-metal compounds?

John Pople: Transition-metal compounds analysis is really just starting. We have developed now the same basis sets for transition metals so we can begin to examine transition-metal compounds. A difficulty there is that the good experimental data is quite limited. Part of the trouble is that doing thermochemistry experimentally went out of fashion some time ago and there is less new material appearing than used to be there. So there is a need for more.

Jack Kay: I was wondering about relativistic corrections and things of that sort.

John Pople: That is another very pertinent question. The theory as I have described it does not include any relativistic corrections apart from the spin-orbit corrections applied to the separated atoms. You can

ask how this changes if relativity is included. And very recently we have generalized the theory to the leading term in the relativistic expansion. You can treat relativity as a perturbation on non-relativity by using the inverse of the speed of light as an expansion parameter and just work out the leading term.

The results are not big but they are not insignificant. The biggest relativistic correction that we have found is for silicon tetrafluoride, which is about 3 kcal/mole. So, if you include relativity you would have to reparameterize these small parameters. But those corrections will get bigger without doubt when we get into transition-metal compounds.

Session 1

Panel Discussion

Richard Hirsh, National Science Foundation: Peter, you discussed the need for new methods, and new methods can take almost any form. They can be new chemistry investigations, new numerical methods, new implementations of the chemistry and the numerical methods to make codes, and so on. I would like to know what you meant by new methods, and beyond that, I would also like to know, given what Paul said about the capability of machines, what new problems those new methods should be or could be applied to.

Peter Taylor: I guess my immediate answer to the list you gave is all of the above. I do not think we should restrict ourselves at this point. But what I was particularly thinking about was not so much reimplementing the existing methodology we have, which I think has a number of disadvantages, some that I discussed and some that were, I think, clear by example in what John presented, but going back to, say, the Schrödinger equation and looking at other strategies for developing approximate solutions of it quite different from what we have now. So I would like to see an effort across the board on that, starting from looking again at the mathematics and looking at different methods for constructing approximate solutions all the way through. We should not exclude looking at the methods we have now and reimplementing them, but that certainly should not be the only thing we do.

And I think one of the key issues here is that most of the effort we have all put in so far toward developing parallel implementations of quantum chemistry has really addressed taking the methodology we have used for a number of years and reformulating this so it runs on parallel hardware. And the scaling of this, as one criterion of how well you do, is good for some methods and good for some implementations but not so good for others. It is not at all clear that the domain of methods we use to try to construct approximate solutions to the Schrödinger equation necessarily includes the most desirable scaling or methods that might perform very well on a machine, say, with 5,000 or 6,000 or even 10,000 processors. I am glad to see, based on Paul's talk, that we are not necessarily thinking of tens of thousands of processors, but it is clear that if we want to use these machines effectively, we have to be able to run on thousands of processors, and that is a lot tougher than getting something going on, say, a 64-way symmetric multi-processor. If we are going to do it effectively we need to look at the largest

possible variety of methods for constructing approximate solutions to the Schrödinger equation, not focus simply on developing, or trying to develop, scalable implementations of what we have now.

Paul Messina: Peter, I remember from a few years ago working with a colleague at Caltech, the chemist Aron Kupperman, that when we were able to provide him a substantially bigger computing environment, he quickly found that the basis set that he was using was, in fact, not adequate, and he had to come up with a more nearly orthogonal basis set to be able to get any type of accuracy. So, would you include in the new methods examination the need to look also at those aspects where now you have a much bigger system and consequently would have to worry more about the capability of the methods?

Peter Taylor: Oh, yes. I think there is no question about that. This is just another aspect of the sorts of improvements we need to try.

Thom Dunning, Pacific Northwest National Laboratory: I think just in general, to make very concrete the kind of suggestion Peter is making, that it has become painfully clear over the past decade how slowly basis-set expansions converge. Yes, we can approach the full solution of the Schrödinger equation with the basis-set technology that we have, but it is a very painful process and it gives rise to some of this very horrible scaling that Peter is talking about. Plus, it gives rise in schemes like what John is talking about, G2 and G3, to some of those large deviations that are observed just because in the basis sets that are used, the convergence is not there for that type of molecule.

John Pople: Yes, there are worries all the time that all the technology we are using, which is based very much on Gaussian functions, may not be the best approach when we come to very large systems and very new technology.

One such possibility to be looked at by chemists is the matter of plane wave expansions that solid state physicists use quite a lot. It is more appropriate for them, perhaps, because a crystal is a periodic system. But nonetheless, one can ask whether you can do better with all the modern techniques of fast Fourier transforms and so forth by taking the molecule in the middle of an empty box and proceed to expand everything in plane waves. That technology could conceivably be a new approach that would possibly eliminate some of the difficulties we have at present.

Evelyn Goldfield, Wayne State University: One of the things that I would like to address was in Peter Taylor's talk when he divided the field into three: electronic structure, reaction dynamics, and molecular dynamics. It seems that one of the hurdles to actually using these codes effectively is that there is a step between electronic structure, when you calculate points, and having a usable potential energy surface. Someone has to laboriously fit a potential, which may or may not be an accurate reflection of the ab initio surface. And then another community, the dynamics community, uses that potential and makes some predictions that may or may not compare to experiment. Among the most challenging things that could be produced are codes that actually integrate these three parts of the problem so that the fitting steps are bypassed. Such efforts are actually a hot topic right now. To deal with realistic and interesting systems is really going to be incredibly computationally intensive and I think could effectively make use of any number of processors.

Peter Taylor: I agree. I think this is a defect in what we currently have—that we do not have enough integration between these steps. While there are, for example, dynamics programs that basically call for the electronic energy or something like the gradient of the potential surface on the fly, we do not have

enough of this. People are put off a bit, even given the sort of hardware we have been talking about, by the dimensionality of it: the fact that if you are doing something with three or four atoms or something similar, this is already fairly expensive; if you have a larger system, where you are treating explicit solvent molecules, say, the idea that you have to evaluate the energy millions of times starts to become a problem, even if you are talking about trying to do this on thousands of processors.

But, to give you an example of what I think of as new methods in this regard, one way, of course, to characterize a potential energy surface is indeed to have millions of energies that you might fit to a functional form, where you might calculate on the fly. Another way to characterize a potential energy surface, though, would be to expand it around a single point. Now, potential surfaces are very complicated functions and so if you were to do, say, a Taylor expansion, the order of derivatives you would need at that point to adequately describe the surface would be very, very high. I would submit, though, that it would be nothing like a millionth-order derivative. It would probably be tens of orders, say. Now, at present we do not know how to do better than fourth-order derivatives of the energy in a way that is more efficient than evaluating individual energies. If we put effort into understanding how to do higher derivatives of the energy efficiently (because we already know that evaluating the derivatives of the energy, at least up through fourth order, is something we can do a lot more efficiently than evaluating lots of individual energies)—if say, we need derivatives through 25th order to characterize an entire potential energy surface and if we could figure out how to calculate the 25th derivative of the energy in an efficient way—then we could do basically one calculation to characterize the whole thing. And that could then be used for the dynamics. That is another example where we need to think about new methods.

John Pople: I think there is an important point here about the interfacing of programs from slightly different areas of chemistry. This is clearly something that is very desirable, and it does bring up the question of the desirability of having source codes public. It is very important, I think, that people should know what the codes do and should be able to carry out some kind of useful interface with other codes. One has to have protection to stop people from getting a code, introducing some bug, and then giving it to somebody else. That clearly would be disastrous. But nonetheless, I think particularly with the development of object-oriented codes we should pay more attention to producing software that can effectively be interfaced with software from some related field so that we can begin to build integrated programs to handle these difficult situations.

Peter Taylor: I absolutely agree.

Paul Messina: Although I certainly did not dwell on this in my presentation, when I was at Caltech and working on the applications that we are trying to tackle, we identified as the biggest challenge the integration of codes from the different disciplines early on. And I think one of the subtexts of ASCI is to be able to do these integrated, multidisciplinary simulations. There are issues of program difficulties that include the code, and certainly of methods—different grading, different approximation techniques—that, if one is not careful, will introduce instabilities in the numerical computation, for example. So that is, indeed, a very important problem that science in general needs to be able to deal with to do the very large simulations.

Judith Hempel, University of California, San Francisco: I am really struck by the fact that we have new challenges in the computing engines and the codes and in the experiments that are needed to validate the codes, but it does seem to me that for a mature field—I think we agree that in some ways the

field is mature—a lot of the impact is in the applications of these theories, and there are new applications that are coming out. So I would like your input.

For example, in the fields of docking and scoring, if you have a biological molecule you can dock many, many different molecules into the binding cleft and score it. And that is a kind of theory in and of itself. So, do you see any particular challenges for the field as a whole, to drop back, as it were, and to simplify in some ways the theory to approach these very large and actually very important problems?

John Pople: I think the main contribution at the present time to that sort of simulation is using these more fundamental methods to work out the energies of interaction of appropriate modeled pieces. We clearly cannot handle the whole systems by these reliable methods but can work at the intermolecular potentials of the pieces, find out something about how configurations may change in the enzymatic environment, and so forth. So, it is really a two-stage step. One level of theory should provide the underlying potentials. Others will then simulate the very large molecular systems using those potentials. So, again, this is an example of integrating the software between different branches of the field.

Judith Hempel: Do you feel that a 1-kilocalorie target is the correct target also for predicting a binding agent?

John Pople: The 1-kilocalorie target is the target that we have for calculating the energy of a bond from scratch, when we do not know the energy and do not know anything about it beforehand. You can do much better than that if you look at the same bond in related circumstances. Even Hartree-Foch theory can work quite well and can be applied, of course, to very large systems, if you do a comparison between the same set of bonds in one molecule and the same set of bonds arranged in a different fashion in another molecule, what we call isodesmic reactions or isodesmic comparisons. An elementary theory can do well there, much better than 1 kilocalorie.

Judith Hempel: Right. These numbers, of course, are affected very much by solvation issues, as Peter pointed out, so there is this integration across many theories.

Peter Taylor: I would also like to address a general point that you raised there. There is a story, I believe told about Wigner, who said that rather than have a very accurate solution to the Schrödinger equation for behavior of electrons in a particular metal, which would necessarily be a very complicated thing, he would sooner have a vivid picture of what the electrons were doing in the metal. I think the two aspects of that statement represent undesirable extremes. A vivid picture is of no use whatsoever if it is a vivid picture of an answer that would be wrong, and having a simple method that lets you get a very nice picture up on the screen that has binding energies that are wrong by 10 kilocalories per mole is really no use to you either. You are not learning anything about nature from that.

This is one of the things Paul emphasized: the need for us to be able to visualize whatever type of calculation we are doing in some way that lets us learn something more about nature from those numbers. I think we learned that the ideal situation would be to have a vivid picture of exactly the right answer. But I would sooner have a vivid picture of nearly the right answer, or actually nearly the right answer and no vivid picture, than I would have a vivid picture of the wrong answer.

Douglas Doren, University of Delaware: It strikes me that the field is, in some ways in terms of our computational needs, at a dividing point. The advances over the last 10 years have made it feasible to get pretty good answers on pretty big systems so that it is now feasible, for example, to calculate the

structure and energetics of some large, inorganic molecules that are actually of a size that a real synthetic chemist would make and be interested in. It is something that we do all the time.

Making something—a method or a machine—that gives us the answer faster is simply going to make these calculations more trivial. There will certainly be some systems where it is important to be able to do very large scale simulations, but the market for those, I think, is getting smaller and smaller. I think for the last 10 years we have been expanding the market for quantum chemistry in general and making it feasible to actually do computational inorganic and organic chemistry in a reasonably useful and reliable way. The market driving the development of methods for extremely large scale solutions, I think, is going to be somewhat smaller. I mean, my synthetic organic chemist friends are not going to start making bigger molecules simply because we are able to calculate them.

There is certainly an important role for these methods. I am interested in them and working on them, but in many ways, fine-grain parallelism and massively parallel systems are going to solve only a subset of all our problems. I will give two examples that I can think of. One, a simple case, is that I have a group of five students. Each student is working on a couple of different problems. They really need separate resources: they need to be able to get all those problems solved all together. Being able to solve a single problem quickly is not their goal. If I could solve each of those problems separately on a workstation, that would be great. Being able to do them sequentially on a massively parallel computer does not necessarily bring my group's set of problems to a solution more quickly.

As another example, I do not think that calculating the 25th derivative of a potential surface at a single point is going to be enough to characterize the whole surface. We are going to need to calculate the potentials and derivatives at lots and lots of points and still have some way of fitting these together. That would work just fine with loosely coupled parallelism.

What are your thoughts? Any sense of where the balance is?

Peter Taylor: What I hope that machines like the ASCI-class machines will do is catalyze development of new methods, even if the immediate application of these to new areas or larger systems or more accurate calculations seems limited. The technology that is developed for handling those problems, the higher accuracy, the larger systems, the greater efficiency, will inevitably fit into the programs that are run on desktop machines and let people do chemistry that almost by definition they are not considering today.

An example of this is that if you go back to the earliest days of Gaussian programs in the early 1970s, the capabilities, in essence, were limited to Hartree-Fock calculations. This is partly, I guess, because of what was feasible at the time, but partly because I think up until the end of the 1960s there was not a very clear understanding of the essential role that electron correlation would play in more accurate predictions. In the mid-1970s, due to the work of various people (some here, like John, the group that I was part of in Australia, Rod Bartlett who is here at the back), more efficient methods of treating electron correlation were developed. Most of the chemists I knew who did calculations at all at that time felt that by and large this was unnecessary and could not see why we were mucking around with something that was only relevant to water or methane or hydrogen fluoride. And yet today no self-respecting chemist, not only quantum chemists, not only theoretical chemists, no self-respecting *chemist* would restrict his or her investigations of something to just the Hartree-Fock level. So, because now we do not see an immediate need for daily use of 5,000 processors and another order-of-magnitude accuracy in our application calculations, I do not think we should conclude from that that the market for these sorts of things, necessarily, is dwindling. I would say it is the technology that comes out of doing those calculations and the way it feeds into what people will use on the desktop in 5 or 10 years time that is the really critical thing.

Thom Dunning: I think also there are applications, as I think Paul could point out in the ASCI program itself, where if you tackle a particular problem it turns out that many of these problems are very data hungry. The combustion problem needs a lot of information on a lot of molecules in a lot of reactions so that you are faithfully representing the chemistry that is going on in a combustion process.

You could say we are going to wait until all those 1,000 species and all those 2,000 reactions are characterized in the laboratory. But, in fact, if you do that you are going to be sitting forever. And the only way to get at some of that data, in fact, is to do it computationally, and you may have to be doing a lot of calculations at a relatively high accuracy that will actually require the kinds of machinery being requested in ASCI. And that is the real driver. That is why those machines are necessary to solve their problem; they have such a complicated system that to describe the various components of that complicated system really represents computational grand challenges that require generating a lot of data that is just not available from experiment. Of course, validating against the experiment, as John said, is done whenever possible, but there are species like radicals and ions that can be very difficult to study in the laboratory, where information on them is absolutely critical to the fidelity of representing that particular physical or chemical process in the total system itself.

David Dixon, Pacific Northwest National Laboratory: We have really not talked much about nuclear motion and its problem, one of the things I think that does argue for looking at much larger architectures. It is much more straightforward to do things in zero Kelvin. If you want to look at chemistry at 298 or 500 degrees, you are after nuclear motion. If you want to treat the water dimer correctly you have to put quantum nuclear motion in, and I think that as we start to look at what we can do with the large architectures, we will actually start thinking about how we treat the nuclear motion problem, how we treat coupled routers, how we treat weakly bound systems. This gets back to the enzyme interaction. That is going to require much larger computer resources than we have today where we are working on single, accurate points.

If you do not get zero point energy at zero Kelvin in methane right, if you just take the zero point energy and cap it with what is known experimentally, you are off by 0.6 or 0.7 kcal, which is almost all of John's error. So the nuclear motion part of it will be critical to really solve, and I think this is going to be one of the arguments we need to have to go to much larger architectures to really understand what is going on. I would appreciate your comments on that.

John Pople: Yes, I agree that is a major feature. We are certainly well aware that anharmonicity is one of the things that is not well treated in the present level of theory. It is swallowed up by one of the empirical parameters. And, indeed, one knows when you come to molecules that are floppy—and most molecules, for example in biological circumstances, are very floppy—where all sorts of rigging around is going on, this is a very important feature. So I fully agree with you that one has somehow to merge the methods that are used for electronic structure with those that are used for handling nuclear motions, and this comes back to the dynamics problem again. There is a lot to be done in interfacing these areas and developing composite programs to handle the whole problem.

Andrew White, Los Alamos National Laboratory: What Paul has talked about, what quantum chemistry has pointed out, are predictive models, whether it is thermochemistry or the safety of an aging stockpile. It seems to me that in your five-step plan you need a sixth that somehow quantifies the uncertainty in the systems if you are really looking into the future—maybe some place you cannot do experiments, like stockpile stewardship or climate or nuclear winter or weather or whatever. Can you talk about what the state of the art is with Gaussian or with any other code relative to how you quantify

this uncertainty? Maybe there some lessons for the rest of us in how to put all this together for predictive models?

John Pople: Well, the uncertainty question is one that we tried to address by validating the method against all the known facts. Now, if you try to do it purely ab initio, then it is rather hard to give an uncertainty. So in principle, but not in practice, you can get upper and lower bounds to the energy to which you could eventually close. Bright Wilson used to be an advocate of that, but there are no signs of that becoming practical.

So, the best that we have been able to do, and I do not think anybody else has come up with an alternative suggestion, is to test a theory fully against everything that is known really well from experimental chemistry, and to do statistics on that. Those are the numbers that we have come up with. I think it is going to be very useful to look at the bad cases and say there is something wrong with our theory and it is showing up because these results are bad; that is a useful form of investigation.

But, that is the best best way that I can think of to describe an uncertainty.

Peter Taylor: Yes, one can perhaps quantify this a bit. The typical total energies of the sorts of molecules in John and Larry's schemes are of the order of hundreds of thousands of kilocalories per mole and one is looking for 1-kilocalorie-per-mole accuracy there. We typically do not do calculations of total energies that are accurate even to tens of kilocalories per mole, and I would claim that we have really no methodology for which it is practical currently to do total energies accurate to 1 kilocalorie per mole.

There is a very large compensation for error in what we do. We have always known about this. We understand where it comes from in our case—that is, formation of a molecule is a relatively small perturbation on a group of atoms and the systematic errors in the atoms cancel out to a significant degree when you form the molecule. But the field is aware of this, and if we really wanted to have hard, small uncertainties, whether it is going the route of calculating upper and lower bounds or whatever, we would have to work very much harder in the calculation of the total energy itself than we are currently set up to do.

Richard Kayser, National Institute of Standards and Technology: I wanted to point out that we have an effort under way to put together what we call a computational chemistry benchmark database. Right now the database contains about 600 compounds for which we believe we have a good handle on the thermochemistry. In addition to that information, the database will contain the results of many different well-defined calculations based on different levels of theory and different basis sets, with the goal of trying to get a handle on the systematic errors that are inherent in different approaches.

Edwin Wilson, University of Akron: One of the things going on at our institution is calculation of conformations and interactions of polymers. What kind of efforts should be happening in that field and how will that interact with the ASCI program?

John Pople: Well, the problems of conformations on molecular geometries were already fairly well handled many years ago at the Hartree-Fock level. It was found that for organic molecules, even if you used a moderate basis like 6-31, Hartree-Fock theory normally gave you the right conformation of individual molecules. So I think there has been fair success with quite elementary theory in dealing with that.

There needs to be refinement along the same lines, but I think the point to make is that conforma

tional problems are somewhat simpler than getting the total energy of a bond. But undoubtedly further work is needed.

Peter Taylor: I would have to be a little less sanguine than my colleague is, I think. If you look at large molecules like polymers, one of the things they can do that the sorts of half-dozen carbon chain molecules in small-molecule organic chemistry cannot is that they can fold around on themselves. And one of the things that is key in some of this folding around on itself is relatively weak interactions, say dispersion-type interactions between different parts of the chain. Those dispersion-type interactions are not treated at the Hartree-Fock level at all, and so I think something significantly better than Hartree-Fock may ultimately be required to do reliable predictions of conformations of longer chains where there is substantial folding or coiling or things like this, assuming you want to do something that goes beyond the level of some parametrized model, some empirical type of force field.

John Pople: Yes, I completely agree with that. My points were referring to things like rotation about single bonds and boat-chair conformations and cyclohexane, which were handled fairly well some years ago.

Edwin Wilson: One also finds that the interaction that occurs at interfaces between different polymers is a fairly important aspect of that area of chemistry.

John Pople: Yes, it is true that there is somewhat of a division among quantum chemists, those who work on individual molecules and those who work on intermolecular forces, and they sometimes do not quite meet. People interested in intermolecular forces tend to look at the long-range limits and they do not know how to join with the people who deal with the strong interactions at closer ranges.

Peter Taylor: Another integration issue!

Jean Futrell, Pacific Northwest National Laboratory: I am not sure whether this is a showstopper or not, but I would like to inquire about the accuracy of present methods for defining transition states and their energetic properties, frequencies, and so on, and what one can expect from this leap forward in technology.

John Pople: One difficulty is that my suggestion that we test theories by comparing with results known experimentally to great accuracy does not really hold for transition structures. We would, indeed, like to do that, but only one or two energies of activation in the literature are really well known, and those energies are very difficult to reproduce theoretically anyway. So, I can only say that is a more difficult problem and we have fewer means of testing the reliability of our results. This is the best that we can do and we hope this is accurate to this level, but we cannot at present completely test it.

Peter Taylor: I would say this is an excellent example of an area where theory needs to do more to meet the experimentalists on their own turf rather than stopping halfway. “Experimental estimates” of barrier heights are derived from all sorts of assumptions made in the interpretation of experiments. Such assumptions may or may not be warranted and may or may not mean what theoreticians think they mean when they are arriving at their own barrier heights. A far more satisfactory way to deal with the issue of reaction mechanisms, and ultimately kinetics, is for the computational chemist to calculate—this follows, really, from Evie Goldfield's point earlier—the rates of the reaction and compare the results

directly with some kinetics from experiments that is not subject to all the varied assumptions made in order to obtain some sort of barrier height.

I think we should not get hung up on the issue of how to calibrate methods for calculating transition states and how to get error bars on the heights of barriers. We should go the next step further, integrate the dynamics into the calculation and then compute what the experimentalists actually measure and compare that with the experiment. That is the way to get reliable calibrations.

Thom Dunning: While there have not been many calculations, there have been some that would certainly indicate that the techniques that we currently have available to us can achieve very high accuracy. It does turn out that calculations of transition states are significantly more challenging than calculations of stable species. The basis sets you have to use are larger and you have to go much closer to convergence to get reliable numbers. But I would say that the best techniques can get errors on the order of tenths of a kilocalorie per mole. But for the few systems that have been checked, the problem there is, as John says, that we really do not have any good information from experiment that pertains directly to what we are calculating. We have to go to the step that Peter is talking about to be able to compare with experiments.

4

The Role of Computational Biology in the Genomics Revolution

Jeffrey Skolnick,
Jacqueline Fetrow,
Angel R. Ortiz,
and
Andrzej Kolinski
Scripps Research Institute

ABSTRACT

The various genome sequencing projects are providing a plethora of protein sequence information, but with no information about protein structure or function. The most effective method for sifting out useful proteins from these genomic databases is the computer prediction of protein function. However, current methods, which are mainly sequence-based, are limited by the extent of similarity between sequences of unknown and known function; they increasingly fail as the sequence identity diverges into and beyond the twilight zone of sequence identity. In practice, between 30 and 60 percent of all proteins can be functionally identified using current sequence-based software. To extend the level of molecular function annotation to a broader class of protein sequences, methods for identification of protein function based directly on the sequence-to-structure-to-function paradigm will need to be developed. One such approach is presented. The idea is to predict the native structure first by using ab initio folding or threading techniques and then to identify its molecular or biochemical function by matching the active site in the predicted protein structure to that in a protein of known function. Application of this approach to genomic screening is then described. Based on these preliminary results, the next 5 to 10 years are likely to see the development of computational tools that will allow for the medium-resolution prediction of the tertiary structure of single domain proteins, the more robust identification of protein ligands, techniques to predict proteins having specific quaternary interactions, and the beginnings of a bottom-up approach to identify important proteins in metabolic and signal transduction pathways.

INTRODUCTION

The various genome sequencing projects are providing a vast quantity of protein sequence data,[1] but what is needed is information about protein function (Rastan and Beeley, 1997). To enhance the efficiency of the drug design process, one must identify the sequences of functionally important proteins

[1] See the GenBank index at <http://www.ncbi.nlm.nih.gov/Web/Genbank/index.html>.

that are hidden in these large databases. For example, microbial genomes contain potential protein targets that can be utilized to kill pathogens or that can be developed into commercially useful enzymes to produce or degrade various substances. By far the most effective method for sifting out useful proteins from these genomic databases relies on the computer-based prediction of protein function (Rastan and Beeley, 1997). However, most current methods, being mainly sequence-based, are limited by the extent of sequence similarity between sequences of unknown and known function (Pearson and Lipman, 1988; Henikoff and Henikoff, 1991; Attwood and Beck, 1994; Bairoch, Bucher et al., 1995; Altschul, Madden et al., 1997; Attwood, Beck et al., 1997). They increasingly fail as the sequence identity between two proteins crosses into and beyond the twilight zone of sequence identity, which is about 30 percent (Fetrow and Skolnick, 1998). In practice, current sequence-based software can identify the molecular or biochemical function of roughly 30 to 60 percent of all proteins in a given genome (Bult, White et al., 1996; Casari, Ouzounis et al., 1996). The full annotation of entire genomes is likely to be a major computational and experimental challenge over the next 5 to 10 years, but one which, when successfully addressed, will provide a revolution in disease diagnosis and treatment as well as in our conceptual understanding of biology. To be fully successful, this will require a multidisciplinary approach involving biology, chemistry, physics, and computer science.

Here, we describe one promising means of extending the ability to annotate the remaining orphan sequences based on the sequence-to-structure-to-function paradigm (Fetrow, Godzik et al., 1998; Fetrow and Skolnick, 1998). Logically, this process can be divided into two parts. First, one employs techniques to determine protein structure from sequence (Godzik, Skolnick et al., 1992; Ortiz, Kolinski et al., 1998 a,b,c). Secondly, one employs tools for function prediction based on the identification of active sites in the predicted or experimental structure. The ability to determine function from structure will be very important given the emerging structural genomics initiatives where the goal is to determine all possible protein folds. This reverses the more traditional approach where one first identifies the function of the protein of interest and then subsequently determines its structure.

PREDICTION OF PROTEIN STRUCTURE FROM SEQUENCE

Currently, there exist two basic theoretical approaches for the prediction of protein structure from sequence when homology modeling (which requires significant sequence identity between the probe sequence and its template structure) (Sali and Blundell, 1993) cannot be applied: threading (Bryant and Lawrence, 1993; Miller, Jones et al., 1996), and ab initio folding (Skolnick, Kolinski et al., 1997; Ortiz, Kolinski et al., 1998 a,b,c). In threading, the idea is to match the sequence of interest to a template structure in a library of known structures (Godzik, Kolinski et al., 1993); thus, this approach is conceptually similar to standard homology modeling, except that now the goal is to match probe sequences to template structures when there is no apparent sequence relationship between the two. In ab initio folding, one attempts to fold a protein starting from a random conformation (Kolinski and Skolnick, 1996). The advantage of threading is its speed and the fact that it can be applied to large proteins. In contrast, ab initio folding is computationally more demanding and is, in practice, currently limited to proteins smaller than 100 residues (Ortiz, Kolinski et al., 1998 a,b,c). However, ab initio folding does not demand that an example of a native structure be already solved. Thus, it can be used to identify proteins having a novel native structure. Recent results indicate that for small proteins (those less than 100 residues), ab initio folding approaches can predict structures at a level of quality (4- to 6-Å coordinate root mean square deviation for the backbone atoms) comparable to that provided by threading (Ortiz, Kolinski et al., 1998a,b).

Description of Ab Initio Protein Folding Methodology

In what follows, we describe a newly developed method for structure prediction, MONSSTER, which attempts to address the aforementioned problems. As depicted in Figure 4.1, prediction of protein structure can be conceptually divided into four stages: (1) restraint derivation; (2) structure assembly; and (3) selection of the native conformation. In addition, for those sequences whose structures are known either before or after the prediction is made, following the structure selection process, (4) objective, rigorous validation criteria are applied to judge the success of the prediction.

For (1), restraint derivation, a multiple sequence alignment with the sequence of interest is generated (Sander and Schneider, 1991). Then, predicted secondary structure restraints are obtained from a standard secondary structure prediction scheme (Rost and Sander, 1993; Rost, Schneider et al., 1993) supplemented by our LINKER algorithm (Kolinski, Skolnick et al., 1997)—a quite accurate technique for predicting where the chain reverses global direction. We term such regions "U-turns" (Kolinski, Skolnick et al., 1997). The predicted secondary structural elements between these U-turns define the predicted core regions of the molecule. Tertiary contacts (restraints), termed "seeds," between these core elements are then predicted from multiple sequence alignments. Multiple sequence information is used to derive such seed side-chain contacts based on patterns of residue conservation (Aszodi, Gradwell et al., 1995; Mumenthaler and Braun, 1995) or residue covariation in a set of homologous sequences

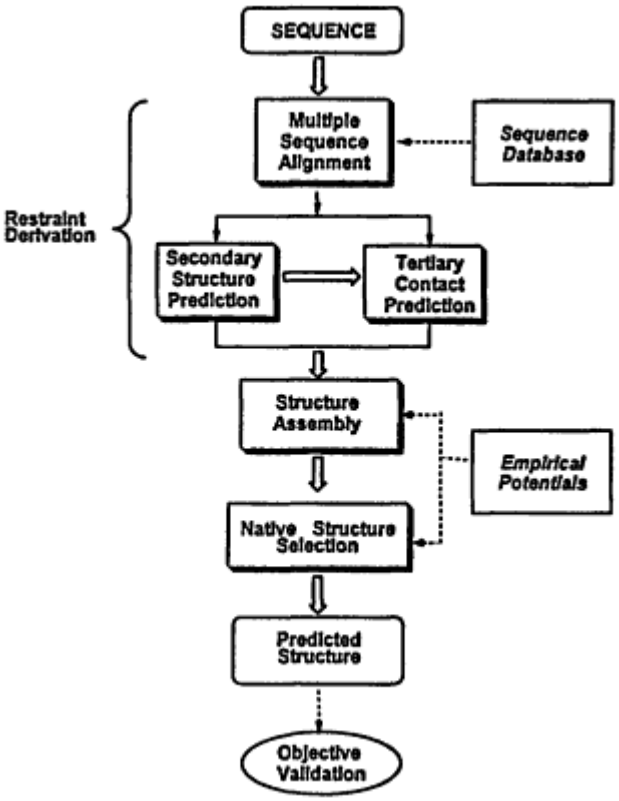


Figure 4.1
Schematic overview of the procedure for tertiary structure prediction.

(Göbel, Sander et al., 1994; Thomas, Cesari et al., 1996; Olmea and Valencia, 1997). Both might be combined for increased sensitivity (Olmea and Valencia, 1997). Here, for the sake of simplicity, we slightly modify the approach of Göbel and coworkers (Göbel, Sander et al., 1994) and calculate the covariation between all residues predicted to be in the putative core of the molecule (Olmea and Valencia, 1997; Ortiz, Kolinski et al., 1998a,b). Unfortunately, there are too few of these seed contacts to assemble a protein from the unfolded state using MONSSTER. Thus, these seed contacts between predicted topological elements (i.e., α-helices and β-strands between U-turns) are enriched by an inverse folding approach that typically produces about N/4 contacts—the number required for successful topology assembly (Olmea and Valencia, 1997; Skolnick, Kolinski et al., 1997; Ortiz, Kolinski et al., 1998a,b).

In (2), the structure assembly step, the set of predicted restraints is used in the MONSSTER method (Skolnick, Kolinski et al., 1997) to drive the conformational search. This uses a reduced-protein-lattice model to assemble the global fold. First, a series of up to 1,000 independent, simulated annealing structure-assembly runs are performed, and the resulting structures are clustered on the basis of their pairwise coordinate root mean square deviation (cRMSD). If the resulting structures do not cluster into several topologies, then no structural prediction is made. If at least a subset of the structures cluster, then we proceed to the structure selection step.

The native structure selection stage (3) consists of long isothermal runs from which the putative native topology is chosen on the basis that it has the lowest average energy. If the differing topologies cannot be selected on this basis, then the prediction consists of several lowest-average-energy representatives of the various generated topologies. In all cases, we report the average cRMSD values corresponding to the lowest-average-energy structures and not the best cRMSD values because in a blind prediction, we would have no means of selecting such structures.

Once the native conformation of the protein of interest is known, we judge the success of the prediction (4) as follows: First, we calculate the global C-α cRMSD between the predicted (lowest average energy) and experimental structures. Since our approach often results in structures whose C-α cRMSD is in the range of 6 Å, there may be substantial topological errors between the native and predicted structure; therefore, a more rigorous assessment of success is necessary. Thus, the predicted fold is subjected to a structural similarity search over a representative database using the DALI (Holm and Sander, 1997) structural superimposition program. We note that a very similar approach has been used to assess the quality of structures predicted by threading techniques, where a sequence is matched to a fold in a library of known structures (Wodak and Rooman, 1993). When a known homologue of the native structure is chosen (or the native structure itself), then the tertiary structural prediction protocol is considered to be successful. If two or more topologies are isoenergetic, both would be subjected to this protocol; if one matches the native topology, we consider this to be a partial success. If the next lowest average energy topology (as predicted by MONSSTER) matches the native fold rather than the lowest average energy structure, this is also considered to be a partial success. Otherwise, by this rigorous criterion, the prediction is unsuccessful.

Validation on Proteins of Known Structure

The above protocol was applied to the set of 19 proteins listed in Table 4.1. On average, for the set of proteins whose native conformation was known in advance, the predicted secondary structure is 69 percent correct; this is slightly less than the reported average for this technique, which is 72 ± 9 percent (Rost and Sander, 1993; Rost, Schneider et al., 1993). Such a large test set is necessary to demonstrate that the current approach can handle a wide variety of folds and different secondary structure types. All

are outside the set of proteins employed in the derivation of the empirical potentials. It is very important to emphasize that all predictions use the identical parameter set and folding protocol. Table 4.2 shows the accuracy of the predicted secondary structure and tertiary contacts, as well as the results from the folding simulations. Only about 78 percent of the native contacts are correct within ±2 residues; these are typical of results seen on an even larger class of proteins. Often, there are also a number of grossly incorrect restraints that can lead to non-native topologies. Using this information, in about 10 to 30 percent of the assembly runs, native-like topologies, as subsequently assessed by their global cRMSD

TABLE 4.1 List of Proteins of Known Structure That Constitute the Validation Set

Protein	Nres	Class	Fold Description	Name
3cti	29	small	disulfide-bound fold, beta hairpin with adjacent disulfide	Trypsin inhibitor from squash (*Cucurbita maxima*)
1ixa	39	small	EGF-like (disulfide-rich fold; nearly all beta)	Factor IX from human (*Homo sapiens*)
protA	47	α	Three-helix bundle	Protein A
1gpt	47	small	disulfide-bound fold, beta hairpin with adjacent disulfide	Gamma-thionine from barley (*Hordeum vulgare*)
1tfi	50	small	Rubredoxin-like (metal-bound fold, with 2 CXXC motifs)	Transcriptional factor SII from human (*Homo sapiens*)
6pti	58	small	BPTI-like (disulfide-rich α+β fold)	Pancreatic trypsin inhibitor from bovine (*Bos taurus*)
1fas	61	small	Snake toxin-like (disulfide rich; nearly all beta)	Fasciculin from green mamba (*Dendroaspis angusticeps*)
1shg	62	β	SH3-1ike barrel (partly opened; n* = 4, S* = 8; meander)	alpha-Spectrin, SH3 domain from chicken (*Gallus gallus*)
1cis	66	α+β	CI-2 family (α+β sandwich; loop across free side of β)	Hybrid protein from barley (*Hordeum vulgare*) hiproly strain
1ftz	70	α	DNA-binding 3-helix bundle (right-handed twist; up-down)	Fushi Tarazu protein from fruit fly (*Drosophila melanogaster*)
1pou	71	α	DNA-binding domain (4 helices, folded leaf, closed)	Oct-I POU-specific domain from human (*Homo sapiens*)
1c5a	73	α	Anaphylotoxins (4 helices; irreg. array, disulfide linked)	C5a anaphylotoxin from pig (*Sus scrofa domestica*)
3icb	75	α	EF-hand (2 EF-hand connected with Ca bind loop)	Calbindin D9K from bovine (*Bos taurus*)
1ubi	76	α+β	β-grasp (single-helix packs against β-sheet)	Ubiquitin from human (*Homo sapiens*)
1lea	84	α	DNA-binding 3-helix bundle (right-handed twist; up-down)	LexA repressor, DNA-binding domain (*Escherichia coli*)
1ego	85	α/β	Thioredoxin-like (3 α/β/α layers; β-sheet order 4312)	Glutaredoxin from bacteriophage t4
1hmd	85	α	Four helical up-and-down bundle(left-handed twist)	Hemerythrin from sipunculid worm (*Themiste dyscrita*)
1poh	85	α+β	α+β sandwich	Histidine-containing phosphocarrier proteins (*Escherichia coli*)
life	100	α+β	IF3-1ike (β-α-β—α-β(2); 2 layers; mixed sheet 1243)	Translation initiation factor IF3 from *Escherichia coli*

and DALI (Holm and Sander, 1997), are recovered for all classes of proteins. But on average, helical proteins are predicted better than alpha/beta proteins, which are predicted better than beta proteins.

In 14 of 19 cases, success or partial success was obtained, with the lowest average C-α cRMSD values ranging from 3.5 to 6.7 Å. For one partial success whose lowest average energy structure has a higher cRMSD from native life, the lowest-energy fold basically adopts the native topology despite its unsatisfactory cRMSD. Here, a strand region is found at the back of the protein rather than at the edge of the fold. The topology with the next higher energy, i.e., the first excited state, is the native one. For the five unsuccessful cases—3cti, 1tfi, 6pti, 1lea, and 1poh—DALI fails to find any structure that is significantly related to the lattice model; thus the prediction is labeled as being unsuccessful. This is true in spite of the fact that the topology of 6pti is native, and for 1poh, a slightly misfolded state is recovered, but by the DALI selection criterion, this simulation is unsuccessful. Furthermore, for mainly helical proteins such as lpou, the alternative low-energy fold is the topological mirror image (where the helices are right-handed, but the chirality of the turns is reversed from the native conformation). In some situations, e.g., for life and 1poh, the alternative topology differs in the placement of one or two topological elements. In other cases, the alternative and native topology do not have much in common.

Blind Predictions

We next present a representative prediction of the tertiary structure of the 81-residue KIX domain of the CREB binding protein, which is involved in gene expression as mediated by AMPc (Brindle and Montminy, 1992; Radhakrishnan, Perez-Alvarado et al., 1997).

As shown in Figure 4.2, the secondary structure prediction scheme suggests that KIX should adopt a three-helix bundle fold. Correlated mutation analysis provides four seed contacts (22-35, 22-73, 35-73, 17-72) that yielded 38 predicted tertiary contacts when enriched; this is a rather large number as compared to other entries in Table 4.2. A series of 10 independent fold assembly simulations were done; all yielded either a left- or right-handed three-helix bundle. As indicated in Table 4.2, on the basis of their average energies, the two topologies are essentially isoenergetic. Decomposing the energy into its constituent contributions (Ortiz, Kolinski et al., 1998c), the pair interactions, secondary-structure preferences, and hydrogen-bond terms favor the right-handed bundle, whereas the burial energy and terms designed to generate protein-like densities favor the left-handed bundle. The difficulty in distinguishing topological mirror images is a problem that this method often experiences with helical proteins, and indicates that improvements in the empirical potential are necessary. When subsequent predictions were done using the subset of restraints that satisfy each of the two topologies, then the native topology was found to be substantially lower in energy than the incorrect alternative.

STRUCTURE TO FUNCTION

In the prediction of protein function from sequence, there are a number of key questions that must be answered. In particular, does one need a protein structure to predict protein function or is sequence information sufficient? If a protein's tertiary structure is needed, how close does it have to be to the native state to permit the protein's function to be identified? Is there a one-to-one relationship between protein structure and protein function? If not, can one construct a library of active sites so that one can search structures for appropriate active sites? In what follows, we address each of these questions in turn.

TABLE 4.2 Summary of Prediction Results

Protein[a]	Type	N[b]	Q_3[c]	N_c[d]	N_p[e]	N_w[f]	d=0[g]	d=2[g]
Proteins with Known Structure in Advance of Prediction								
3cti	small	29	50.5	39	6	0	83.3	100.0
1ixa	small	39	70.5	48	5	0	100.0	100.0
1gpt	small	47	70.6	70	13	0	46.1	100.0
1tfi	small	50	60.0	84	37	0	21.6	88.8
protA[o]	α	47	77.6	91	17	0	0	70.5
1ftz	α	56	63.5	149	12	1	25.0	58.3
1c5a	α	66	85.8	105	43	1	24.4	73.3
1pou	α	71	78.8	122	49	0	28.6	89.8
3icb	α	75	82.3	154	25	0	28.0	68.0
1hmd	α	85	90.6	157	20	2	10.0	65.0
1shg	β	57	67.1	109	39	0	28.2	100.0
1fas	β	61	67.1	98	25	1	26.3	78.9
6pti	αβ	56	58.8	92	19	0	68.4	100.0
1cis	αβ	66	64.7	144	23	0	8.6	78.2
1lea	αβ	73	63.5	131	41	2	9.7	75.6
1ubi	αβ	76	62.3	153	17	0	23.5	94.1
1poh	αβ	85	65.9	162	36	3	8.3	55.5
1ego	αβ	85	60.0	223	33	0	15.1	93.9
1ife	αβ	100	75.3	148	21	3	14.2	38.0
Result from Blind Prediction								
KIX	α	81	87.7	320	37	11	26.3	57.9

NOTES:
[a] Protein refers to the Brookhaven National Laboratory's Protein Database (PDB) access number for the protein studied.
[b] N is the number of residues in the protein in the PDB file.
[c] Q_3 is the percent of correctly predicted secondary structure. All proteins have a Q3 within one standard deviation of the average.
[d] N_c is the number of contacts in the native structure.
[e] N_p is the number of predicted contacts.
[f] N_w is the number of contacts that are incorrect when no native contact is found within ±5 residues of a predicted contact.
[g] Percent of predicted contacts within d residues of a native contact.
[h] rms_n is the cRMSD deviation in angstroms from the native structure.
[i] E_n is the lowest average energy (in kT) after refinement for the nativelike topology
[j] rs_n is the number of restraints satisfied in the nativelike topology.
[k] rms_w is the cRMSD deviation from native in angstroms of the alternative topology of lowest energy.
[l] E_w is the lowest average energy (in kT) in the alternative topology after refinement runs.
[m] rs_w is the number of restraints satisfied in the alternative topology.
[n] Relationship of lowest average energy structure to the native conformation if known. S indicates that the full structural selection criterion as assessed by the energy and DALI are "successful," PS indicates that the tertiary structure prediction is "partially successful," and U indicates that the tertiary structure prediction is "unsuccessful."
[o] The B domain of protein A.

Native Topology			Lowest Energy, Nonnative Topology			
rms_n[h]	E_n[i]	rs_n[j]	rms_w[k]	E_w[l]	rs_w[m]	Final Score[n]
3.8	-107	6	6.7	-103	6	U
5.6	-130	5	7.7	-131	5	S
5.9	-276	9	6.6	-142	10	S
5.9	-202	28	7.0	-191	31	U
3.1	-246	2	9.4	-240	10	S
5.1	-277	11	10.1	-270	15	S
4.2	-194	20	9.8	-182	26	S
3.5	-418	18	11.9	-364	22	S
4.5	-406	21	12.6	-342	11	S
4.6	-458	3	9.3	460	13	PS
4.5	-420	19	6.7	-397	18	S
6.2	-330	19	9.37	-284	20	S
4.7	-410	19	9.7	-397	18	U
6.4	-240	7	7.6	-232	7	S
6.1	-136	26	9.4	-115	27	U
6.1	-238	9	11.5	-203	8	S
6.5	-336	42	11.7	-299	23	U
5.7	-417	20	9.0	-396	16	S
6.7	-419	15	8.2	-482	16	PS
5.8	-477	19	10.7	-479	26	PS

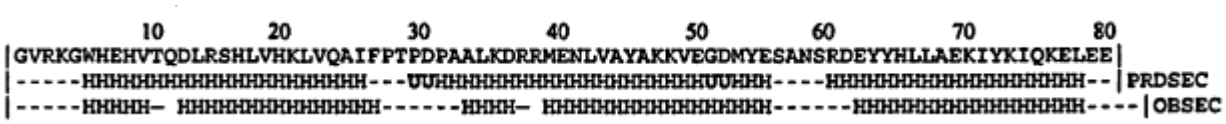


Figure 4.2
For KIX, the primary sequence and a comparison of the predicted and observed secondary structure. Here, H denotes a helix, U a U-turn; PRDSEC (OBSEC) is the predicted (observed) secondary structure from PHD and LINKER.

Limitations of Sequence-based Methods

As residue identity falls into the twilight zone, standard sequence-alignment algorithms will pick up false positive sequences as well as miss false negative sequences. Similarly, as sequence diversity increases, the local sequence signatures found in the Prosite (Bairoch, 1990; Bairoch, Bucher et al., 1995), Blocks (Henikoff and Henikoff, 1991), and Prints (Attwood and Beck, 1994; Attwood, Beck et al., 1997) databases will no longer be strong enough to recognize protein sequences as belonging to a functional family, even though the specific active site residues might be strictly conserved. (See Table 4.3.) To illustrate this inability to recognize local sequence signatures as the sequences diverge, we performed an analysis of the Prosite database (Release 13.0, November 1995). Of 1,152 patterns in this release of Prosite, 908 (79 percent) of the patterns were absolutely specific for their sequences (using the set of true and false positives and negatives as identified by the Prosite developers). However, as the number of instances of a local pattern increases, the number of false positives also tends to increase. For 10.5 percent of the patterns, 90 to 99 percent of the selected sequences were true positives, while for the remaining 10.5 percent of the patterns, fewer than 90 percent of the selected sequences were true positives. To overcome this deficiency, the developers of the Prosite database have begun to use weight matrices or profiles for detection of domains. Unlike the typical Prosite, Blocks, and Prints methods, they create profiles of sequence information such as residue type and solvent accessibility (Gribskov, McLachlan et al., 1987) based on the complete protein sequence, not just a small segment. As with domain-matching methods, problems inherent in matching highly divergent parts of the sequence, as well as the highly conserved functional regions, still reassert themselves.

Similarity of Global Tertiary Structure Does Not Always Imply Similarity of Function

In principle, additional information might be provided by comparing the complete tertiary structures of proteins; however, comparison of overall structure is also not enough to classify protein function unequivocally. The structural databases such as SCOP (Murzin, Brenner et al., 1995), CATH, and DALI (Holm and Sander, 1997) show significant redundancy in domain structures. Proteins such as the barrels and the sandwiches can exhibit very similar structures even though they have very different functions. Valuable information can be obtained from overall tertiary structure comparison (Murzin, 1996), but two proteins with the same global tertiary structure do not necessarily have the same function.

Proteins with Similar Function Conserve the Local Structure Around the Active Site, Even If the Global Fold Is Dissimilar

As the families become more diverse, the sequence similarity among many proteins in the family falls into and below the twilight zone. Then, standard sequence alignments have difficulty establishing a significant relationship between sequences even though one might exist. For example, the mammalian and bacterial serine proteases demonstrate that proteins with very similar functions can have very different three-dimensional structures (Branden and Tooze, 1991). The geometry of the active site would not be recognized by local sequence signatures or by overall comparison of global tertiary structures, but only from an analysis of the structure of the functional residues around the active site.

TABLE 4.3 Data for Classification of Possible Thioredoxin Sequences by the Prosite, Prints, and Blocks Algorithms

	Prosite[a,b]			Prints[b,c]			Blocks[b,d]		
	TP	FP	FN	TP	FP	FN	TP	FP	FN
Possible Thioredoxins									
DSBC_HAEIN			X			X	X		
THIO_CHLLT			X	X(2)[e]			X		
THIO_CHRVI			X	X			X		
THIO_RHORU			X	X					X
YX09_MYCTU			X	X					X
Y039_MYCTU			X			X			X
YB59_HAEIN			X			X	X		
May Be Thioredoxins (treated here as if they are thioredoxins)									
BS2_TRYBB	X			X			X		
FIXW_RHILE	X					X	X		
GSBP_CHICK	X			X			X		
RESA_BACSU	X			X(2)[e]			X		
YME3_THIFF [f]			X			X	X		
Probably Not Thioredoxins									
YNC4_CAEEL		X							
POLG_PVYC		X							
POLG_PVYN		X							
POLG_PVYHU		X							
POLG_PVYO		X							

[a] Prosite: recent Prosite database online (thioredoxin examples updated 9/10/97).
[b] TP = true positives; FP = false positives; FN = false negatives.
[c] Prints: search of OWL26.0 database.
[d] Blocks: search of SwissProt32.
[e] Prints uses three different sequence signatures to recognize the thioredoxins. "2" means that this sequence was recognized by only two of the three signatures.
[f] A plasmid in *E. coli* expressing this gene product complements a thioredoxin mutant, providing experimental evidence that this protein may be a glutaredoxin or thioredoxin.

Development of a Three-dimensional Library of Functional Motifs

What these examples suggest is that one might be able to excise the local structure around the active site and use this local conformational signature to identify function. In fact, proteins function because of the arrangement of specific residues in three-dimensional space, The residues involved in protein function, particularly those at enzyme active sites, will be highly conserved throughout evolution. This statement seems obvious and it was clearly demonstrated experimentally by the serine protease presented above. The problem with recognizing these residues by sequence alignment is that they are likely to be distant along the sequence, even if they are close together in three-dimensional space. This makes recognition by multiple-sequence-alignment methods problematic. If protein function relies on the

specific tertiary placement of residues, then one should use that geometric information to describe functional families. We term these geometric (e.g., distances and angles) and conformational (e.g., a residue must be in a helix) descriptors "fuzzy functional forms" (FFFs). These methods do not rely on evolutionary conservation of local sequence as do the local sequence signature methods, but instead involve the construction of three-dimensional descriptors of protein function.

There are several distinct advantages to using geometric and conformational descriptors rather than local sequence signatures to describe protein function. It permits classification of proteins into families, even if there is little or no sequence identity to other proteins in the database. Thus, proteins that fall below the twilight zone of sequence identity will still be amenable to analysis. Nor does it rely on matching of the overall protein structure. Thus, proteins with similar structures but different functions will be classified differently by this method. Note that the term "function," as used here, is defined very narrowly; what is meant is the biochemical activity of the protein of interest.

The one major disadvantage of this method is that the structure of the protein must be known. However, as described below, FFFs are specific and unique enough that the structure does not have to be known to high resolution. Low- to moderate-resolution structures are sufficient for functional recognition, and current state-of-the-art prediction algorithms can often predict protein structure at sufficient resolution to allow identification of function using the FFFs. Finally, these prediction algorithms can be scaled up to analyze complete genomes.

Representative Case: The Glutaredoxin/Thioredoxin Family

Overview

In what follows, we consider the glutaredoxin/thioredoxin protein family. These proteins were selected because members of these families have tertiary structures that have been predicted by ab initio methods (e.g., in Table 4.2, 1ego is a glutaredoxin). This family also satisfies the requirement that the functional motif is not simply local in sequence, which could mean that difficulties might be expected in identifying all members of the family from sequence based-methods. Members of the glutaredoxin/ thioredoxin protein family are small proteins that catalyze thiol-disulfide exchange reactions via a redox-active pair of cysteines in the active site. While glutaredoxins and thioredoxins catalyze similar reactions, they are distinguished by their differential reactivity. Glutaredoxins contain a glutathione binding site, are reduced by glutathione (which is itself reduced by glutathione reductase), and are essential for the glutathione-dependent synthesis of deoxyribonucleotides by ribonucleotide reductase. Thioredoxins are reduced directly by the specific flavoprotein thioredoxin reductase and act as more general disulfide reductases. Ultimately, however, reducing equivalents for both proteins come from NADPH. Protein disulfide isomerases (PDIs) have been found to contain a thioredoxin-like domain and thus also have a similar activity.

The active site of the redoxin family contains three invariant residues: two cysteines and a *cis*-proline. Mutagenesis experiments have shown that the two cysteines separated by two residues are essential for significant protein function. The side chains of these two residues are oxidized and reduced during the reaction (Yang and Wells, 1991; Bushweller, Aslund et al., 1992). However, this local sequence signature is not sufficient to specifically select the members of the family. These two cysteines are also located at the N-terminus of an a-helix. Peptide studies suggest that the positive pole of the helix macrodipole affects the ionization of the cysteines and is important for protein function (Kortemme and Creighton, 1995, 1996). Another unique feature of the redoxin family is the presence of a *cis*-proline located close to the two cysteines in structure, but not in sequence. While this proline is

structurally conserved in all glutaredoxin and thioredoxin structures (Katti, Robbins et al., 1995) and is invariant in aligned sequences of known glutaredoxins and thioredoxins, its functional importance is unknown. Other residues, particularly charged residues, are also important for the Specific thiol ionization characteristics of the cysteines, but are not essential and can vary within the family (Dyson, Jeng et al., 1997).

The FFF for the glutaredoxin/thioredoxin family is based on the three-dimensional structural comparison of bacteriophage T4 glutaredoxin, 1aaz (Eklund, Ingelman et al., 1992), human thioredoxin, 4trx (Kay, Clore et al., 1990), and proline disulfide isomerase, ldsb (Martin, Bardwell et al., 1993), as well as on literature searches to find residues and structures shown to be functionally important. It consists of two cysteines separated by two residues at the N-terminus of a helix and close to a proline residue. The exact distances are described elsewhere (Fetrow and Skolnick, 1998).

Ability of the FFF to Identify the Active Site in Experimentally Determined Structures

The FFF is sufficient to distinguish proteins belonging to the redoxin family uniquely from a data set of 364 non-redundant proteins from the Brookhaven database. For this set of 364 proteins, 13 have the sequence signature -C-X-X-C-. Of these, three have a proline within the requisite distances. Of these three, only 1thx (a thioredoxin) and 1dsb (chain A, a disulfide binding protein) have the cysteines at or near the N-terminus of a helix. These two proteins are the only two true positives in the test data set, showing that this simple FFF is quite specific for the redoxin protein family. Thus, the FFF can be applied to experimental structures to identify active sites.

Application of the FFF to Predicted Structures

Is this FFF sufficient to identify the function of an inexact model of a protein, or is a high-resolution crystal or solution structure required? The structure of glutaredoxin, 1ego, was predicted with a 5.7-Å cRMSD by MONSSTER (Ortiz, Kolinski et al., 1998a,b). The sequence of this glutaredoxin exhibits less than 30 percent sequence identity to any of the three structures used to create the FFF. The redoxin FFF was applied to 25 correct structures and 56 incorrect or misfolded structures generated by MONSSTER on the 1ego sequence during the isothermal runs. It specifically selects all 25 ego-like structures as belonging to the redoxin family and rejected all 56 misfolded structures. A set of 267 correctly and incorrectly predicted structures produced by the MONSSTER algorithm for five different proteins was then created. The glutaredoxin/thioredoxin FFF was specific for the correctly folded ego structures and did not recognize any of the other correctly or incorrectly folded structures.

Screening of Entire Genomes

This sequence-to-structure-to-function concept has been applied to the analysis of the complete *E. coli* genome; i.e., all *E. coli* open reading frames (ORFs) are screened for the thiol-disulfide oxidoreductase activity of the glutaredoxin/thioredoxin protein family. The method can identify the active site residues in 10 sequences that are known to or proposed to exhibit this activity. Furthermore, oxidoreductase activity is predicted in two other sequences that have not been previously identified. These results are summarized in Table 4.4. The method distinguishes protein pairs with similar active sites from protein pairs that are just topological cousins, i.e., those having similar global folds, but not necessarily similar active sites.

TABLE 4.4 Glutaredoxins and Thioredoxins Identified in E. coli Strain K-12

	Functional Motif[a]						
Database Name[b]	Thrd/FFF[c]	Blst/FFF[d]	ps	pps	pb	b	Database Description
GLR1_ECOLI	x	x	x	x	x	x	glutaredoxin 1
GLR2_ECOLI	x		x		x[e]	x	glutaredoxin 2
GLR3_ECOLI	x	x	x	x	x	x	glutaredoxin 3
THIO_ECOLI	x	x	x	x	x	x	thioredoxin
DSBA_ECOLI	x	x	x		x[f]	x	thiol-disulfide interchange protein
DSBC_ECOLI	x		x		x[e]	x	thiol-disulfide interchange protein
DSBD_ECOLI	x	x	x	x	x	x	c-type cytochrome biogenesis protein (inner-membrane Cu tolerance protein)
DSBE_ECOLI	x		x	x[e]	x	x	thiol-disulfide interchange protein; (cyto c biogenesis protein CCMG)
YFIG_ECOLI	x	x	x	x	x	x	hypothetical thioredoxin-like protein
NRDH_ECOLI	x				x[f]	x	glutaredoxin-like NRDH protein
NRDG_ECOLI	x						anaerobic ribonucleoside triphosphate inactivating protein
B0853	x						ORF; putative regulatory protein
YIEJ_ECOLI		x					hypothetical protein in tnaB-bglB intergenic region

[a] **Functional motif:** Search of each sequence found by either BLAST/FFF or Thread/FFF protocols against the local signature databases Prosite, Prints using the Prosite scoring method, Prints using the Blocks scoring method, or Blocks. Each motif database was searched with the given sequence, and the returned scores were analyzed to see if the thioredoxin or glutaredoxin families were identified.
[b] **Database name:** This is the database identifier for each sequence. All sequences come from the SwissProt database, except B0853, which is the label given by the *E. coli* genome database. This sequence can also be accessed by the GenBank accession number ECAE000187.
[c] **Thrd/FFF**: Alignment of *E. coil* open reading frame (ORF) to the sequences of 1ego, 1dsb (chain A), or 2trx (chain A) using a threading algorithm, followed by analysis of the resulting sequence-sequence alignment for the active site residues specified by the fuzzy functional form (FFF) for the thiol-disulfide oxidoreductase activity of the glutaredoxin/ thioredoxin family. Threading results are for a combination of three different scoring methods, sq, br, and tt, as described by Godzik and coworkers (Jaroszewksi et al. 1998).
[d] **Blst/FFF**: Alignment of each *E. coli* ORF to the sequences of the lego, ldsb, chain A, and 2trx, chain A proteins using the BLAST search protocol, followed by analysis of the resulting sequence-sequence alignment for the active site residues specified by the thiol-disulfide oxidoreductase activity of the glutaredoxin/thioredoxin family. Results reported here are for a combination of the gapped-BLAST protocol and the PSI-BLAST alignment protocols. All sequences marked are found by both gapped- and PSI-BLAST, except YIFJ-ECOLI, which is found only by gapped-BLAST.
[e] Prints has three patterns for glutaredoxin/thioredoxin activity. This sequence hits only one of the patterns.
[f] Prints has three patterns for glutaredoxin/thioredoxin activity. This sequence hits only two of the patterns.

COMPUTATIONAL REQUIREMENTS FOR GENOME SCALE STRUCTURE/FUNCTION PREDICTION

The computational requirements of this type of genomic screening analysis are quite substantial. For example, contemporary ab initio protein-folding methods are applicable to single domain proteins—up to about 150 or so residues in length—and can identify possible novel protein folds. Threading is

significantly less expensive. Table 4.5 gives a summary of the CPU requirements for protein structure prediction on the genomic scale. Thus, given the extensive CPU requirements and the large number of genomic sequences, this type of sequence-to-structure-to-function paradigm would greatly benefit from the availability of teraflops-class machines. This would allow for the construction of low- to moderate-resolution predicted structures of a substantial fraction of proteins in the genome, as well as the prediction of their molecular function. Since these calculations are basically data parallel, they should be done on a machine composed of a large number of loosely coupled processors; e.g., farms of PCs are one means of achieving this. This is typical of many but not all types of calculations at the interface of chemistry and biology.

TABLE 4.5 CPU Requirements for Protein Structure Prediction on the Genomic Scale

Genome	Number of ORFS	Number of ORFS <150 Residues	Ab Initio Folding CPU Time[a]	CPU Time[b]
M. genitalium	408	82	4,920	2
H. influenzae	1,680	369	22,410	8.4
M. jannaschii	1,735	425	25,500	8.9
E. coli	4,290	879	52,740	21.5
S. cerevisiae	5,885	1,433	85,980	29.4

[a] Assumes an average of 60 CPU days to perform 1,000 folding simulations per sequence on a single processor of an SGI ORIGIN 200 running at 180 megahertz.
[b] 200 sequences threaded through 1,000 structures takes 1 CPU day.

OUTLOOK FOR THE FUTURE

While low- to moderate-resolution models can be used to predict protein biochemical activity, they are too crude to be used in drug ligand design. Techniques that allow for refinement of these low-resolution to higher-resolution models must be developed. One can imagine a hierarchical approach where the overall topology of the protein is predicted using a reduced protein model, and then atomic detail is added. Such simulations being done at atomic detail will be very CPU-intensive and can profitably exploit the parallelism of current molecular dynamics codes such as AMBER (Pearlman, Case et al., 1991) or CHARMM (Brooks, Bruccoleri et al., 1983). Recently there has been encouraging progress along this direction both for folding of small proteins at atomic detail (Duan, Wang et al., 1998) and for the refinement of protein structures starting from a reduced protein model and finishing with molecules at atomic detail (Simmerling, Lee et al., 1998). To accomplish this goal, in general, will require the development of more efficient conformational sampling algorithms as well as better potentials that can discriminate the native conformation from the myriad of alternative structures.

In the area of structural genomics, where the objective is to determine the structure of all possible types of protein folds (Holm and Sander, 1996), computation will also play a key role. This will happen in sequence selection where the goal is to identify sequences likely to adopt novel folds and where ab initio techniques may prove to be particularly useful, as well as in the development of techniques that will allow for more rapid structure determination. Here, approaches that combine a limited amount of experimental data with structure prediction may prove to be particularly powerful (Monge, Friesner et al., 1994; Aszodi, Gradwell et al., 1995; Monge, Lathrop et al., 1995; Mumenthaler and Braun, 1995; Dandekar and Argos, 1996; Skolnick, Kolinski et al., 1997; Kolinski and Skolnick, 1998). Such

experimental data may come from nuclear magnetic resonance, from electron microscopy, and from low-resolution X-ray crystal structures.

Another promising area of investigation will be in the prediction of protein binding regions. This will be the first step toward identifying multidomain interactions, both in the sense of predicting which proteins interact as well as where they interact. Then, the simulation of more complex interactions involving the components of various signaling pathways and metabolic cascades will have to be addressed. The very elegant studies of Schulten and coworkers on the light harvesting complex are an excellent example of the power of such approaches (Hu and Schulten, 1998). More generally, the simulation of membrane proteins and the prediction of their structure and function will also be a very important, computer-intensive area of investigation (Milik and Skolnick, 1992, 1993; Heijne, 1994, 1995; Stowell and Rees, 1995; Casadio, Fariselli et al., 1996) and will be the active focus of future research in the next 5 to 10 years. In addition to studies at full atomic detail, hierarchical approaches that represent the system at different levels of detail will be developed. In this regard, an interesting preliminary study is found in the simulation of virus coat protein assembly (Rapaport, Johnson et al., 1998).

Another very important area of investigation that touches on the areas of computer science, biology, and chemistry will be in the development and presentation of large databases containing all that is known about a given protein, its structure, and molecular and physiological function. Basically, since so much information is and will be available, means must be developed to make it usable and understandable to both the specialist and the nonspecialist alike. This is a very outstanding unsolved problem, but it is a reasonable guess that Web-based tools are going to be very important.

SUMMARY

These studies demonstrate that protein function prediction based on the sequence-to-structure-to-function paradigm can successfully compete with more standard sequence-based approaches and may well identify the function of additional proteins in the twilight zone of sequence identity. What is very encouraging is that low-resolution structures as provided by state-of-the-art tertiary structure predictions can identify active sites by using appropriate three-dimensional conformational descriptors, the fuzzy functional forms. Future methodological developments may allow for the prediction of protein structures at the resolution required for automated drug design. This will enable the sequence-to-structure-to-function paradigm to realize its full potential. More generally, large-scale simulations that describe the interactions of large protein (and/or membrane) aggregates will be undertaken in the near future. Such simulations will not only provide fundamental insights into how various cellular processes work at the microscopic and mesoscopic level, but may also suggest therapeutic approaches at the molecular level for the treatment of numerous diseases. These advances in algorithms and techniques at the interface of biology and chemistry will rely on the use of large numbers of inexpensive computers. Often, these can be loosely coupled, but other problems demand closely coupled, parallel machines. Whatever the mode of parallelism, advances in computational biology will, depending on the specific problem, require the availability of 1 to 100 teraflops-class machines. Given the advances in raw CPU power as well as theoretical understanding, there is every reason to believe computational biology and chemistry will play a major role in the genomics revolution.

Literature Cited

Altschul, S., T. Madden, et al. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res*. **25**: 3389-3402.

Aszodi, A., M.J. Gradwell, et al. (1995). Global fold determination from a small number of distance restraints. *J. Mol. Biol* **248**: 308-326.
Attwood, T. and M. Beck (1994). PRINTS—A protein motif fingerprint database. *Protein Eng*. **7**: 841-848.
Attwood, T., M. Beck, et al. (1997). Novel developments with the PRINTS protein fingerprint database. *Nucleic Acids Res*. **25**: 212-216.
Bairoch, A. (1990). *Prosite: A Dictionary of Protein Sites and Patterns* . Department de Biochimie Medicale, Universite de Geneva, Geneva.
Bairoch, A., P. Bucher, et al. (1995). The PROSITE database, its status in 1995. *Nucleic Acids Res*. **241**: 189-196.
Branden, C. and J. Tooze (1991). *Introduction to Protein Structure* . New York and London, Garland Publishing, Inc.
Brindle, P., K. and M.R. Montminy (1992). The CREB family of transcription factors. *Curr. Opin. Genet. Develop*. **2**: 199-204.
Brooks, B.R., R. Bruccoleri, et al. (1983). CHARMM: A program for macromolecular energy minimization, and molecular dynamics. *J. Comp. Chem*. **4**: 187-217.
Bryant, S.H. and C.E. Lawrence (1993). An empirical energy function for threading protein sequence through folding motif. *Proteins* **16**:92-112.
Bult, C.J., O. White, et al. (1996). Complete genome sequence of the methanogenic archaeon *Methanococcus jannaschii. Science* **273**: 1058-1073.
Bushweller, J.H., F. Aslund, et al. (1992). Structural and functional characterization of the mutant *Escherichia coli* glutaredoxin (C14-S) and its mixed disulfide with glutathione. *Biochemistry* **31**: 9288-9293.
Casadio, R., P. Fariselli, et al. (1996). A predictor of transmembrane a-helix domains of proteins based on neural networks. *Eur. Biophys. J*. **24**: 165-178.
Casari, G., C. Ouzounis, et al. (1996). GeneQuiz II: Automatic function assignment for genome sequence analysis. *The First Annual Pacific Symposium on Biocomputing*. World Scientific, pp. 708-709.
Dandekar, T. and P. Argos (1996). Identifying the tertiary fold of small proteins with different topologies from sequence and secondary structure using the genetic algorithm and extended criteria specific for strand regions. *J. Mol. Biol*. **256**: 645-660.
Duan, Y., L. Wang, et al. (1998). The early stage of folding of villin headpiece subdomain observed in a 200 nanosecond fully solvated molecular dynamics simulation. *Proc. Natl. Acad. Sci. U.S.A*. **95**: 9897-9902.
Dyson, H.J., M.F. Jeng, et al. (1997). Effects of buried charged groups on cysteine thiol ionization and reactivity in *Escherichia coli* thioredoxin: Structural and functional characterization of mutants of Asp 26 and Lys 57. *Biochemistry* **36**: 2622-2636.
Eklund, H., M. Ingelman, et al. (1992). Structure of oxidized bacteriophage T4 glutaredoxin (thioredoxin). Refinement of native and mutant proteins. *J. Mol. Biol*. **228**: 596-618.
Fetrow, J., A. Godzik, et al. (1998). Functional analysis of the *Escherichia coli* genome using the sequence-to-structure-to-function paradigm: Identification of proteins exhibiting the glutaredoxin/thioredoxin disulfide oxidoreductase activity. *J. Mol. Biol*. **282**: 703-711.
Fetrow, J. and J. Skolnick (1998). Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases. *J. Mol. Biol*. **281**: 949-968.
Göbel, U., C. Sander, et al. (1994). Correlated mutations and residue contacts in proteins. *Proteins* **18**: 309-317.
Godzik, A., A. Kolinski, et al. (1993). De novo and inverse folding predictions of protein structure and dynamics. *J. Comp. Aided Mol. Design* 7: 397-438.
Godzik, A., J. Skolnick, et al. (1992). A topology fingerprint approach to the inverse folding problem. *J. Mol. Biol*. **227**: 227-238.
Gribskov, M., A.D. McLachlan, et al. (1987). Profile analysis: Detection of distantly related proteins. *Proc. Natl. Acad. Sci. U.S.A*. 84: 4355-4358.
Heijne, G.v. (1994). Membrane proteins: From sequence to structure. *Annu. Rev. Biophys. Biomol. Struct*. **23**: 167-192.
Heijne, G. v. (1995). Membrane protein assembly: Rules of the game. *Bioessays* **17**(1): 25-30.
Henikoff, S. and J. Henikoff (1991). Automated assembly of protein blocks for database searching. *Nucleic Acids Res*. **19**: 6565-6572.
Holm, L. and C. Sander (1996). Mapping the protein universe. *Science* **273**: 595-602.
Holm, L. and C. Sander (1997). Dali/FSSP classification of three dimensional protein folds. *Nucleic Acids Res*. **25**: 231-234.
Hu, X. and K. Schulten (1998). Model for the light harvesting complex I (B875) of *Rhodobacter spheroides. Biophys. J*. **75**: 683-694.

Jaroszewski, L,Rychlewski, L., et al. (1998). Fold prediction by a hierarchy of sequence, threading, and modeling methods. *Protein Sci*. **7**: 1431-1440.

Katti, S.K., A.H. Robbins, et al. (1995). Crystal structure of thioltransferase at 2.2 Å resolution. *Protein Sci*. **4**: 1998-2005.

Kay, J.D.F., G.M. Clore, et al. (1990). Studies on the solution conformation of human thioredoxin using heteronuclear15N-1H nuclear magnetic resonance spectroscopy. *Biochemistry* **29**: 1566-1572.

Kolinski, A. and J. Skolnick (1998). Assembly of protein structure from sparse experimental data: An efficient Monte Carlo model. *Proteins* **32**: 475-494.

Kolinski, A., J. Skolnick, et al. (1997). A method for the prediction of surface U-turns and transglobular connections in small proteins. *Proteins* **27**: 290-308.

Kolinski, A.K. and J. Skolnick (1996). *Lattice Models of Protein Folding, Dynamics and Thermodynamics*. Austin, Tex., R.G. Landes Company.

Kortemme, T. and T.E. Creighton (1995). Ionisation of cysteine residues at the termini of model alpha-helical peptides. Relevance to unusual thiol pKa values in proteins of the thioredoxin family. *J. Mol. Biol*. **253**: 799-812.

Kortemme, T. and T.E. Creighton (1996). Electrostatic interactions in the active site of the N-terminal thioredoxin-like domain of protein disulfide isomerase. *Biochemistry* **35**: 14503-14511.

Martin, J.L., J.C. Bardwell, et al. (1993). Crystal structure of the DsbA protein required for disulphide bond formation in vivo. *Nature* **365**: 464-468.

Milik, M. and J. Skolnick (1992). Spontaneous insertion of polypeptide chains into membranes: A Monte Carlo model. *Proc. Natl. Acad. Sci. U.S.A*. **89**: 9391-9395.

Milik, M. and J. Skolnick (1993). Insertion of peptide chains into lipid membranes. An off-lattice Monte Carlo dynamics models. *Proteins* **15**: 10-25.

Miller, R.T., D.T. Jones, et al. (1996). Protein fold recognition by sequence threading: Tools and assessment techniques. *Federation of American Societies for Experimental Biology (FASEB) Journal* **10**: 171-178.

Monge, A., R.A. Friesner, et el. (1994). An algorithm to generate low-resolution protein tertiary structures from knowledge of secondary structure. *Proc. Natl. Acad, Sci. U.S.A*. **91**: 5027-5029.

Monge, A., E.J.P. Lathrop, et al. (1995). Computer modeling of protein folding: Conformational and energetic analysis of reduced and detailed protein models. *J. Mol. Biol*. **247**: 995-1012.

Mumenthaler, C. and W. Braun (1995). Predicting the helix packing of globular proteins by self-correcting distance geometry. *Prot. Sci*. **4**: 863-871.

Murzin, A.G. (1996). Structural classification of proteins: New superfamilies. *Curr. Opin. Struct. Biol*. **6**: 386-394.

Murzin, A.G., S.E. Brenner, et al. (1995). Scop: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol*. **247**: 536-540.

Olmea, O. and A. Valencia (1997). Improving contact predictions by the combination of correlated mutations and other sources of sequence information. *Folding & Design* **2**: S25-S32.

Orengo, C.A., A.D. Michie, et al. (1997). CATH—A hierarchic classification of protein domain structures. *Structure* **5**: 1093-1108.

Ortiz, A., A. Kolinski, et al. (1998a). Fold assembly of small proteins using Monte Carlo simulations driven by restraints derived from multiple sequence alignments . *J. Mol. Biol*. **277**: 419-448.

Ortiz, A., A. Kolinski, et al. (1998b). Nativelike topology assembly of small proteins using predicted restraints in Monte Carlo simulations. *Proc. Natl. Acad. Sci. U.S.A*. **95**: 1020-1025.

Ortiz, A., A. Kolinski, et al. (1998c). Tertiary structure prediction of the KiX domain of CBP using Monte Carlo simulations driven by restraints derived from multiple sequence alignments. *Proteins* **30**: 287-294.

Pearlman, D.A., D.A. Case, et al. (1991). Assisted Model Building with Energy Refinement (AMBER) code. University of California, San Francisco.

Pearson, W. and D. Lipman (1988). Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. U.S.A*. **85**: 2444-2448.

Radhakrishnan, I., G.C. Perez-Alvarado, et al. (1997). Solution structure of the KIX domain of CBP bound to the transactivation domain of CREB: A model for activator:coactivator interactions. *Cell* **91**: 741-752.

Rapaport, D.C., J.E. Johnson, et al. (1998). Supramolecular self-assembly: Molecular dynamics modeling of polyhedral shell formation. *Comput. Phys. Commun*., submitted.

Rastan, S. and L. Beeley (1997). Functional genomics: Going forwards from the databases. *Curr. Opin. Genet Devel*. **7**: 777-783.

Rost, B. and C. Sander (1993). Prediction of secondary structure at better than 70% accuracy. *J. Mol. Biol.* 232: 584-599.

Rost, B., R. Schneider, et al. (1993). Progress in protein structure prediction? *TIBS* **18**: 120-123.

Sali, A. and T. Blundell (1993). Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**: 779-815.
Sander, C. and R. Schneider (1991). Database of homology derived protein structures and the structural meaning of sequence alignment. *Proteins* **9**: 56-68.
Simmerling, C., M. Lee, et al. (1998). Combining MONSSTER and LES/PME to predict protein structure from amino acid sequence: Application to the small protein CMTI-1 *J. Am. Chem. Soc*., submitted.
Skolnick, J., A. Kolinski, et al. (1997). MONSSTER: A method for folding globular proteins with a small number of distance restraints. *J. Mol. Biol*. **265**: 217-241.
Stowell, M.H.B. and D.C. Rees (1995). Structure and stability of membrane proteins. *Adv. Protein Chem*. **46**: 279-311.
Thomas, D.J., G. Cesari, et al. (1996). The prediction of protein contacts from multiple sequence alignment. *Protein Eng*. **11**: 941-948.
Wodak, S.J. and M.J. Rooman (1993). Generating and testing protein folds. *Curr. Opin. Struct. Biol*. **3**: 247-259.
Yang, Y.F. and W.W. Wells (1991). Identification and characterization of the functional amino acids at the active center of pig liver thioltransferase by site-directed mutagenesis. *J. Biol. Chem*. **266**: 12759-12765.

DISCUSSION

William Winter, SUNY-ESF, Syracuse: Glycosylation has to play a major role in the final selection of a particular protein conformation in many proteins where it does occur. Are you doing anything at all to use that kind of information to make further selections once you have determined a family of possible structures?

Jeffrey Skolnick: Not yet, but we are aware of the problem. So far we have picked molecular functions that are basically self-contained by design because we did not pick the hardest case first. But you are absolutely right, glycosylation is extremely important. The problem there is that not a lot is known. Even the potentials that you should put in to describe the conformational spectrum are not well established. People are still developing these, so that field is very much in its infancy. Our view has been, yes, we recognize it is important, and especially in a biological context it is very, very important; it protects the proteins and keeps them from being chewed up, but we quite frankly wanted to consider the simplest cases first to see if the basic approaches could work—choose molecular functions or biochemical functions where it is apparently not believed to be important and then work our way up. But, yes, you are absolutely right. One day we or someone else will have to deal with that problem, but I think it is premature at this stage of the game.

David Dixon, Pacific Northwest National Laboratory: Jeff, have you looked at or have you started thinking about the fact that there is also spatial resolution within a cell, and have you looked at how you connect your proteins up into cell signaling pathways?

Jeffrey Skolnick: Yes, we have already started, at least on a very schematic level, simulating peptide insertion and protein insertion into membranes, treating the system, you know, with spatial anisotropy. You have a membrane region that could be treated at various levels of detail in the interfacial regions, bulk regions, but only on a very, very schematic level at this point. As it is, these kinds of calculations really tax any resources that we can get hold of, and we are not sure about adding additional details other than on a very simplified level. And then we are not even sure that the descriptives are sufficiently good that it would be worthwhile. I mean, we are trying to proceed on a very building-block basis: establish something that works, validate it, move on, make it more complicated, move on. My guess is the next thing we are going to do is membrane protein tertiary structure prediction, and there there are some encouraging results.

5

Needs and New Directions in Computing for the Chemical Process Industries

W. David Smith, Jr.
E.I. DuPont

This paper is organized into six sections. The first provides some general background information and highlights the main factors driving the chemical process industries (CPI) today in order to provide a context for all that follows. The second presents my view of the total process simulation requirements of the CPI and defines the scope of the problem. The next section then provides a brief overview of the current tool set that is being used in the CPI. The fourth section introduces the European CAPE OPEN project, funded by Brite-Euram and nearing completion; it also discusses briefly the follow-on project, GLOBAL CAPE OPEN. The fifth section considers the closely related topic of process control, and the final section lists conclusions.

BACKGROUND AND FACTORS DRIVING THE CHEMICAL INDUSTRY

Let's begin with a few general observations about the CPI. Chemical and petroleum plants require huge, long-term capital investments. The lifetime of a plant, once built, will range from 25 to 50 years. Processes and process designs are almost always customized. We (DuPont) have quite a few nylon plants, but no two are alike. Each time a new plant is built, we incorporate improvements that have been learned from the operation of the last plant that was built. This places some constraints on the costs that

AUTHOR'S NOTE: This paper did not go through the approval process that is normally required by DuPont. Therefore, this should not be viewed as an official DuPont perspective on the needs and new directions in computing for the chemical process industries. Rather it is my personal view of the subject that is based on 20 years of university teaching and 20 years of experience with DuPont, the latter being most relevant to the topic at hand. Furthermore, at times it is necessary to refer to commercial software products and make contrasts between their respective features. These references are not meant to be either a recommendation or a condemnation of a particular product but are meant to illustrate different approaches that have been taken to try to satisfy the simulation needs of the chemical process industries. In fact, it is safe to say that the complete simulation tool for all of our needs does not and is never likely to exist because of the enormous diversity within a company like DuPont and the industry as a whole. Also, generalizations will inevitably be used to make a point about the field for which someone will be able to provide an exception or a counterexample, but in a broad-brash overview of any field that is to be expected.

one is willing to accept for a particular plant. For example, if you are designing an airplane and you expect to build a thousand of them, then you can afford to design and build in a super-sophisticated control system because the cost will be spread out over many units. That does not happen in the CPI.

Our plants are large, very complex, and, I am sorry to say, still relatively poorly understood in the sense that we don't know all the details of the chemical reactions and their associated thermodynamics and transport properties needed for the design of these facilities. Because of this lack of knowledge and the inherent safety concerns of handling some very dangerous chemicals, our plants tend to be over-designed to ensure that both safety and production goals will be met. Even for some of our oldest and best-known products we are still improving our fundamental understanding. This is happening largely because of vastly improved analytical instrumentation. In some cases we have benefited significantly from the application of computational chemistry to provide basic data that otherwise would have been too costly and time consuming to measure in the laboratory.

Most of our large plants, particularly those that make polymers, were justified on economies of scale. Initially that made sense because we were only expecting to make one or two products. Over time, because of good chemistry and market demand, the number of products has grown, which has often meant that the plant has had a much broader range of operating conditions. In many cases we have designed and built very good “battleships” but now, because of the proliferation of products, we are forced to try and run them as if they were "PT boats." Another characteristic of those polymer plants that make filament and sheet products is that they typically have a very wide range of time scales. At the front end, the reactor might have a time constant that is measured in hours while at the back end of the process where the filament or sheet is being formed, the time constant may be measured in milliseconds or seconds. The combination of multi-product transitions with wide ranges of operating conditions and time constants leads to some very challenging manufacturing problems.

The essential problem facing the chemical process industries is shown in Figure 5.1. The graph shows the trend of capital productivity with time for both total manufacturing and chemicals. Capital productivity tries to measure how effectively capital is being utilized. Simply stated, for each dollar

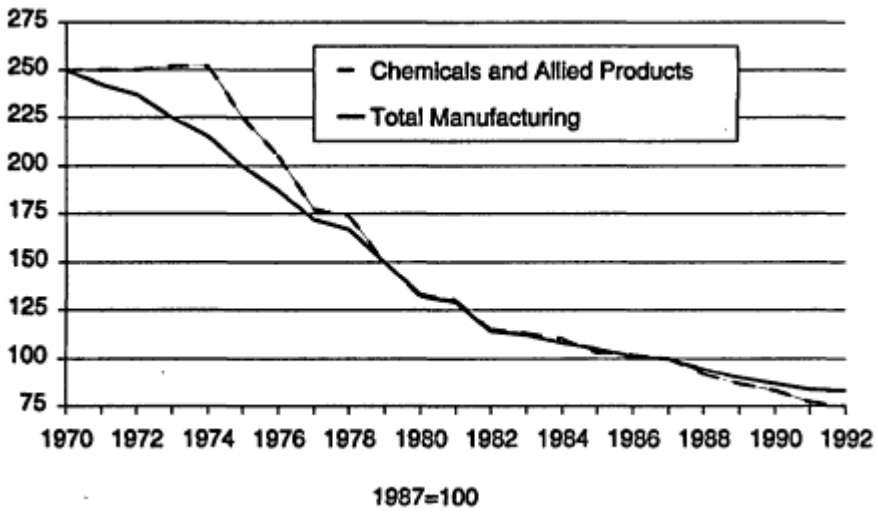


Figure 5.1
Capital productivity for all manufacturing and for chemicals.

invested in manufacturing facilities, how much profit do I make? The break-even point was assigned arbitrarily a value of 100, so it easy to see that since 1987 the chemical industry has not been the best place to invest your money. The major contributors to the decline have been increased energy costs after the oil embargo, overcapacity and tougher global competition, increased costs for environmental programs, and new business ventures outside of core competencies that have been costly and unprofitable.

In this setting, what are chemical companies focusing on? Certainly the answer will vary from company to company but, using DuPont as an example, the top five items are asset productivity, growth, quality, environmental issues, and shorter development cycle. The middle three really don't need any explanation. Asset productivity deals with the problem of squeezing even greater profits from the large investment we already have in the ground. Three of the major factors that govern the productivity of a plant are yield, up-time, and instantaneous production rate. Yield, unfortunately, can be defined in many ways and is often defined to suit the needs of the definer. In an effort to eliminate that ambiguity we use first-pass, first-quality yield, which counts only the amount of the on-aim desired product made in one pass of the fresh feed through the reactor. One wants to eliminate rework and blending of off-specification material as well as product reclassification. In multi-product plants one also tries to minimize transition losses. Up-time focuses on preventive maintenance, as you can't make a product if the plant isn't running. Assuming you can sell what you make, one always wants to run the plant at the maximum instantaneous rate, which is usually based on the best past performance of the plant.

At a recent meeting, a speaker from Dow reported that from the time the chemist comes up with a new idea at the bench to the time that the first pound comes out of a commercial-scale plant, is typically about 12 to 13 years. That process development cycle is just a little longer than the 10- to 11-year average that DuPont likes to claim. Now that is much too long in today's very competitive and rapidly changing marketplace. If it takes you that long to go from idea to finished plant, you will probably have competition from a lower-cost replacement-in-kind product or the market could be dramatically altered by the appearance of another completely new product. Either possibility could invalidate the whole basis for the investment that you have made. In one recent example in DuPont, cycle time was reduced to 5 years. The challenge we face now is to institutionalize what we learned in that very successful project.

PROCESS SIMULATION NEEDS OF THE CPI

What is the process of designing a plant, and what software tools are needed? Figure 5.2 presents a block diagram of the major steps in the process of designing a chemical plant. The impetus to build a plant generally comes from one of three sources: a new discovery from the lab, a response to a customer request, or a needed capacity increase for an existing product. In the third case, the design process should be easier because a lot is already known from the existing plant. Since, in principle, the first two options require more work and essentially the same steps, we will assume that we are dealing with a potential new product that has been discovered in the lab.

Borrowing from Edward Tufte[1] in the preparation of Figure 5.2, we intend the thickness of the arrows to convey the bandwidth of the data and information that are transferred between the activities represented by the boxes. The idea for a new product and the process to manufacture it would originate in either Box 1 or 3. Ideally, one would get a process engineer, skilled in process synthesis, working

[1] Tufte, Edward. 1992. *The Visual Display of Quantitative Information* . Chesire, Conn.: Graphics Press.

with the chemist(s) as soon as the potential new product begins to look interesting and preliminary cost estimates for a process to manufacture it are needed. At this stage it is important to examine alternate raw materials and different reaction conditions such as temperature, pressure, choice of solvent, and choice of catalyst. The selectivity, yield, and nature of the by-products of the reaction determine the complexity and cost of the rest of the process. The rapid testing of alternatives at this stage is the key to successful process development and plant design. The heavy arrows connecting Boxes 2 and 3 indicate the intensity of that interaction. Unfortunately, there is no commercially available software to aid in the process synthesis step. One very interesting program, PIP-II, has been developed at the University of Massachusetts by James Douglas and his students to address these activities. The difficulty in using commercial process simulators for process synthesis is that the equipment modules they provide require complete kinetic and thermodynamic data that are never available at this early stage of process design. In some cases one can make effective use of molecular modeling, Box 4, to help develop estimates of the missing data. That played a prominent role in the DuPont example cited above with a development cycle time of only 5 years. This part of the overall design process usually produces a small number of "attractive" process alternatives that have been obtained by using consistent but approximate design and economic calculations to screen many potential processes. During the process synthesis step many calculations were done on the basis of ideal vapor-liquid equilibria data. Since the next step requires

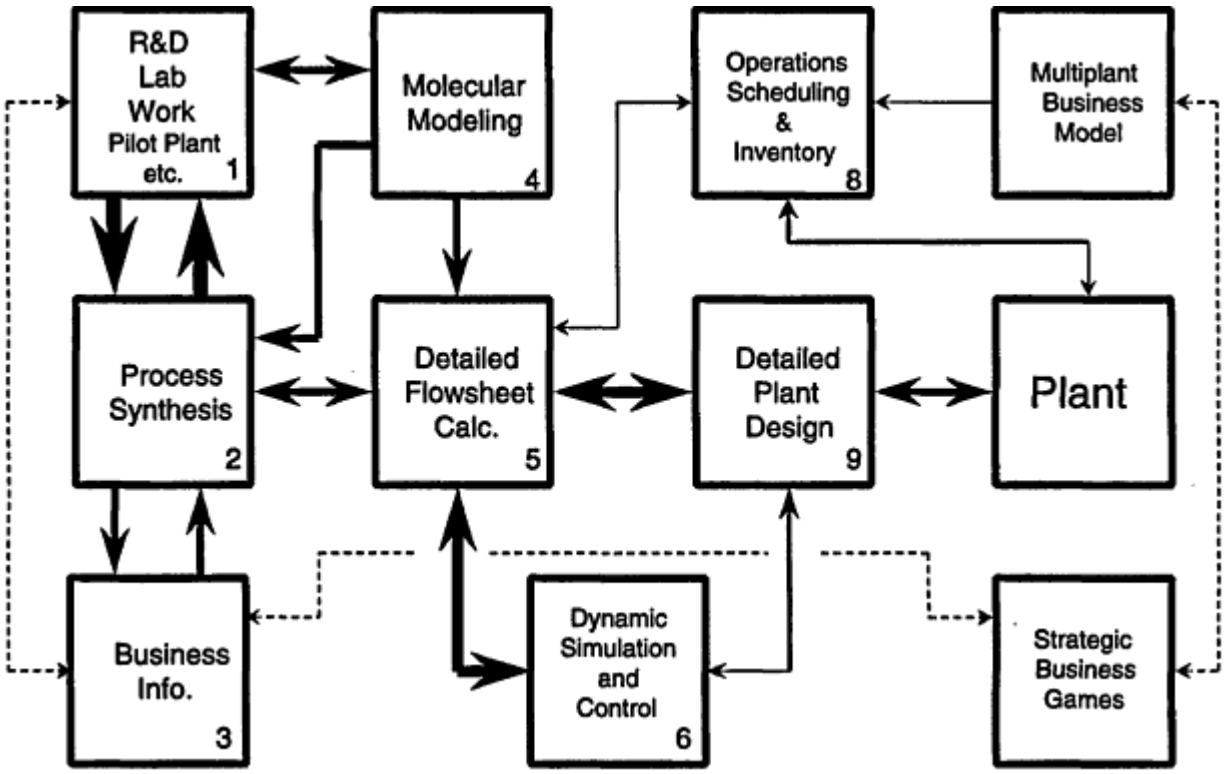


Figure 5.2
Process of plant design.

more accurate kinetic and thermodynamic data, every opportunity should be taken during this stage of the development process to extract this information from the available experimental data.

At this time in the process, one needs to develop more rigorous and detailed estimates of process performance and cost so that the final choice can be made between the alternatives developed in the process synthesis step. This is generally accomplished by the use of commercial process simulators from the three major vendors, such as Aspen Plus from Aspen Tech, HYSYS from Hyprotech, and Pro II from Simulation Sciences (see Box 5). A lot of work gets done here, and it is not unusual for about 80 percent of the effort to be spent on getting reasonably accurate kinetic and thermodynamic data. To the extent that this has been done during the process synthesis step it will speed up this part of the process. While the steady-state process design is being evaluated it is also very important to evaluate the controllability and operability of the candidate processes. This is accomplished by developing an appropriate dynamic model of the process (Box 6). Several of the commercial simulation programs now claim to be able to convert the steady-state simulation into a dynamic simulation. For a single piece of equipment, that claim may be correct, but for a process flow sheet that is neither possible nor, in general, desirable. In a steady-state flow sheet the lines connecting the different pieces of equipment do not have to have any diameters or lengths specified since their role is simply to pass information to the next unit. In a dynamic simulation, the size of the lines is essential to track the propagation of disturbances from one vessel to the next. Furthermore, in most process control analyses, approximate or low-fidelity models are perfectly adequate and one would not want to use models of the complexity typically used in a detailed flow sheet calculation.

In the case where the design is for a multi-product plant, we should try to ensure that the equipment is designed such that the product scheduling can be accomplished with minimum cost. At present there are no design methodologies to solve this problem. The design is typically evaluated by doing case studies (Box 8). When controllability and operability with scheduling considerations are satisfied, a basic data package is assembled and transmitted to the organization that will do the detailed equipment design and construction (Box 9).

After the plant is built, we still have to run the plant and worry about what our competitors are doing. If the plant has been built to increase capacity of an existing product, then it is very likely that the other plants will be distributed around the world. We have global operating, scheduling, and inventory decisions that have to be made. These are usually formulated as supply chain optimization problems (Box 9).

Every box in Figure 5.2 requires a different computer program—if one exists at all—and none of them communicate in any reasonable way. This leads to inefficiency in the design process, as some of the output from one program needs to be re-entered into one or more additional programs. The programs come from different vendors, and they invariably use different methods for the evaluation of thermodynamic properties, which adds to our problems. It should also be obvious that this is a significant interdisciplinary process that requires skills that range from wet chemistry to molecular modeling to the optimization of large supply chain problems.

Commercial physical property databases do not cover our needs. We run reactions that range from less than 100 degrees in the liquid phase to solid reactions at 1,200 degrees to plasma reactors at 5,000 degrees. We run polymer reactions that span a pressure range from one to 2,000 atmospheres. Most companies have their own proprietary physical property databases that have to be incorporated into one or more of the simulation programs that are used in Boxes 2, 5, 6, and 9. All of the current commercial simulation products are closed, proprietary programs, which makes the job of incorporating proprietary thermodynamics or equipment modules much more difficult.

Another serious problem that none of the vendors has addressed in a significant way is the visualization of results. Simulation programs generate masses of numbers, and there is no easy way to take the output from these simulation packages and look at it in a way that makes sense. A recent development has been to allow the export of results to an Excel™ spreadsheet. I suppose that is a small step in the right direction, but I don't consider it to be a general solution to this problem.

AN OVERVIEW OF CURRENT SIMULATION PROGRAMS

To make sense out of some of the comments that follow, it would be helpful to look at Figure 5.3. In my view, there are at least three important views of software. Two views belong to the user community; one group is the casual or infrequent user, while the second is the power user who tries to squeeze as much capability out of the program as he or she can. The final view is that of the person or team designing and writing the software. Issues that the developers need to resolve are experience level of the user for whom the product is intended, choice of computer platform(s), design for easy maintenance, management of versions and updates, provision of help and error diagnostics to the user, etc. For this discussion the first issue is the key one. Is the product intended for the casual user or the power user? If the choice is made to design for the casual user, a lot of care will be taken to prevent the user from inadvertently gaining access to sections of the code that might govern the choice of convergence algorithm or criteria for convergence or other performance parameters to prevent the user from "making mistakes." Once the decision has been made to "bulletproof" the code in this way, the ability of the power user to use the software in sophisticated ways has been significantly curtailed. The power users tend to be frustrated by these imposed limitations. The major concern of the vendors is to be profitable so they need to concentrate on the largest market, which, without question, is the casual user. In DuPont the ratio of casual to power users is about 20/1.

Having decided which category of user to target, there are two choices for the simulation "technology" to implement. They are the sequential modular approach and the equation-based approach. The

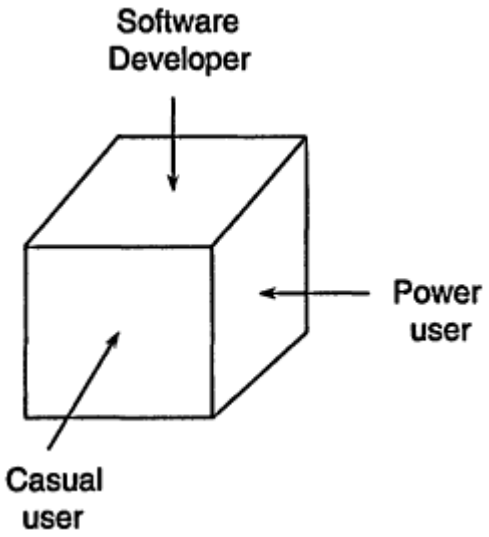


Figure 5.3
Three important perspectives from which to view software.

sequential modular approach is the oldest and most common approach and is used by all the major vendors. In this approach the vendor provides a library of equipment modules that the user can connect in a sequence that best represents the process. The problem is that the vendors have the experience and ability to provide only a limited range of equipment modules to industry. In particular, the ability to model reactors in these systems is very, very limited.

Sequential modular simulators are best suited to do equipment rating instead of equipment design. Equipment rating means that one wants to evaluate the performance of a piece of installed equipment. Taking a distillation column as an example, the user would know the number of trays, location of the feed stream, and the appropriate set of physical properties and would wish to determine the effect of changes in feed rate, reflux ratio, or reboiler heat input on column performance. In contrast, the design problem is usually posed as how big the column has to be to separate a feed stream of known rate and composition to products of specified purity. It is clumsy to do design calculations in sequential modular simulators, but it can be done by trial and error. In my opinion, the sequential simulator has been designed for the casual user. Key word input is giving way to click and drag input on PCs. Convergence is still a problem for large flow sheets, and the diagnostics that are available when convergence fails are not extremely useful.

Equation-based simulators are for the power user. The learning curve is much steeper. One big advantage is that both steady-state and dynamic models are possible. Here the Achilles heel is that it is essential to scale the equations that describe your equipment or you face severe stiffness problems. Until very recently the diagnostics in these systems were abysmal (that is an understatement), and they can still be better.

One of the other advantages of the equation-based approach is that, for a specialized piece of equipment, it is much easier for the engineer to write down the equations that describe it, put them into the equation-based simulator, and let it solve them. Before the solution can proceed, the user has to specify which variables are independent and which ones are fixed. This means that, once the equations have been scaled and entered in the simulator, the user can easily go from the rating problem to the design problem just by specifying what the dependent and the independent variables are in a particular problem. Other advantages of the current generation of equation-based simulators are that partial differential equations can now be handled much more easily and the dynamic models that can be developed in them can easily be linked to a distributed control system to provide operator training systems.

Since they didn't appear in Figure 5.2, some comment should be made about codes for computational fluid mechanics. They involve a very steep learning curve and tend to be used only by specialists. One of the major problems they have is a very serious lack of physical property support. One important application area is multi-phase reactor problems, but those are still very difficult to solve. Accuracy can be a problem because of accumulated round-off errors in these very lengthy calculations. We solve some problems on a Clay computer, running for anywhere from 10 to 24 hours. When a run is complete, the question is, How many significant digits do you believe you really have in the answer? Even with double-precision calculations on the Cray we believe that in some problems we can count on only two significant digits in the answer.

Optimization codes also deserve some comment. These are still largely the domain of specialists and power users. Global optimization methods for general problems are still an active research area. Algorithms often need to be tailored to the structure of the problem to get reasonable solution times.

THE EUROPEAN CAPE OPEN PROJECT

CAPE is an acronym for computer-aided process engineering. OPEN refers to having the vendors support non-proprietary open standards for commercial simulators. The funding from Brite-Euram was approximately $3.5 million for 3 years, and this first phase of the project ends in the middle of 1999.

Although this is a European activity, the only major software vendors at the time were all U.S. or Canadian. So by default they were invited to participate. Since that time Hyprotech and Simulation Sciences have been purchased by large British companies, leaving Aspen as the only North American simulation company. QuantiSci, a small consulting company, is also a member. The universities participating are Aachen, Imperial College, and Toulouse. The companies involved are BASF, Bayer, BP, ELF, ICI, IFP, and DuPont Iberica.

Why were people interested in CAPE OPEN? The driving force for this activity was provided by the large German chemical companies under the leadership of Bayer. They had a real concern about the large quantity of old but valuable code that they wanted to keep on using in existing commercial simulators without having to rewrite it. This is often referred to as legacy code.

Every large chemical company that I know probably uses at least two of the three available commercial simulators; some use all three. This is expensive, and when you use two of these simulators to look at the same problem in different parts of the company you waste a lot of time trying to figure out which one of the two is really giving you the right answer, because they will be different. Since the current commercial simulators are closed proprietary systems, it means that if Hyprotech has a good azeotropic distillation model and Aspen has the required thermodynamics, you could not use the two together to solve the problem. It was also recognized that no matter how hard the vendors try, they are not going to be able to provide all the modules that we need to model the wide range of systems that we deal with on a routine basis.

The goal of the CAPE OPEN project is to develop a non-proprietary framework for simulation software that is based on software components. That requires standards for the functional interfaces between classes of software components and the use of a binary standard such as Microsoft's Active X or CORBA. A software component is compiled software, written in any language, that supports the binary standard. A software component can be either a server or a client to any other software that supports the binary standard.

The functional interfaces for which we have developed standards are thermodynamics, unit operations, and numerics. Prototypes exist for these three interfaces. They are in the final stages of testing now and they will be released in the next 6 months. What does this effort buy us? It buys us "plug-and-play" capability. It means that I will be able to take the thermodynamics from Aspen and use them in a Hyprotech simulation and it will run. We have demonstrated that capability, and the vendors are working with us to complete the interface testing. What will a CAPE OPEN-compliant simulator look like? I have tried to indicate this in Figure 5.4.

The part of the simulator that will be proprietary, which enables the vendors to compete, will be the executive. The rest of the simulator is divided into three main sections: the unit operations, which includes reactors and other specialized pieces of processing equipment; thermodynamics and transport properties; and numerical routines of all sorts. Each unit operation will be a software component that will be able to communicate with the other major pieces of the simulator to get any services or information it needs to complete its calculations through the standard interfaces that have been developed. This means that all executives will be able to "use" a unit operations software component that has the standard interfaces regardless of its origin. Clearly similar statements can be made about thermo

dynamics and numerical components. The end result is a plug-and-play simulation environment that will contain a much broader choice of components.

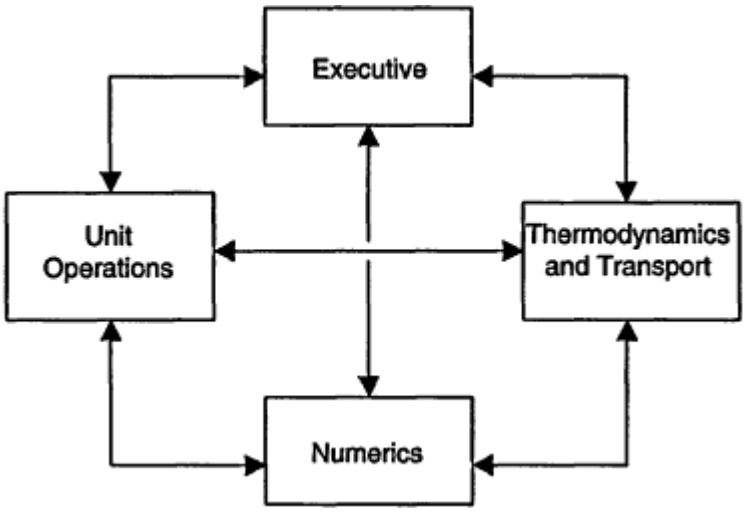


Figure 5.4
A CAPE OPEN-compliant simulator.

One of the most important aspects of having CAPE OPEN software is that it will facilitate the transfer of technology from universities to industry. DuPont had supported research on azeotropic distillation at the University of Massachusetts and we were interested in extending the contract to work on reactive distillation. Recent research had shown that a change of variables would convert the reactive distillation equations into the same form as the azeotropic equations. Near the end of the first project, the university had negotiated a deal with Hyprotech to commercialize the azeotropic code. Hyprotech was using object-oriented programming (OOP) as was DuPont, so during the planning session for the reactive distillation project the question was raised about using the inheritance features of OOP to build the reactive distillation code from the azeotropic code. Two DuPont staff members, neither accomplished programmers, went to Hyprotech and, with help from Hyprotech programmers to understand the structure of the azeotropic code, they had in 2 weeks a working reactive distillation module that would run in HYSYS. We had budgeted 1 year of time for an experienced postdoc to code the problem in FORTRAN.

One of the questions that this workshop was intended to address was how the advances in computer and communication technology would change the way industry works. You can imagine that a pump vendor, for instance, would have on its Web site a CAPE OPEN-compliant software component for each of its products. If an engineer were doing a simulation and wanted to test a new pump or consider a different pump, he or she could go to the Web site and either download the appropriate software component or run it over the Web on the server, complete the simulation, and evaluate the results. If the results were satisfactory the engineer could check to determine the availability of that model pump and perhaps place an order directly. One can imagine that all equipment suppliers to the CPI would provide similar services.

What is Global CAPE OPEN? The people involved in CAPE OPEN are very conscious of the fact that theirs is a relatively small European activity. They would like the standard interfaces that have been

developed to be accepted as universal or global standards and extended to cover a broader range of applications. Some that have not been addressed for lack of time are solids handling, polymer processing, and pharmaceutical/biological systems. Brite-Euram has granted additional funding for another 2.5 to 3 years. The membership has been extended to 18 partners from Europe, Japan, and North America. There is a problem, however, as the Brite-Euram funding is only for the European partners and, as in the past, most of it is directed to university research. Other participants have to get local funding, and many universities in the United States would like to be a part of this research program but are faced with the problem of finding a U.S. funding agency that is willing to deal with a manufacturing question like this. Industrial participation from North America has not been established as we are still looking for a regional coordinator.

PROCESS CONTROL IN THE CPI

Most processes in the chemical industry now have distributed control systems (DCSs) and all new plants would almost certainly have one included in the design specification. Furthermore, as a rule of thumb, roughly 80 percent of the control loops can be handled well with simple PID controllers. The other loops usually require some form of advanced control. As the power of the process control computers has increased, more and more advanced process control functionality has been incorporated into the DCS. Now several of the major vendors provide the ability to do moderately sized linear model predictive control applications. On the surface it would appear that most of industry's process control needs have been addressed by current DCS technology.

The process control options that are available are all based on linear control theory because that is what we know best. Unfortunately, almost all chemical processes are nonlinear, and established linear techniques work well only for mild nonlinearities. The business drivers that I described earlier are pushing the operation of our processes into regions where the nonlinearities are emphasized, so the popular linear techniques may no longer be good enough.

In a new application the first question is, Do I need to use nonlinear control? There is no simple answer to this question nor is there a simple calculation that would provide the answer. In practice the question is decided when standard linear techniques do not work, but then another question immediately arises—What nonlinear control strategy should be used? Again experience or trial and error will provide an answer. The final question becomes, How will the nonlinear control algorithm be implemented in the DCS? Currently, this question is always present because the DCS vendors only provide linear control options.

Nonlinear control technology is inherently more complicated and is generally inaccessible to most control practitioners. Nonlinear model development is several orders of magnitude more difficult than that for linear models. Each application is usually unique and a specialized implementation is required. In spite of these difficulties we believe that there is a vast untapped potential economic benefit in cases of moderate nonlinearities where linear methods can be used but where nonlinear methods would yield significant performance improvements.

There is interesting work on nonlinear control being done in the universities that we would like to try, but there is a major barrier to be overcome: the monolithic code that the DCS vendors have implemented in their DCSs. This makes it very difficult for them, let alone industrial control practitioners, to add or try new control algorithms on any commercial DCS. If the DCS vendors adopted the component-based software development strategy of the simulation software vendors, this problem could be solved and we would get a much faster and effective transfer of advanced process control technology from the universities to industry. Everyone would benefit.

CONCLUSIONS

1. Both process simulation and process control will benefit from component-based redesign of their software.
2. Technology transfer in both areas will improve dramatically.
3. Component-based software operating as clients or servers over the Web will significantly reduce the cycle time for new process development. Component-based software has the potential to alter the way that equipment suppliers market and sell their products to the CPI.
4. The potential benefit from the application of nonlinear control is large, and vendors need to provide the capability for easier implementation of these techniques.

DISCUSSION

Jack Pfeiffer, Air Products and Chemicals, Inc.: Dave, regarding your diagram on the modeling process that you said moved from left to right—one of the things that we have been toying with is a recycle loop that would bring information back from plant operations into the R&D or the process synthesis step. Have you thought about the value of that, and especially have you thought about how the software that we have today, or the directions that you are proposing, may enhance that capability?

David Smith: Well, I do not think there is anything that prevents you from doing that. I think the tools to analyze historical process data do exist but they are not integrated with the tools that I have discussed. That is one area that has not been considered by CAPE OPEN. However, I think the biggest problem you actually have is the cultural problem of what the production or operations people get rewarded for. In DuPont, that is significantly different from what R&D folks get rewarded for, so the problem ultimately becomes the willingness of the plant people to accept change. They get rewarded for continuity of operation, no labor problems, no safety problems, and pounds of product shipped. If you go to the typical operations superintendent with a proposal to improve operations, there will usually be little interest because it means downtime, retraining, and a whole bunch of things that they do not want to deal with. So, it is the cultural problem that is much more difficult than the problem of having that recycle loop of process information.

Jack Pfeiffer: As a further clarification on that, one of the things we are thinking about is that we collect a pot full of data in plants, and these data may be valuable if analyzed using the creative concepts of R&D.

David Smith: That is definitely true. We spend an inordinate amount of money collecting data that we never really analyze. The question is, Why do we do that? One of the major reasons is that we have reduced the number of people at the plants to the minimum, and those people are so busy that they do not have time to go back and look at historical data until the plant is having problems.

Gintaris Reklaitis, Purdue University: If Global CAPE OPEN is such a wonderful thing for the industry, why is the industry not "ponying up" and supporting the activity? It does not appear that there are key technology bottlenecks in terms of how to structure object-oriented codes, how to develop them. Why not "pony up" and do it?

David Smith: The issue was that it was very difficult to convince the vendors to accept the concept of open, nonproprietary simulation systems. Also the business of defining standard interfaces that will support the broad range of modeling activities encountered in the CPI was a very difficult task. One of the reasons for forming Global CAPE OPEN was to have greater participation in testing and extending those standards. Finally, my management says we have "ponied up" for 3 years with three people from DuPont participating in and leading parts of this activity. My current management feels that it is time for another U.S. company to step forward and lead the North American effort in Global CAPE OPEN.

Christos Georgakis, Lehigh University: Dave, what do you perceive as the computational challenges in achieving industrial-scale green chemistry, where green plants produce no pollution, only products? CAPE OPEN is among the challenges, but what other computational challenges exist?

David Smith: CAPE OPEN-compliant simulation software that is integrated as I suggested in Figure 5.2 will help because it will allow the evaluation of more process alternatives and thus increase the chances of finding process designs that will have minimal environmental impact. However, I believe the fundamental challenge is still one of chemistry. Greg McRae and his students have looked at some approaches to this problem. One approach is to start with different raw materials that might make a broader range of salable products but have much less or no waste. I think our businesses would not support this approach, as the yield to products other than the desired product could be significant. I think there is more hope in developing more selective catalysts and bio-catalysts.

Tom Edgar, University of Texas: I had an industrial chemist ask me recently about a problem that he is encountering. He says that his business people are on his case because every time they design a plant they find out they have about 20 percent overcapacity because of the intrinsic conservatism in designing the plant. Do you think one of the bottlenecks is this software problem?

David Smith: No, the problem is not with the existing software, nor will CAPE OPEN-compliant software solve the problem. I think the problem we all have is the fundamental uncertainty about kinetics, thermodynamics, and transport data during the design. It is costly, time-consuming work that is perceived as slowing down the design process. Since you do not know those properties accurately enough you tend to overdesign the plant. Initially that is viewed as a "bad" thing, but then 10 years down the road when you need incremental capacity you look like a hero because you get it cheaply.

6

Vision 2020: Computational Needs of the Chemical Industry

T.F. Edgar
University of Texas
D.A. Dixon
Pacific Northwest National Laboratory
and
G.V. Reklaitis
Purdue University

INTRODUCTION

There are a number of forces driving the U.S. chemical industry as it moves into the 21st century, including shareholder return, globalization, efficient use of capital, faster product development, minimizing environmental impact, improved return on investment, improved and more efficient use of research, and efficient use of people. As the chemical industry tries to achieve these goals, it is investigating the expanded use and application of new computational technologies employed in areas such as modeling, computational chemistry, design, control, instrumentation, and operations. The key technology driver over the past 20 years has been the continuing advances in digital computing. The 100-fold increase in computer speed, and the same in software, each decade has led to significant reductions in hardware cost for computers of all types and has increased the scope of applications in chemistry and chemical engineering.

A forecast of future advances in process modeling, control, instrumentation, and optimization is a major part of the recently completed report *Technology Vision 2020: Report of the U.S. Chemical Industry*. This report was sponsored by five major societies and associations (American Institute of Chemical Engineers [AIChE], American Chemical Society [ACS], Council for Chemical Research [CCR], Chemical Manufacturers Association [CMA], and Society of Organic Chemicals Manufacturers Association [SOCMA]) and involved more than 200 business and technical leaders from industry, academia, and government. It presents a road map for the next 20 years for the chemical and allied industries.

The collaboration among the five societies, as well as government agencies (Department of Energy, National Institute of Standards and Technology, National Science Foundation, and U.S. Environmental Protection Agency), has spawned many additional workshops, generating more detailed R&D roadmaps on specific areas of chemical technology. Several workshops pertinent to this paper have been held during 1997 and 1998, covering the areas of instrumentation, control, operations, and computational

chemistry. Other Vision 2020 workshops have been held on subjects such as separations, catalysis, polymers, green chemistry and engineering, and computational fluid dynamics.[1]

This paper reviews the computational needs of the chemical industry as articulated in various Vision 2020 workshops. Subsequent sections of this paper deal with process engineering paradigm in 2020, computational chemistry and molecular modeling, process control and instrumentation, and process operations.

PROCESS ENGINEERING IN 2020

Increased computational speeds have spurred advances in a wide range of areas of transport phenomena, thermodynamics, reaction kinetics, and materials properties and behavior. Fundamental mathematical models are becoming available due to an improved understanding of microscopic and molecular behavior, which could ultimately lead to ab initio process design. This will enable design of a process to yield a product (e.g., a polymer) with a given set of target properties, predictable environmental impact, and minimum costs. Ideally one would want to be able to start with a set of material properties and then reverse-engineer the process chemistry and process design that gives those properties.

Historically the chemical industry has used the following sequential steps to achieve commercialization:

1. Research and development,
2. Scale-up,
3. Design, and
4. Optimization.

Note that steps (1) and (2) generally involve several types of experimentation, such as laboratory discovery, followed by bench-scale experiments (often of a batch nature), and then operation of a continuous flow or batch pilot plant. It is at this level that models can be postulated and unknown parameters can be estimated in order to validate the models. A plant can be designed and then optimized using these models. If the uncertainty in process design is high, pilot-scale testing may involve several generations (sizes) of equipment. With the advent of molecular-scale models for predicting component behavior, some laboratory testing can be obviated in lieu of simulation. This expands upon the traditional relationship of scientific theory and experiment to form a new development/design paradigm of process engineering (see Figure 6.1).

The development of mathematical models that afford a seamless transition from microscopic to macroscopic levels (e.g., a commercial process) is a worthy goal, and much progress in this direction has occurred in the past 10 years in areas such as computational fluid dynamics. However, due to computational limitations and to some extent academic specializations, process engineering research has devolved into four more or less distinct areas:

1. Process design,
2. Structure property relationships,
3. Process control,
4. Process operations.

[1] See <http://www.chem.purdue.edu/v2020/>, the Vision 2020 Web site for workshop reports.

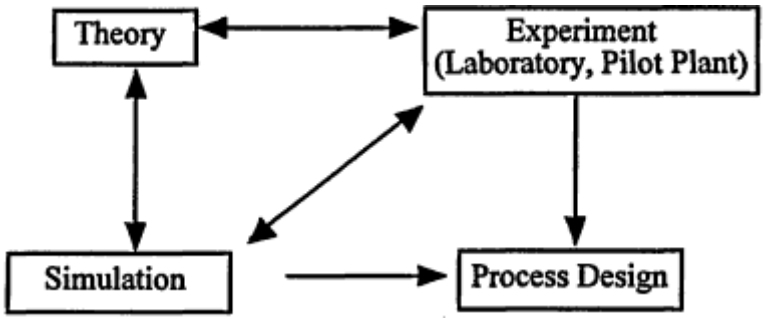


Figure 6.1
Process engineering paradigm for the 21st century

In fact, research conferences will be held during the next 2 years in each of these areas, but only a few hardy souls will participate in cross-fertilizing the areas by attending multiple conferences. Consider the interaction of process design and control; process design decisions can be made that simultaneously optimize plant profitability and the controllability of the plant, rather than the traditional two-step approach of designing the most profitable plant and then considering how to control it in a subsequent design phase. The different models, problem scope, and terminology used in each of these areas is an indicator that no lingua franca has emerged. Actually areas (1), (3), and (4) fall under a broad umbrella of systems technology, but until these three areas begin to use a common set of mathematical models, progress toward a more catholic view of process design will be impeded.

A molecular-level understanding of chemical manufacturing processes would greatly enhance the ability of chemical engineers to optimize process design and operations as well as ensure adequate protection of the environment and safe operating conditions. Currently there is considerable uncertainty in thermodynamic and reaction models, so plants are normally overdesigned (above required capacity) to allow for this uncertainty. Also plants are operated conservatively because of an inadequate understanding of dynamic process behavior and the dire consequences if an unsafe condition arises. Chemical reactors are at the heart of this issue, with uncertainties in kinetic mechanisms and rate constants and the effects of reactor geometry (such as catalyst beds) on heat and mass transfer. Clearly the availability of better microscopic mathematical models for macroscopic plant simulation will help the chemical industry operate more profitability and more reliably in the future.

Besides providing fundamental data for process simulations, computational chemistry plays an important role in the molecular design process beginning at the basic research level. By predicting accurate thermochemistry, one can quickly scope out the feasibility of reaction pathways as to whether a reaction is allowed or not. Computational chemistry can also reliably predict a wide range of spectroscopic properties to aid in the identification of chemical species, especially important reaction intermediates. Electronic structure calculations can also provide quantitative insights into bonding, orbital energies, and form, facilitating the design of new molecules with the appropriate reactivity.

COMPUTATIONAL CHEMISTRY AND MOLECULAR MODELING

The computational chemistry subgroup of Vision 2020 under the sponsorship of the CCR has outlined a set of computational "grand challenges" or "technology bundles" that will have a dramatic impact on the practice of chemistry throughout the chemical enterprise, especially the chemical industry. The computational "grand challenges" are given in Box 6.1.

BOX 6.1 COMPUTATIONAL "GRAND CHALLENGES" FOR MATERIALS AND PROCESS DESIGN IN THE CHEMICAL ENTERPRISE

A. Reliable prediction of biological activity from chemical structure
B. Reliable prediction of environmental fate from chemical structure
C. Design of efficient catalysts for chemical processes
D. Design of efficient processes in chemical plants from an understanding of microscopic molecular behavior
E. Design of a material with a given set of target properties

"Grand challenge A" or "bundle A" in Box 6.1 has received recent emphasis because this area includes drug design. However, the biological activity due to a specific chemical is needed for other areas such as agricultural pesticide design and predictive toxicology. The potential for toxic impact of any chemical must be addressed before a chemical is manufactured, sold to the public, or released to the environment. Furthermore, the toxic behavior must be evaluated not only for human health issues but also for its potential ecological impact on plants and animals. Examining chemical toxicity is currently an extremely expensive process that can take a number of years of detailed testing. Such evaluations usually occur late in development, and the inability to anticipate the evaluation of toxicological testing can place large R&D investments at risk. Also, the possibility exists that unanticipated toxicological problems with intermediates and by-products can create liabilities. The cost of toxicology testing is generally too high to complete testing early in the development process. Thus reliable, cost-effective means for predicting toxicological behavior would be of great benefit to the industry.

Grand challenge B in Box 6.1 is focused on the need to predict the fate of any compound that is released into the environment. For example, even if a compound is not toxic, a degradation product may show toxic behavior. Besides being toxic to various organisms, chemicals released into the environment can affect it in other ways. A difficulty in dealing with the environmental impact of a chemical is that the temporal and spatial scales cover many orders of magnitude from picoseconds to 100,000 years in time, and from angstroms to thousands of kilometers in distance. Furthermore, the chemistry can be extremely complex and the chemistry that occurs on different scales may be coupled. For example, chemical reactions that occur on a surface may be influenced not only by the local site but also by distant sites that affect the local electronic structure or the surrounding medium.

Grand challenges C and D in Box 6.1 are tightly coupled but are separated here because different computational aspects may be needed to address these areas. Catalysis and catalytic processes are involved in manufacturing most petroleum and chemical products and account for nearly 20 percent of the U.S. gross domestic product. Improved catalysts would increase efficiency, leading to reduced energy requirements, while increasing product selectivity and concomitantly decreasing wastes and emissions. Considerable effort has been devoted to the ab initio design of catalysts, but such work is difficult because of the types of atoms involved (often transition metals) and because of the fact that extended surfaces are often involved. Besides the complexity of the materials themselves, an additional requirement is the need for accurate results. Although computational results can often provide insight into how a catalyst works, the true design of a catalyst will require the ability to predict accurate thermodynamic and kinetic results. For example, a factor of two to four in catalyst efficiency can

determine the economic feasibility of a process. Such accuracies mean that thermodynamic quantities should be predicted to within 0.1 to 0.2 kcal/mol and rate constants to within ~ 15 percent—certainly difficult, if not impossible, by today's standards. Even for the nominally simple area of acid/base catalysis, many additional features may have to be included in the model, for example, the effects of solvation.

Another example of complexity is found in zeolites, where the sheer size of the active region makes modeling studies difficult. Modeling of the surfaces present in heterogeneous catalysts is even more challenging because of the large numbers of atoms involved and the wide range of potential reactive sites. If the catalyst contains transition metals, the modeling task is difficult because of the problems in the treatment of electronic structures of such systems with single-configuration wave functions in a molecular orbital framework.

A molecular-level understanding of chemical manufacturing processes would greatly aid the development of steady-state and dynamic models of these processes. As discussed in subsequent sections, process modeling is extensively practiced by the chemical industry in order to optimize chemical processes. However, one needs to be able to develop a model of the process and then predict not only thermochemical and thermophysical properties but also accurate rate constants as input data for the process simulation. Another critical set of data needed for the models are thermophysical properties. These include such simple quantities as boiling points and also more complex phenomena such as vapor/liquid equilibria phase diagrams, diffusion, liquid densities, and the prediction of critical points. The complexity of process simulations depends on whether a static or dynamic simulation is used and whether effects such as fluid flow and mass transfer are included. Examples of complex phenomena that are just now being considered include the effects of turbulence and chaotic dynamics on the reactor system. A key role of computational chemistry is to provide input parameters of increasing accuracy and reliability to the process simulations.

Grand challenge E in Box 6.1 is extremely difficult to treat at the present time. Given a structure, we can often predict at some level what the properties of the material are likely to be. The accuracy of the results and the methods used to treat them depend critically on the complexity of the structure as well as the availability of information on similar structures. For example, various quantitative structure property relationship (QSPR) models are available for the prediction of polymer properties. However, the inverse engineering design problem, designing structures given a set of desired properties, is far more difficult. The market may demand or need a new material with a specific set of properties, yet given the properties it is extremely difficult to know which monomers to put together to make a polymer and what molecular weight the polymer should have. Today the inverse design problem is attacked empirically by the synthetic chemist with his/her wealth of knowledge based on intuition and on experience. A significant amount of work is already under way to develop the "holy grail" of materials design, namely, effective and powerful reverse-engineering software to solve the problem of going backwards from a set of desired properties to realistic chemical structures and material morphologies that may have these properties. These efforts are usually based on artificial intelligence techniques and have, so far, had only limited success. Much work needs to be done before this approach reaches the point of being used routinely and with confidence by the chemical industry.

The achievement of the goals outlined in Box 6.1 will require significant advances in a number of science and technology areas. Box 6.2 summarizes the important scientific research areas needed to accomplish the goals outlined in Box 6.1, and Box 6.3 summarizes technical issues that need to be addressed. Below we highlight some of these issues.

There are a number of methods for obtaining accurate molecular properties. One can now push the thermochemical accuracy to about 0.5 kcal/mol if effects such as the proper zero-point energy, core/

BOX 6.2 RESEARCH AREAS FOR IMPLEMENTATION OF GRAND CHALLENGES

1. Accurate methods for calculating thermochemical and thermophysical properties, spectroscopy, and kinetics (A,B,C,D,E)[1]
2. Efficient methods for generating accurate potential functions for molecular mechanics-based methods (A,B,C,D,E)
3. Improved methods for molecular dynamics simulations at long times for large ensembles (A,B,C,D,E)
4. Improved methods for including quantum effects (A,B,C,D,E)
5. Improved methods for including environmental effects such as solvent effects (A,B,C,D,E)
6. Efficient and accurate computational methods for treating solid state structures (B,C,D,E)
7. Improved optimization strategies for the determination of large, complex structures such as predicting protein structure from sequence (A,B,C,D,E)
8. Accurate methods for treating the scaring-up problem: molecular ⇒ microscopic ⇒ mesoscopic ⇒ macroscopic (A,B,C,D,E)
9. New techniques for materials design and bulk property prediction (E)
10. New methods for predictive toxicology (A,B)
11. Integration of computational fluid dynamics (including lattice-Boltzmann approaches) with physics, chemistry, and biology to predict the behavior of reacting flows at different spatial and temporal scales[2] (B,D)

[1] Impact on grand challenge from Box 6.1 given in parentheses.
[2] Additional research area for implementation of grand challenges.

BOX 6.3 TECHNOLOGY NEEDS FOR IMPLEMENTATION OF GRAND CHALLENGES

1. High-performance, scalable, portable computer codes for advanced (massively parallel) computer architectures (A,B,C,D,E)[1]
2. Improved problem-solving environments (PSEs) to make computational tools more widely accessible (A,B,C,D,E)
3. Improved database and data-analysis technologies (A,B,C,D,E)
4. Computer-aided synthesis methods with a focus on materials (E)
5. Computer architectures, operating systems, and networks (A,B,C,D,E)

[1] Impact on grand challenge from Box 6.1 given in parentheses.

valence effects, and relativistic effects are considered. Predicting kinetics can be considered as an extension of thermochemical calculations if one uses variational transition-state theory. Instead of just needing an optimized geometry and calculated second derivatives at one point on the potential energy surface, this information is required at up to hundreds of points. It is necessary to incorporate solvent effects in order to predict reaction rate constants in solution. The prediction of rate constants is critical for process and environmental models. Predicted rate constants (computational kinetics) have already found use in such complex systems as atmospheric chemistry, design of chemical vapor deposition reactors, chemical plant design, and combustion models. Spectroscopic predictions are increasing in their accuracy, but it is still difficult to predict NMR chemical shifts to better than a few parts per million, vibrational frequencies to a few cm^{-1}, or electronic transitions to a few tenths of an electron volt for a broad range of complex chemicals.

There is a real need for accurate methods for predicting accurate thermophysics for gases and liquids. For gases, certain properties can be predicted with reasonable reliability based on the interaction potentials of molecular dimers and transport theory. For liquids, such properties can be predicted by using molecular dynamics and grand canonical Monte Carlo (GCMC) simulations. The GCMC simulations are quite reliable for some properties for some compounds, but they are very dependent on the quality of the empirical potential functions. Such predictions, today, are much less reliable for mixtures or if ions are present.

The whole area of potential functions needs to be carefully addressed. Potential functions are needed for all atomistic simulations, (e.g., molecular dynamics and energy minimizations of materials, polymers, solutions, and proteins), Monte Carlo methods, and Brownian dynamics. However, reliable potential functions are not available for all atoms and all bond types or for a wide range of properties such as polarization due to the medium. At present, it is very time-consuming to construct potential functions. A robust, automated potential function generator for producing a polarizable force field for all atom types needs to be developed. It needs to be able to incorporate both the results of quantum mechanical calculations and the empirical data.

There is a critical need to be able to take atomistic simulations such as molecular dynamics to much longer time scales. At present, it is routinely possible to study atomistic systems (or systems represented as interacting atoms, such as proteins and polymeric systems) for periods on the order of nanoseconds. However, much longer time scales are needed for the study of such problems as phase transitions, rare events, kinetics, and long time protein dynamics for protein folding. Even today, long runs on current computing systems create as-yet-unresolved data issues due to massive amounts of data generated. For example, a single time step of a million-atom simulation easily manipulates tens of megabytes of data. While a reasonable strategy for short simulations of small systems is to dump configurations every 10th or 50th time step for later analysis, this is clearly not an option for large-scale simulations over long time frames. Methodologies for implementing and modifying data analysis "on-the-fly" must be developed and refined.

The question of reaching macroscopic time scales from molecular dynamics simulations cannot be solved solely by increases in hardware capacity, since there are fundamental limitations on how many time steps can be executed per second on a computer, whether parallel or serial. One can scale the size of the problem with increasing numbers of processors, but not to longer times. To cover macroscopic time scales measured in seconds while following molecular dynamics time steps of 10^{-15} seconds requires the execution of on the order of 10^{15} time steps. Even with five-orders-of-magnitude increases in clock cycles, the required computations will take days. Between now and 2020, clock rates will undoubtedly increase, but not by this magnitude. Hence, the long time problem in molecular dynamics will not be solved purely by hardware improvements. The key is the development of theoretically

sound, time-coarsening methodologies that will permit dynamics-based methods to traverse long time scales. Brownian dynamics with molecular-dynamics-sampled interactions and dynamic Monte Carlo methods are promising possibilities for this purpose.

A technology issue that will have an enormous impact on computational chemistry is that of computer architectures, operating systems, and networks. The highest performance today is pushing 1.0 teraflops of sustainable performance on highly tuned code. The biggest technical issue is how to deal with nonuniform memory access and the associated latency for data transfer between memory on distributed processors. The latest step for large-scale computers is massively parallel computing systems based on symmetric multi-processors (SMPs). The goal is tens-of-petaflops performance by 2020. This will be achieved by improvements in the speeds of individual chips, which have been doubling every 18 months, although the cost of building plants to produce them may lead to a lengthening of the time to double processor speed. There will have to be significant improvements in switches as well as in memory speeds, and I/O devices (disks) will need to be much faster and cheaper. There is a real need for significant advances in application software for usable teraflops to petaflops performance to be achieved as well as improvements in operating systems (OSs). One major issue will be the need for single-threaded OSs that are fault-tolerant, as the reliability of any single processor means that some will fail on a given day. It is the issue of operating systems, especially for large-scale batch computing, that is likely to hold up the ability to broadly address the computational grand challenge issues raised above.

In summary, rapid advances on many fronts suggest that we will be able to address the complex computational grand challenges outlined above. This will fundamentally change how we will do chemistry in the future in research, in development, and in production. Getting there will not be simple and will require novel approaches, including the use of teams from a range of disciplines to develop the software, manage the computer systems, and perform the research.

PROCESS CONTROL AND INSTRUMENTATION

The process control and instrumentation issues identified in Vision 2020 include changes in the way plants operate, computer hardware improvements, the merging of models for design, operations, and control, development of new sensors, integration of measurement and control, and developments in advanced control. In the factory of the future, the industrial environment where process control is carded out will be different than it is today. In fact, some forward-thinking companies believe that the operator in the factory of the future may need to be an engineer, as is the case in Europe. Because of greater integration of the plant equipment, tighter quality specification, and more emphasis on maximum profitability while maintaining safe operating conditions, the importance of process control will increase. Very sophisticated computer-based tools will be at the disposal of plant personnel. Controllers will be self-tuning, operating conditions will be optimized frequently, fault detection algorithms will deal with abnormal events, total plant control will be implemented using a hierarchical (distributed) multivariable strategy, and expert systems will help the plant engineer make intelligent decisions (those he or she can be trusted to make). Plant data will be analyzed continuously and will be reconciled using material and energy balances and nonlinear programming, and unmeasured variables will be reconstructed using parameter estimation techniques. Digital instrumentation will be more reliable and will be self-calibrating, and composition measurements that were heretofore not available will be measured online. There are many industrial plants that have already incorporated several of these ideas, but no plant has reached the highest level of sophistication over the total spectrum of control activities.

We are now beginning to see a new stage in the evolution of plant information and control architectures. Over the last 20 years, progress in computer control has been spurred by acceptance across a wide

spectrum of vendors of the distributed control hub system for process control, which was pioneered during the 1970s by Honeywell. A distributed control system (DCS) employs a hierarchy of computers, with a single microcomputer controlling 8 to 16 individual control loops. More detailed calculations are performed using workstations, which receive information from the lower-level devices. Set points, often determined by real-time optimization, are sent from the higher level to the lower level. With the focus now on enterprise integration, automation vendors are now implementing Windows NT as the new solution for process control, utilizing personal computers in client-server architectures rather than the hub-centric approach used for the past 20 years. This promotes an open application environment (open control systems) and makes accessible the wide variety of PC object-oriented software tools (e.g., browsers) that are now available.

The demand for smart field devices is rising rapidly. It is desirable to be able to query a remote instrument and determine if the instrument is functioning properly. Of course digital-based rather than analog instruments have the key advantage that signals can be transmitted digitally (even by wireless) without the normal degradation experienced with analog instruments. In addition, smart instruments have the ability to perform self-calibration and fault detection/diagnosis. Smart valves include proportional-integral-derivative (PID) control resident in the instrument that can permit the central computers to do more advanced process control and information management. It is projected that installations of smart instruments can reduce instrumentation costs by up to 30 percent over conventional approaches. There has been much recent activity in defining standards for the digital, multidrop (connection) communications protocol between sensors, actuators, and controllers. In the United States, the concept is called "fieldbus control," and vendors and users have been working together to develop and test interoperability standards via several commercial implementations.

When data become readily available at a central point, it will be easier to apply advanced advisory systems (e.g., expert systems) to monitor the plant for performance as well as detect and diagnose faults. Recent efforts have built on the traditional single variable statistical process control approach and extended it to multivariable problems (many process variables and sensors) using multivariate statistics and such tools as principal component analysis. These techniques can be used for sensor validation to determine if a given sensor has failed or exhibits bias, drift, or lack of precision.

In the area of process modeling, industrial groups are beginning to examine whether it is possible to achieve a seamless transition between models used for flow-sheet design and simulation and models used for control. The CAPE OPEN industrial consortium in Europe and other groups in the United States are working toward an open architecture for commercial simulators to achieve "plug and play" using company-specific software such as physical property packages. The extension of these steady-state flow-sheet simulators to handle dynamic cases is now becoming an active area (e.g., linking Aspenplus to Speedup). The goal is to have models for real-time control that run at 50 to 500 times real-time, but this will require increased computational efficiency and perhaps application of parallel computing.

A new generation of model-based control theory has emerged during the past decade that is tailored to the successful operation of modern plants, addressing the "difficult" process characteristics encountered in chemical plants shown in Box 6.4. These advanced algorithms include model predictive control (MPC), robust control, and adaptive control, where a mathematical model is explicit in developing a control strategy. In MPC, control actions are obtained from online optimization (usually by solving a quadratic program), which handles process variable constraints. MPC also unifies treatment of load and set-point changes via the use of disturbance models and the Kalman filter. MPC can be extended to handle nonlinear models, as shown in Figure 6.2.

The success of MPC in solving large multivariable industrial control problems is impressive. Model

BOX 6.4 PROCESS CHARACTERISTICS THAT MUST BE TREATED BY ADVANCED CONTROL

- Time delays
- Nonminimum phase disturbances
- Unmeasured variables
- Noise
- Time-varying parameters
- Nonlinearities
- Constraints
- Multivariable interactions

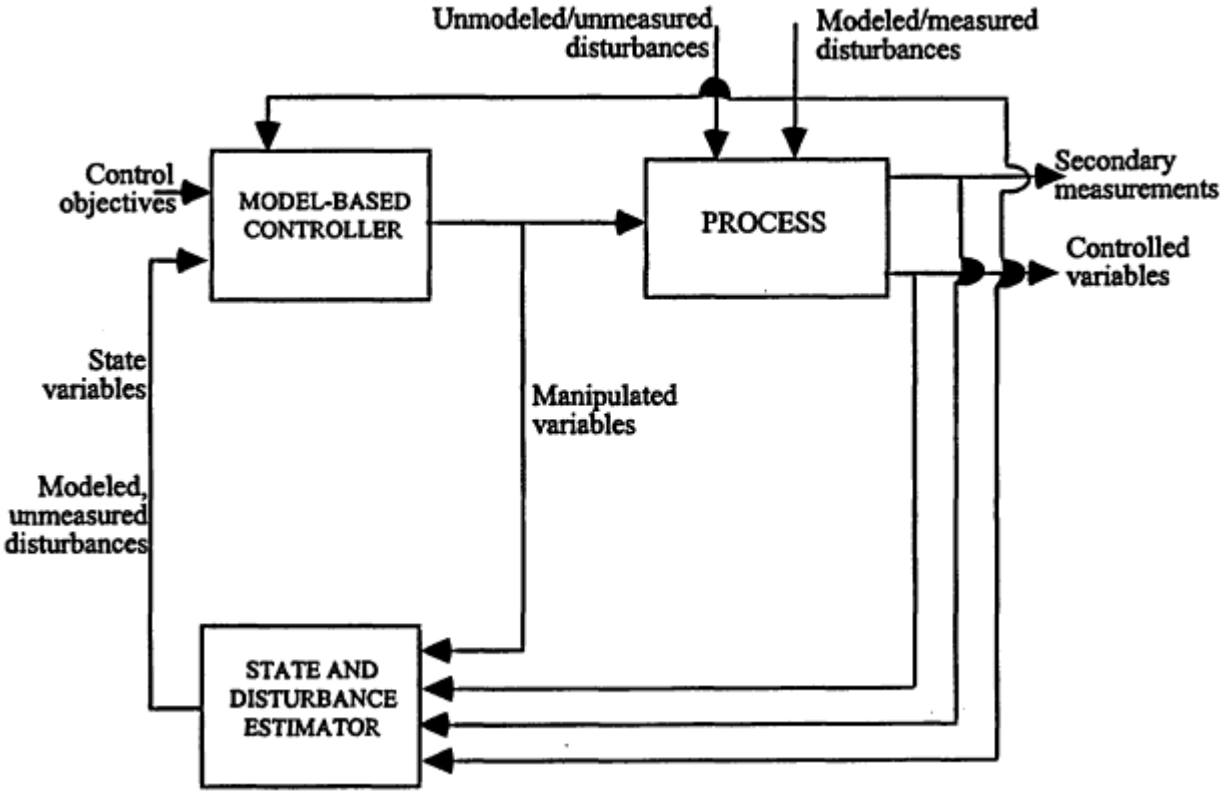


Figure 6.2
Generalized block diagram for model predictive control.

predictive control of units with as many as 10 inputs and 10 outputs is already established in industrial practice. Computing power is not causing a critical bottleneck in process control, but larger MPC implementations and faster sample rates will probably accompany faster computing. Improved algorithms could easily have more impact than the improved hardware for the next several years. MPC will appear at the lowest level in the DCS, which will reduce the number of PID loops implemented.

Adaptive control implies that the controller parameters should be adapted in real-time to yield optimal performance at all times; this is often done by comparing model predictions with online plant data and updating the process model parameters. The use of nonlinear models and controllers is under way in some applications. Some of the new versions of MPC are incorporating model adaptation, but so far adaptive control has not had much impact. This is due to problems in keeping such loops operational, largely because of the sensitivity of multivariable adaptive controllers to model mismatch.

Recent announcements by software vendors indicate that the combination of process simulation, optimization, and control into one software package will be a near-term reality, i.e., a set of consistent models across R&D, engineering, and production stages, with increased emphasis on rigorous dynamic models and the best control solutions. Software users will be able to optimize plant-wide operations using real-time data and current economic objectives. Future software will determine the location and cause of operating problems and provides a unified framework for data reconciliation and parameter estimation in real time.

There are still many questions to be answered regarding the connection between modeling and control. This includes the explicit modeling information needed to achieve a particular level of control performance, the fundamental limitations on control performance even for perfect models, and the trade-offs between modeling accuracy, control performance, and stability.

Process Measurement and Control Workshop

In recognition of the needs and challenges in the areas of process measurement and control, a workshop entitled “Process Measurement and Control: Industry Needs" was convened in New Orleans, March 6-8, 1998.[2] The goals of the workshop were as follows:

1. To survey the current state of the art in academic research and industrial practice in the areas of measurement and control, particularly as they apply to the chemical and processing industries. The extent of integration of measurements with control is a particular focus of the survey.
2. To identify major impediments to further progress in the field and the adoption of these methods by industry; and
3. To determine highly promising new directions for methodological developments and application areas.

The workshop emphasized future development and application in eight areas:[3]

- Molecular Characterization and Separations,
- Nonlinear Model Predictive Control,
- Information and Data Handling,
- Controller Performance Monitoring,

[2] Material from the workshop will appear in Vol. 23, Issue No. 2 (1999) of *Computers and Chemical Engineering*.
[3] See <http://fourier.che.udel.edu/~doyle/V2020/Index.html> for further information on workshop findings.

- Sensors,
- Estimation and Inferential Control,
- Microfabricated Instrumentation Systems, and
- Adaptive Control and Identification.

As an example of a specific road map, the second topic (nonlinear model predictive control [NMPC]) has been mainly of academic interest so far, with a few industrial applications involving neural networks. What is needed is an analysis tool to determine the appropriate technology (NMPC vs. MPC) based on the process description, performance objective, and operating region. There is also a desire to represent complex physical systems so that they are more amenable to optimization-based (and model-based) control methods. The improved modeling paradigms should address model reduction techniques, low-order physical modeling approaches, maintenance of complex models, and how common model attributes contribute pathological features to the corresponding optimization problem. Hybrid modeling, which combines fundamental and empirical models, and methodologies for development of nonlinear models (e.g., input sequence design, model structure selection, parameter adaptation) deserve attention. More details are contained in the Web site for this workshop.

Chemical Instrumentation

Chemical analysis is a critically important enabling technology essential to every phase of chemical science, product and process development, and manufacturing control. Advances in chemical measurement over the past two decades have greatly accelerated progress in chemical science, biotechnology, materials science, and process engineering. Chemical measurements also play a key role in numerous related industries, such as pharmaceuticals and pulp, paper, and food processing. During recent years, impressive advances have been made in the resolution, sensitivity, and specificity of chemical analysis. The conduct of analytical chemistry has been transformed by advances in high-field superconducting magnets, multiple-wavelength lasers, multiplexed array detectors, atomic-force microscopes, scanning spectral analysis, and the integration of computers with instrumentation. These methods have been extended to the detection and spectral characterization of molecular structure at the atomic level.

A Vision 2020 workshop was held in March, 1997, to assess future directions for R&D in chemical instrumentation. Research needs identified included:[4]

- Transfer of analytical laboratory capabilities into plants, incorporating ease of maintenance and support, utilizing new technology and molecular-scale devices;
- Improved real-time characterization of polymers (molecular weight distribution, branching);
- Improved structure/property/processing modeling capability, especially macromolecular products such as biomolecules and biopolymers;
- Physical/chemical characterization of solids and slurries;
- Online characterization of biotechnological processes;
- New approaches for sampling and system interlinks to control and information systems;
- Self-calibrating and self-diagnostic (smart) sensors;
- Identification of processes needing microfabricated instruments and development of corresponding models/control systems;

[4] For more details see <http:www.nist.gov/cstl/hndocs/ExternalTechnologyBundles.html>

- Integration of data from multiple sensors for environmental compliance, product development, and process control, including soft sensors; and
- Advanced measurement techniques to support combinatorial chemistry in catalysis and drug discovery.

PROCESS OPERATIONS

Three of the four technology thrust areas of the Vision 2020 document—namely, supply chain management, information systems, and manufacturing and operations—address the business and manufacturing functions of the chemical enterprise. This clearly reflects the importance that efficient production and distribution of chemical products have on the economic viability of the enterprise now and over the next 25 years. In this section, we highlight the role that technical computing and information systems play as technology enablers for effective operation and present the most important challenges and needs that must be addressed in the future. The discussion of research issues draws on "R&D Needs in Systems Technologies for Process Operations," a workshop that was convened in July 1998. For full details of the workshop report the reader is invited to consult the Vision 2020 Web site.[5]

In the present context, "process operations" refers to the management and use of human, capital, material, energy, and information resources to produce desired chemical products safely, flexibly, reliably, cost-effectively, and responsibly for the environment and community. The traditional scope of operations encompasses the plant and its associated decision levels, as shown in Figure 6.3.

The key information sources for the plant operational decision hierarchy are the enterprise data, consisting of commercial and financial information, and the process itself. The unit management level includes the process control, monitoring and diagnosis, and online data acquisition functions. The plant-wide management level serves to coordinate the network of process units and to provide cost-effective set points via real-time optimization. The scheduling decision layer addresses time-varying capacity and manpower utilization decisions, while the planning level sets production goals that meet supply and logistics constraints. Ideally there is bi-directional communication between levels, with higher levels setting goals for lower levels and the lower levels communicating constraints and performance information to the higher levels. In practice the information flow tends to be top-down, invariably resulting in mismatches between goals and their realization.

In recent years this traditional view of operations has been expanded to include the interactions between suppliers, multiple plant sites, distribution sites and transportation networks, and customers. The planning and management of this expanded network, referred to as the supply chain, pose challenging decision problems because of the wide temporal scale and dynamics of the events that must be considered, the broad spatial distribution and dimensions of the entities that must be managed, and the high degree of uncertainty due to changing market factors and variable facilities uptimes and productivity. Clearly the supply chain is a highly complex dynamic system. Nonetheless, the vision proposed for the operational domain is that in 2020 the success of a chemical enterprise will depend upon how effectively it generates value by dynamically optimizing the deployment of its supply chain resources. The seven factors critical to the achievement of the vision are:

- Speed to market—time from piloting to the market place;
- Efficient operation in terms of operational cost and asset utilization;

[5] See <http://www.chem.purdue.edu/v2020/>, the Vision 2020 Web site for workshop reports.

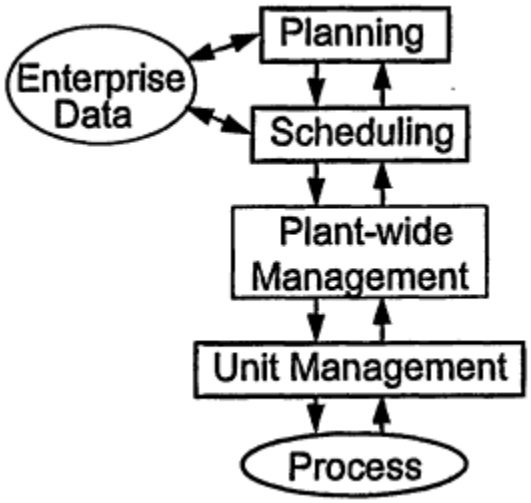


Figure 6.3
Plant decision hierarchy.

- Health, environment, and safety: factors affecting workers and the community;
- Quality work force—management, engineering, and process and business literate operational staff;
- Technology infrastructure—processes, instrumentation, and equipment as well as information systems and technical computing;
- Quality and product integrity: work process for producing the product right the first time; and
- Functional integration: bi-directional linkage of all decision levels of the supply chain.

To allow the vision of the dynamically optimized supply chain to be realized under each of these factors, innovations extending beyond developments in information and computing technology alone are required. However, it is clear that the infrastructure for storing and sharing information and technical computing tools that exploit such information constitute the key enabling technology. The information that must be stored and shared includes transactional information, resources costs and availabilities, plant status information, models, and model solutions. This diversity of information types must be effectively organized and must be sharable using reliable high-speed networks. The enabling technical computing components include model-building methods and tools; solution algorithms using numerical, symbolic, and logic-based methods; visualization and interpretation methods; interfaces for use and training; and integration of all of these components into usable decision support tools.

PRESENT STATUS

At present, the essential elements of information technology to support operations are at hand, in terms of both data infrastructure and network connectivity. Commercial database management systems and transactional systems are common in the industry. Plant information systems and historians are in widespread use, and enterprise-wide database system installations are growing explosively. UNIX- or Windows NT-based networks are common, and Internet and Web-based applications are growing rapidly. Despite this growth there is only limited integration of business and manufacturing data and

tools to facilitate effective use of this data. Indeed, the general consensus is that corporations are drowning in a sea of data. The challenge is to extract information and knowledge and thus to derive value from this data.

The present status of technical computing of relevance to process operations can best be characterized as a patchwork of areas at different levels of development. At the planning level, multi-time-period linear programming tools capable of handling large-scale systems are well developed and have been in use, especially in the petroleum/petrochemical sector, since the 1970s. Real-time, plant-wide optimization applications using steady-state process models are growing rapidly in the petrochemical domain, although some of the statistical and computational formulations and algorithms remain under active development. The methodology for scheduling of multipurpose batch and continuous production facilities has been under investigation since the late 1970s, initially using rule-based and heuristic randomized search methods and more recently using optimization-based (mixed integer linear programming) methods. Application of the latter in industry is limited but growing. Successful solutions of problems involving more than one hundred equipment items and several hundred distinct production tasks have been reported, although the deployment of the technology still requires high levels of expertise and effort. As noted in the section above titled "Process Control and Instrumentation," tools for abnormal situation management are in their infancy, although significant industry-led developments are in progress. Linear model predictive control has been practiced in the field since the early 1980s, although the theoretical supports for the methodology were developed later. Plant data rectification has been practiced since the mid-1980s, but typically applications have been confined to linear models and simple statistical descriptions of the errors in the measurements.

CHALLENGES

The long-term challenges for the application of computing technology can be divided into four major areas:

- Conversion of data into knowledge,
- Support tools for the process,
- Support tools for the business, and
- Training methodologies.

The development of tools that would facilitate conversion of the extensive data contained in enterprise information systems into actionable information and ultimately knowledge is of highest priority. Some of the capabilities that need to be pursued include soft-sensors, data rectification techniques, trend analysis and monitoring methods, and data visualization techniques. Soft-sensors are critical to simplifying the detection of erroneous measurements by localizing the detection logic. Data rectification refers to the process of condensing and correcting redundant and inaccurate or erroneous process data so as to obtain the most likely status of the plant. Trend analysis and monitoring refers to the process of using process knowledge and models to identify and characterize process trends so as to provide timely predictions of when and what corrective action needs to be taken. Data visualization is an essential element for facilitating understanding of process behavior and tendencies.

The decision support tools for the process include streamlined modeling methodology, multi-view systems for abnormal situation management, nonlinear and adaptive model predictive control, and process optimization using dynamic, and especially hybrid, models. Model building is generally perceived to be a key stumbling block because of the level of expertise required both to formulate process

models and to implement them using contemporary tools. The goal is to make model building and management rapid and reliable and to create environments in which the models associated with the various levels of the operational decision hierarchy will be consistent and unified. The role of abnormal situation management systems is to identify plant trends, to diagnose likely causes and consequences, and to provide intelligent advice to plant personnel. While components that address portions of this entire process have been under investigation for the past decade, full integration of the various qualitative and quantitative support tools remains to be realized. Needed developments in process control have been discussed in an earlier session and hence will not be reiterated here, except to note that control of batch and other intentionally dynamic processes needs to be given considerably more attention. Finally, the optimization of models consisting of differential algebraic systems, and especially differential algebraic systems with discrete elements, is essential to the realization of the vision for process operations. The latter type of so-called hybrid systems is particularly relevant to processes that involve batch and semicontinuous operations.

The overall goal of these decision support methodologies for the process is to realize the integrated model-centered paradigm for process operation shown in Figure 6.4. Under this paradigm all of the

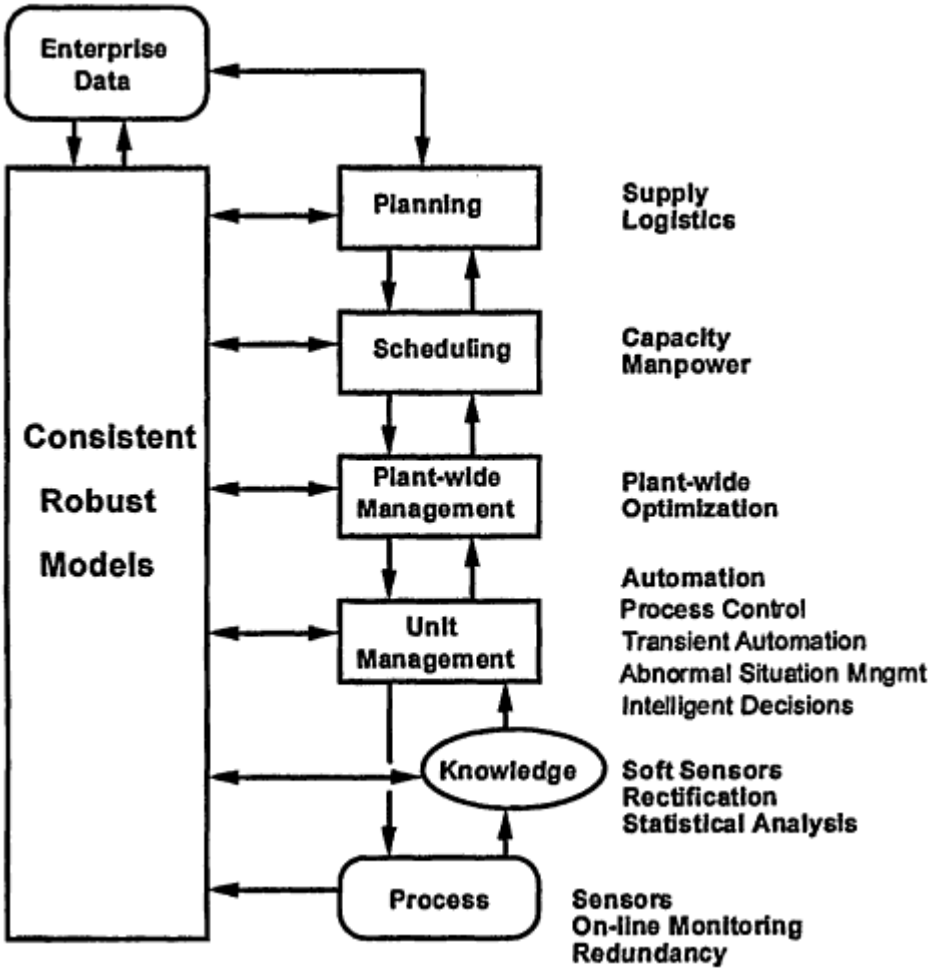


Figure 6.4
Integrated model-centered operation.

decision levels of the operational hierarchy are fully integrated through the shared use of consistent, robust models. Models serve as the central repository of process knowledge. Information flows from the lower levels to the higher levels to ensure that decisions fully consistent with the status and capacity of the production resources are made.

The third area of need is in the development of tools to support the overall business decision processes. The objective is to expand the envelope beyond the process itself and to encompass the business processes that are essential to driving manufacturing and the entire supply chain. The tools include improved sales and market forecasting methodologies, supply and logistics planning techniques, methodologies for quantitative risk assessment, optimization-based plant scheduling methods, business modeling frameworks, and approaches to dynamic supply chain optimization. Optimization-based scheduling requires the solution of very high dimensionality models expressed in terms of discrete 0-1 variables. The key need is to be able to solve scheduling problems with hundreds of thousands of such variables reliably and quickly. Such capabilities need to be extended to allow treatment of models that encompass the entire supply chain and to quantitatively address business issues such as resource and capital planning associated with the supply chain, siting of new products, and the impact of mergers and acquisition on the supply chain.

Finally, in order to realize the benefits of the developments in the other three areas, it is necessary, indeed essential, to create training methodologies for the work force. These computer-based training methodologies must make efficient use of students' time, recognize differences in levels of expertise, and employ extensive visualization tools, including virtual reality components. Methods must also be developed to aid process staff in the understanding of models and the meaning of the solutions resulting from the various decision support tools that are based on these models. Such understanding is critical both to the initial adoption of such models and to the continuous improvement process, as it is only from understanding the constraints of the existing operation and their implications that cost-effective improvements can be systematically generated.

In conclusion, the process-operations-specific information systems and technical computing developments outlined above are essential to the realization of the goal of the dynamically optimized supply chain. Continuing increases in computing power, network bandwidth, and availability of faster and cheaper memory will no doubt facilitate achievement of this goal. However, the scope and complexity of the underlying decision problems require methodological developments that offer effective gains orders of magnitude beyond the likely increases in raw computing power and communication bandwidth. Process-oriented technical computing really does play the pivotal role in the future of process operations.

Session 2

Panel Discussion

Robert Lichter, Camille & Henry Dreyfus Foundation: During this joint presentation, all of you in one way or another talked about the way in which people will be an integral part of the whole revolution by 2020. I would like to hear your comments on how we will get there. That is, what are the implications for the education of the people who would be part of that revolution that would presumably begin now?

David Dixon: I will try and make a stab at it, although I am not in the formal education business these days. Much of what we are going to see will actually be on-the-job training. When I was at DuPont, for example, we did not see the broad training you need to actually solve some of these problems in new staff coming directly out of universities.

In the future, we are going to be seeing teams of people working together. We are going to have to change fundamentally the concept of how individual research is done. If we want to solve complex software issues and things like that, we need to look at team approaches.

We have been developing very complex parallel code at PNNL, and the only way we have been able to do it is by putting teams of applied mathematicians, computer scientists, and chemistry domain specialists and users together to solve the problem. Fundamentally, academics will have to change to allow teaming to happen and for that to be a profitable part of the university curriculum.

David Smith, DuPont: Can I just try to add to that incrementally? The issue seems to me that the university must and will continue to rely on the individual contribution in the thesis area, and that is an essential part of getting your Ph.D. On the other hand, when you come to work at a company like DuPont, and I am sure, Dow, teamwork is important, and I think that most of the exciting work that is going on in our corporation today is not so much in any one particular field, but in the intersection between fields where interesting things are happening. So, the ability to have people from different disciplines work together and to be able to understand one another's language is really an important part of the process, and it takes time, and that time comes from DuPont's time. I do not think that there is any way the university can give that to us.

Gintaris Reklaitis: I would like to add to that point from the perspective of the software tools. The question that I have is, How do you teach people to use the computing tools that are getting increasingly more complex? I really think we need a lot of new ideas in this area. I am struck by the fact that we have tools like process simulators that are very time consuming to master. If you look at the user's manual for complex software tools, the first thing the user is asked to do is to spend 2 hours going through all the menus and clicking on all of the commands. Have you tried this with a group of undergraduates? It lasts about 5 minutes. This is absolutely not the way anyone wants to learn to use a software tool. We have to come up with intelligent ways around this training problem.

Specifically, I think that there is a lot of creative work that we need to do with our education colleagues to devise new models for learning and intelligently using complex software tools that have many options and possibilities. This is particularly relevant for the casual user, which is what most industrial users are. Most engineers are focused on projects rather than tools and will only revisit the tools periodically as the need arises. Such a practitioner surely does not want and does not have the time to reinvestigate all of the menus every time to recall what is available.

Evelyn Goldfield, Wayne State University: I want to follow up on the previous question because I think that a lot of universities, including my own, are rethinking some of the ways that we train graduate students and really do want to focus on some of these interdisciplinary team approaches, particularly in computational science. We found that there were a lot of different people at the university who should be and could be training together and working together but until recently were almost totally isolated. There tends to be duplication of effort, of reinventing the wheel. Also, too many students are not taking certain courses that would benefit them because these courses are offered in a different department or a different part of the university, whereas they are all using basically the same algorithms. And so, I wondered if you had any comments for how that these sorts of team and interdisciplinary approaches at the university would have any beneficial effects on your projects?

Gregory McRae, Massachusetts Institute of Technology: I would like to address that question from an MIT perspective because, in fact, just last week a whole new division has been formed in the Engineering School called the Engineering Systems Division, which is specifically directed at dealing with the issues that you talked about. It is across disciplines. It involves not only interaction with the Engineering School but also with the social sciences as well as with the management sciences.

It is by no stretch of the imagination an easy thing to do in a university, but there are people with vision to basically say this has to be done and they are putting the faculty slots on the table to actually make it happen. So, I think that there are some universities quite committed to doing that.

David Smith: I would just like to comment that the Design Center at Carnegie Mellon University had exactly the same strategy and was very, very effective for those of us who participated in it. It was a very good experience.

Judith Hempel, University of California, San Francisco: I was just going to ask you, the panel, what you see for the year 2020 in the sort of division that we currently see in the chemistry modeling area between what some people call materials research and on the other side, biological materials research. Many of the techniques are very similar when you go all the way from pharmaceutical modeling over to materials modeling. In 2020 will there be a division, do you think, of this kind? Or will it come together?

David Dixon: In principle, one would hope that they would come together. I would not guarantee it at

this point, and the reason is that it is not clear to me yet that the potential function needs for both are going to be the same. I think you will have similarities, but right now if you look at industry, for example, I would say that in the use of computational chemistry throughout industry, including chemistry and pharmaceuticals, you probably have 80 percent of the people in the pharmaceutical/biological and about 20 percent in the materials/chemical side. And I think it is going to depend on a lot of features. One would hope that they would come together, but one does not know that yet.

Stanley Sandler, University of Delaware: The one thing I was surprised to see on your list was the comment for the year 2020 of the need for efficient methods for generating accurate potentials for molecular simulation. Intermolecular potentials generally imply two-body effects, not considering pairwise nonadditivity. Wouldn't you think by 2020 we would have done away with that completely, and be using quantum mechanics and simulation together to calculate the total energy functions, and not calculating individual two-body potentials?

David Dixon: One would hope that we would be able to do the quantum mechanics then, but, as Peter Taylor mentioned this morning, it is extremely expensive to get those very small energies correctly. It is not clear when we will be able to carry out the kind of simulations that Peter Cummings is doing in order to obtain thermophysical properties that will not require potential functions.

The potential function issue has come up at about three or four workshops we have been at over the last year as being one of the key foci for the next 10 years, if we are going to make headway in predicting thermophysical properties.

Stanley Sandler: How do you take into account nonpairwise additivity or non two-body effects?

David Dixon: That is part of generating the potential functions. Do we put in polarization potentials, and how do we put in three-body terms? There is a whole branch of science in how to do this correctly. People are now putting in polarization potentials so you can actually treat the electrostatics of a molecule as it interacts with other molecules and solve for the electric field changing over time with the dynamics. We need to worry about three-body effects, four-body, and up to n-body effects. These are areas that have to be studied in order for the field to progress.

Christos Georgakis, Lehigh University: Besides the industrial need for people who have been trained in several disciplines, maybe one other issue that we need to discuss is the type of academic degree(s) these people should have: B.Sc., M.S., or Ph.D.? More specifically, do we need more M.S. graduates than Ph.D.s?

David Smith: I will make an attempt at answering that. The first thing I would like to comment on is that chemical plants have a very long lifetime. They are on the ground for anywhere from 25 to 50 years. No matter what we do today in research, those plants are still going to be making adipic acid well into the next century, so we still need students trained in traditional chemical engineering who are going to go out and run those plants and manage our businesses in the traditional way. We just cannot walk away from that responsibility. There are too many billions of dollars of equipment on the ground to think that we should change the way we have been doing chemical engineering because in the next 10 years we are going to be building biological processes to replace them. This is just unrealistic.

Now, because of where we believe the growth to be, there is a trend in DuPont toward hiring more in the biological sciences, but that does not mean that we are going to stop hiring Ph.D.s, well trained in

chemical engineering and other disciplines, especially chemistry, to do our work. So I do not see DuPont R&D shifting away from hiring Ph.D.s.

I see us perhaps not hiring quite as many as we did 10 years ago, but there is another complicating factor here, and that is the demographics. If you walk into the DuPont cafeteria at lunchtime and take a look at the people there, there are a lot of chronologically gifted people present. We face a problem in terms of maintaining our technical capability in the next decade as these people retire. We have had a hard time getting through to our HR people that they cannot solve that problem in one year. There are not enough high-quality Ph.D.s in chemical engineering graduating in any one year for us to replace the people we are going to lose at that particular time. There is a fundamental problem but I am not sure that it is being addressed in the best possible way.

Tom Edgar: We talked a little bit about what the factory of the future is going to look like and the sort of technical demands that are going to be placed on people who are there. There is also this ongoing pressure to reduce the number of personnel in chemical plants as they become more automated.

What is the operator of the future going to look like? Is that person going to need a B.S. in chemical engineering rather than a community college degree? Someone in an editorial recently likened that person to essentially having the same responsibility as an airplane pilot in terms of the financial implications as well as possible safety implications. So, it is, again, something to think about. The differential costs between an operator and a B.S. chemical engineer are not all that great.

David Smith: I would like to say that the salary for a senior operator it is not that much different from the starting salary for a Ph.D. I had a young Ph.D. go to the plant and discover what the lead operator was making, and he came back absolutely incensed and wanted to know why I wasn't paying him more. I simply asked him, “Can you run the plant as well as he can?" The answer, of course, was no.

Gregory McRae, Massachusetts Institute of Technology: There is also another dimension of your question that I do not think we should ignore, especially if we are looking at 2020. The training that students have in chemical engineering makes them very attractive to a lot of people other than those who make chemicals. In a typical graduating class at MIT, not many of those kids finish up as chemical engineers because of the tremendous salary offers they are getting from many other disciplines. So, it is not only the question of how you can feed people into the existing industry to deal with the problem that Dave is describing, it is also how to make the industry itself attractive to these people, because they are able to go to many other different places.

For example, the last three of my Ph.D. students have gone to Wall Street, and for one of them, his first-year bonus is more than I will make in my lifetime.

Christos Georgakis: Let me retry a more focused question. Fathers and mothers see the incentive to pay for a B.S. education. University professors who have to do research see the incentive to educate Ph.D. students. Where is the financial incentive to educate M.S. students, which, I personally believe, will be in much larger demand? These M.S. graduates are needed to operate model-directed plants, or to utilize and apply the sophisticated computer technologies that the 2020 framework envisions.

David Smith: Christos, I am not sure whether it is a chicken and egg problem. Today if I wanted to, I could not go out and hire master's degree candidates in the numbers that we might be interested in hiring. So if they were available and we had experience dealing with that situation I might be able to give you a meaningful answer. In point of fact, I think that we are going to hire more B.S. engineers to

run our plants. I do not see us requiring more master's degrees. I think a lot of the mining that we want is going to be on-the-job training because of the increasing interdisciplinary nature of the work we will be doing in the future. It is too difficult to target that. I just think we are going to rely on more on-the-job training.

Gintaris Reklaitis: One might ask whether industry will continue to be wanting to invest in on-the-job training because at universities, we hear from our industrial colleagues that they want graduates ready to hit the ground running. They do not want to invest in training or in career development.

David Smith: Oh wow. I do not work at one of those places, I am happy to say.

Sam Kounaves, Tufts University: Parents pay for B.S. degrees and faculty pay for the Ph.D. degrees. At our university nobody wants to pay for M.S. degrees because no master's student is willing to encumber another $20,000 tuition bill on top of the bill for the B.S. degree, and we cannot support our master's students because they are not there long enough to do any research for us.

At least that is my perspective on the problem. The other question I have is more general. Greg talked about a balanced approach to doing modeling, both in terms of software and hardware, and I was just curious about your opinion on the allocation of resources for supporting computational chemistry. At first it sounded like we need all the high-powered hardware to do advanced modeling in all branches of chemistry and then later it sounded like the problem is actually in the software. I am curious about what your opinions are in terms of the resources. Are we at the point now where the hardware is really all right except for specific cases and that a larger effort needs to be put into funding programming and algorithms? I think people would rather put funding into hardware in the computer sciences, but software has been relegated down to the bottom. But that is exactly where the resources need to be put in order to address some of the problems, like interfacing and operating systems and interoperability.

David Dixon: Actually, I do not think that there has been a decision actually to put all the money only into hardware. I think one needs to continue to have hardware that is going to give us new ways of solving larger and larger problems, but you have to have a balance between the software and the hardware. I think ASCI and the Strategic Simulation Plan are trying to be balanced on both. I think the point that I was trying to make earlier is that the operating system part of the hardware has been the weakest part for us in terms of making it available to a broad range of users. We are significantly investing in all of the software pieces, the algorithm development and the theory, and we are trying to have a balanced approach. I think everybody has been trying to do that.

Paul Messina, California Institute of Technology and Department of Energy: But, yes, in terms of the investments, even the ASCI program, which is known for buying big machines, is spending less than half of its money on hardware. The rest of it is on developing the software and the applications. And your comment about computer scientists wanting the hardware, I am afraid, is wrong, because computer scientists typically do not want any hardware.

Tom Edgar: There is one other aspect of computational speeds that I would like to clarify. In real-time process control, there is an imperative to try to generate an answer within a sampling time or within a time constant of a process. That is a little bit different than for a simulation environment that is off-line. So there actually is a fair amount of pressure to have hardware that is really fast. But, of course, having said that, with the doubling of processor speeds every 18 months according to Moore's law, you can

always figure that with a good algorithm you can just wait 18 or 36 months and the computers will be fast enough so you can use your algorithm then.

Gintaris Reklaitis: From the perspective of combinatorial problems arising in supply-chain management and scheduling, doubling of computer time every 18 months will not suffice. These are problems for which computational effort grows exponentially with problem size, at least in the worst case. For the kinds of increasingly larger applications that people want to solve in these domains, waiting for the hardware to become faster is not the solution. The improvements must be found through algorithm research.

David Smith: I do not know. I used to have a group that did supply chain optimization, and in the reorganization that group went elsewhere. I would say that for a lot of very good sized, really significant supply chain problems that we were tackling, we were getting solutions in a couple of hours. The problem that we ran into was that, for reasons I could not understand, the people who wanted the answers were upset because they had to wait 2 hours for them; they did not realize that the business time scale that were dealing with probably involved days or weeks. These guys are used to running Excel spreadsheets in minutes, and the fact that they had posed a problem to the computer and had to wait for 2 hours for an answer was a very difficult cultural thing for them to deal with.

In general, business people are not trained in optimization, and they are usually very defensive when we bring those kinds of solutions to them. So it is really an educational problem that we have internally.

Judith Hempel: Are they senior staff of long standing?

David Smith: No. Neither were the people who were solving those problems for me. The young people in support positions in DuPont supplying that kind of information to business managers do not have the right kind of training and background to solve those problems. Solving the problem for them turns out to be an iterative process because they do not really understand how much information they have to give us so we can give them a good solution to their business problem.

7

Collaboratory Life: Challenges of Internet-mediated Science for Chemists

Thomas A. Finholt
University of Michigan

ABSTRACT

Since the birth of modern chemistry in the early 19th century there has been tremendous growth in the knowledge and the practical application of chemical principles. However, in many important ways, the practice of chemistry research and teaching has remained unchanged. The advent of the Internet as a worldwide mechanism for conducting scientific communication challenges this status quo. Specifically, innovations like collaboratories, or network-based virtual laboratories, remove constraints of distance and time on scientific collaboration. In particular, collaboratories increase access to scarce instruments, accelerate the flow of information, and place new demands on senior scientists to mentor students. Chemists need to appreciate how these new ways of doing scientific work will influence the conduct of chemistry research so that they can effectively anticipate and influence the development of emerging Interact technologies.

COLLABORATORY LIFE: CHALLENGES OF INTERNET-MEDIATED SCIENCE FOR CHEMISTS

The Internet,[1] the World Wide Web,[2] and sophisticated collaboration technologies[3] represent the raw ingredients for a revolution in the practice of chemistry. Yet today, for many chemists, this revolution is only partially realized or has not begun. As a result, the field of chemistry rests squarely

[1] See Hafner, K., and Lyon, M. (1996). *Where Wizards Stay Up Late: The Origins of the Internet*. New York: Simon & Schuster; Hauben, M., and Hauben, R. (1997). *Netizens: On the History and Impact of Usenet and the Internet*. Los Alamitos, CA: IEEE Computer Society Press; and Neil, R. (1997). *The Soul of the Internet: Net Gods, Netizens and the Wiring of the World*. Boston: Thomson Computer Press.

[2] Schatz, B.R., and Hardin, J.B. (1994). NCSA Mosaic and the World Wide Web: Global hypermedia protocols for the Interact. *Science, 265*, 895-901.

[3] Olson, G.M., and Olson, J.S. (1997). Research on computer supported cooperative work. In M.G. Helander, T.K. Landauer, and P.V. Prabhu (eds.), *Handbook of Human-Computer Interaction*, Second edition (pp. 1433-1456). New York: Elsevier.

on techniques and methods, many of which are over a century old. That is, while the content of chemical knowledge has advanced dramatically in the last 200 years, the organization of chemical research and education has remained relatively constant. By contrast, other disciplines race to embrace change, such as physicists' invention and rapid adoption of the World Wide Web and the widespread use of the Web for data dissemination among biomedical scientists. An apt metaphor to describe the challenge of the Internet for chemistry is Paul Gauguin's masterpiece (Figure 7.1), *Where Do We Come From? What Are We? Where Are We Going?*

In the title of his painting, Gauguin evokes the fear and uncertainty that accompany the transition from the past (*Where do we come from?*), through the present (*What are we?*), and into the unknowable future (*Where are going?*). The style and content of the painting also underline Gauguin's personal status as a bridging figure between impressionism and modernist schools, such as cubism and fauvism. While Gauguin was captivated by the impressionists early in his career, and worked and showed with them, later in his career he broke away and defined a new kind of art, often labeled post-impressionism. In this later work, Gauguin experimented with the use of color and symbolism in a way that paved the way for those who followed, including Matisse, Picasso, and Munch. Therefore, at many levels, this painting represents the tension of being caught between familiar traditions and the birth of new ways.

Chemists confront a similar tension between tried and true practices from the past and unknown alternative practices made possible through advances in information technology. In this sense *Where do we come from?* is a question about the traditions and conventions that have defined chemistry, especially with regard to the organization of research and education. The question *What are we?* offers an opportunity to reflect on the present state of the Internet, while the question *Where are we going?* forces consideration of the various new paths that Interact-mediated chemistry might follow into the future. The “Gauguin problem," then, is a statement about the difficulty any community faces when past and current success precludes full examination or experimentation with potentially transformational practices and approaches. In chemistry, the Gauguin problem can be framed as the enduring legacy from innovation at the dawn of modern chemistry in the late 18th and early 19th centuries, the mixing of inherited tradition with capabilities provided by the Internet that is occurring today, and alternative views of the future defined by new uses of the Internet.

Figure 7.1
Where Do We Come From? What Are We? Where Are We Going? Paul Gauguin, 1897. Tompkins Collection. Courtesy, Museum of Fine Arts, Boston.

WHERE DO WE COME FROM?

As an emergent discipline in the late 18th and early 19th centuries, chemistry had little received tradition in terms of either how to conduct chemical research or how to teach about chemical relationships. Therefore, the first 30 years of the 19th century saw development of many of the research and pedagogical practices that are still in use today. Three particularly important innovations were the creation and elaboration of the research laboratory, the use of laboratory classes in chemical education, and the use of lecture demonstrations to illuminate and clarify chemical principles.

The roots of the modern research laboratory, that is, a physical concentration of personnel and apparatus dedicated to Systematic chemical research, can be found in Humphry Davy's laboratory at the Royal Institution at the turn of the 19th century. Under the patronage of Count Rumford, the founder of the Royal Institution, Davy simultaneously mastered the arts of building voltaic devices for the discovery of new elements as well as raising the funds to construct new devices.[4] The impact of Davy' s efforts in terms of new knowledge is obvious in terms of his identification of sodium, potassium, and so forth. Nearly as important, however, is the model that Davy's lab established for subsequent scientists. That is, the halls of the Royal Institution defined not only a physical space that housed critical instruments but also a social organization that produced tradition, fostered networks, and became a place to train future generations of researchers. Indeed, among Davy's greatest Contributions was his mentorship of Michael Faraday. The broader success of the Royal Institution as a research organization is represented in the work of the Institution's nine Nobel laureates. Davy's approach carries through to the present largely unchanged. That is, the imperatives that drive modern research laboratories—hiring good people, installing state-of-the-art equipment, and getting funding—differ in magnitude and sophistication, but not in fundamental character, when compared to Davy's day. In fact, it seems likely that were Davy to travel forward in time to a lab at Michigan, or DuPont, or Cambridge, he might be amazed at the focus of research and the instruments in use, but the organization of scientists and resources to conduct the research would be entirely familiar.

A second great innovation from chemistry's founding era was the invention of the laboratory class. The introduction of laboratory classes is associated with Justus yon Liebig and the organic chemistry curriculum he developed at the University of Giessen, beginning in 1824. Prior to Liebig, chemistry was taught largely through example and demonstration and not through active participation at the lab bench. While the demonstration approach was developed to a high art, it did not produce many chemists, since few students gained hands-on experience in manipulating chemicals.[5] By contrast, Liebig's "practical" approach immersed students in exercises that were reasonable analogs to techniques used by practicing chemists. Therefore, students in Liebig's lab gained not only experience but also a sense of the thrill and challenge of conducting research. Pictures of Liebig's laboratory drawn in 1842 show a scene not too different from a modern undergraduate chemistry lab. Students clustered around lab benches produce and analyze specified compounds, guided by an instructor or lab supervisor. Again, as with Davy, we could bring Liebig forward in time and most of what he would observe about the organization of modern lab courses would be similar to his own era.

A final innovation was the public lecture and accompanying demonstrations. This practice origi

[4] Fullmer, J. (1989). Humphry Davy: Fund raiser. In F.A.J.L. James (ed.), *The Development of the Laboratory: Essays on the Place of Experiment in Industrial Civilization* (pp. 11-21). London: MacMillan Press.

[5] Fenby, D.V. (1989). The lectureship in chemistry and the chemical laboratory, University of Glasgow, 1747-1818. In F.A.J.L. James (ed.), *The Development of the Laboratory: Essays on the Place of Experiment in Industrial Civilization* (pp. 22-33). London: MacMillan Press.

nated with Faraday's Christmas lectures at the Royal Institution, which began in 1826. The Christmas lectures have continued to the present and are now broadcast via television and available via download from the Web. The essence of this tradition is for prominent scientists to communicate the excitement and value of their research to a lay audience, particularly young people. The main mechanism used in the Christmas lecture is explanation accompanied by interesting demonstrations. This pedagogical approach is not just the foundation of the Christmas lecture series; it is also the continuing basis for most secondary and undergraduate instruction in chemistry. Indeed, most large university chemistry lecture halls have an accompanying room where special equipment, just for doing demonstrations, is prepared under the guidance of a demonstrations supervisor. Therefore, as in the two preceding examples, there is not much about contemporary lectures that would be surprising or strange to someone from Faraday's time.

These three innovations do not represent a comprehensive treatment of the historical tradition in chemistry. However, they do signify important cornerstones of past and present practice in chemical research and education. More important, the legacy of the research laboratory, the laboratory course, and the public lecture defines the starting place for thinking about the organization of alternative approaches. Specifically, with the advent of the Internet, the Web, and collaboration tools, chemists confront a choice between extrapolation from a known and successful past (i.e., joining the capabilities of these new technologies to familiar practices) and exploration of entirely new ways of doing and teaching chemistry.

WHAT ARE WE?

The previous section examines chemistry's past and how that past determines the present. This section considers the present, particularly the new opportunities available to chemists through the expansion of computer and Internet technologies. These opportunities can be thought of in terms of the raw performance of computer processors, the capacity of communication networks, the scope of networks, and the evolution of software.

The engine of progress in computing is Moore's law, or the observation by former Intel CEO Gordon Moore that the performance of computer processor chips—when measured as the number of transistors per processor—doubles roughly every 2 years. Figure 7.2 illustrates the progress of processor development over the last 26 years, using Intel chips as a benchmark. A corollary of Moore's law says that for a constant price, a computer purchaser gets twice as much power every 2 years. Either way, this is a trend toward phenomenal increases in computing capability over time. For instance, it is often observed that current Pentium 2 workstations are roughly comparable in speed to supercomputers sold in the early 1980s.

Recent explosive growth in the size of the Internet points to a new metric of computing performance: network bandwidth. Contemporary performance, or lack of performance, within the Internet is legendary—hence the popular observation that WWW stands for "World Wide Wait." Examining plans for installation of high-capacity fiber across both the Pacific and the Atlantic oceans, however, suggests that current network delays may soon disappear. For example, capacity across the Pacific is slated to increase to 300 gigabits per second (Gbps) by the year 2000 from a 1998 level of 25 Gbps, while capacity across the Atlantic will increase to 250 Gbps from 110 Gbps.[6] For comparison purposes, I

[6] Staple, G. (1998). The global bandwidth boom: Something's happening here . . . but what? Bandwidth Economy Conference, Columbia Business School. May 1, 1998.

Gbps is equivalent to 70,000 phone calls. Estimates are that the expansion of international bandwidth will easily meet projected growth in voice traffic and that the bulk of increased use will be data traffic. This means that many applications that impose prohibitive bandwidth overhead today (such as desktop video conferencing) may, in the near future; become more practical. Plans are already launched in the United States for next-generation network technologies that will exploit increasing bandwidth, such as the University Consortium for Advanced Internet Development (UCAID; <http://www.ucaid.org>). UCAID hopes to deliver network performance of 150 megabits per second, which in most cases will represent dramatic improvement in network throughput. This could mean, for instance, easy transfer of large data sets around the Internet, routine use of bandwidth-intensive applications (such as audio or video), and increased use of applications that require high quality of service (such as remote manipulation of instruments).

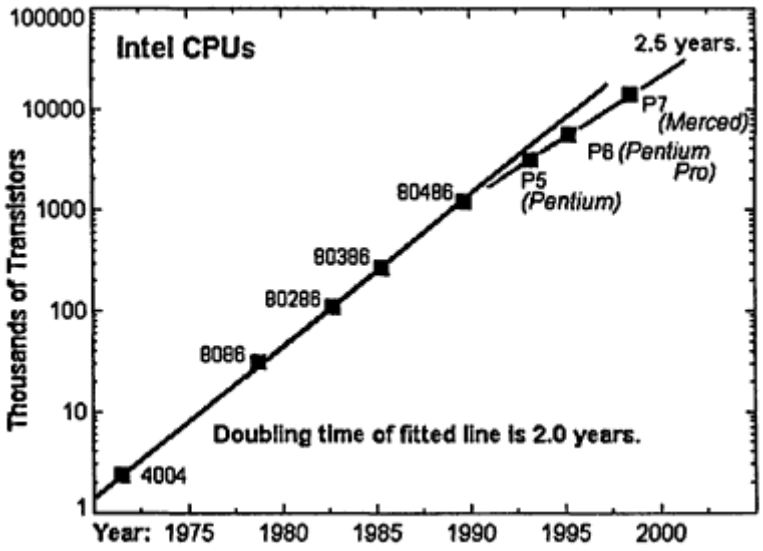


Figure 7.2
Illustration of Moore's law using Intel CPU chips. Courtesy of George Watson, University of Delaware, <http://www.physics.udel.edu/~watson/scen103/intel.html>.

A third hallmark of change in the Internet is the tremendous expansion of hosts and the worldwide penetration of the Internet. Recent examination of connected countries shows host domains for all but three nations, with even the tiny island nations of Nauru and Comoros linked to the global network.[7] This scope means that, to an unprecedented extent, scientists with access to the Internet can reasonably expect to communicate—and possibly collaborate—with colleagues located anywhere on the face of the Earth. Similarly, through the Internet, scarce resources, such as libraries and rare instruments, can be made available to larger populations of users.

The changes noted above are momentous, but each represents more potential progress than realized progress, at least in terms of practice and behavior. This may be because unprecedented increases in

[7] Kaiser, Jocelyn, ed. (1999). Netwatch. *Science*, *283*, 295.

computing power, network bandwidth, and network scope have not been matched by corresponding improvement in the usability of applications and software. Figure 7.3 represents this trend. The y-axis indicates raw performance of computing technology, such as the benchmarks listed above (processor power, bandwidth capacity). The x-axis indicates time. The upward slope of each curve shows overall improvement. However, the contrast between the curve labeled "raw performance" and the curve labeled "real performance" reflects the difference between what we could do with information technology (shown at the extreme left with the “hype" curve) and what we can actually do. This difference, called the "reality gap," is in part why scientists may be reluctant to launch or adopt bold information technology innovations. For instance, an oft-cited reason for staying with a specific platform or application is the cost of learning a new program. Some have argued that this essential difficulty is the root of the apparent productivity paradox in computing, where massive investment in computing technology has often failed to produce significant increases in output or performance.[8] A way out of this bind might be broader application of user-centered design philosophies that, in contrast to traditional development approaches, attempt to evolve applications with constant feedback drawn from users in authentic settings.

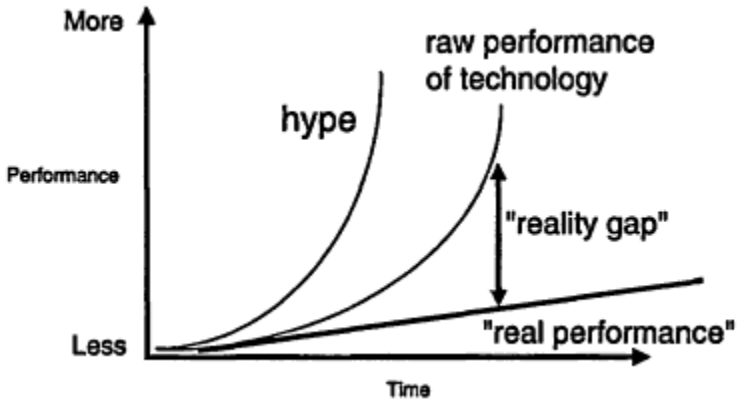


Figure 7.3
The "reality gap" in information technology performance. Courtesy of Dan Atkins, University of Michigan.

In summary, the present is a time of fantastic change in the raw capabilities of information and network technologies. However, the impact of these changes is somewhat reduced by the difficulty of effectively harnessing the potential of new technologies. For instance, producing usable software is still more difficult than it should be, and many software applications don't produce benefits to justify the often painful process of learning to use them. A particular challenge for chemists will be finding ways to train the next generation of chemical software designers to more effectively design code, such that broad populations of users can adopt applications quickly and easily—therefore more fully tapping the rich potential of advances in hardware and network systems.

[8] Landauer, T.K. (1995). *The Trouble with Computers*. Cambridge, MA: MIT Press.

WHERE ARE WE GOING?

The Collaboratory Concept

One way to think about the future is to seek examples in the present of possible new modes of doing research and teaching. A dream of Internet proponents has been the creation of collaboratories, or centers without walls ". . . in which the nation's researchers can perform their research without regard to geographical location—interacting with colleagues, accessing instrumentation, sharing data and computational resources, [and] accessing information in digital libraries."[9] The earliest and most extensive collaboratory R&D project is the Space Physics and Aeronomy Research Collaboratory (SPARC). This project began in late 1992, and by early 1993 produced the world's first operational collaboratory.[10]

The SPARC project started by addressing the needs of space physicists who used observational data collected from a suite of ground-based instruments at the Sondrestrom Upper Atmospheric Research Facility, located in Greenland. Between 1992 and 1995, SPARC evolved to provide data viewers for five Sondrestrom instruments: (1) the 60-meter incoherent scatter radar, (2) an all-sky camera, (3) a Fabry-Perot interferometer, (4) an imaging riometer, and (5) a local magnetometer. During this period, the primary use of the collaboratory involved real-time access to these instruments either for ionospheric observations or for instrument testing. At this early stage, collaboratory-based science resembled traditional research practices, although mediated by the Internet.

Between 1995 and 1997 SPARC transformed dramatically to accommodate three major changes. First, early success with the collaboratory led to increased interest from scientists and to demands to include more instruments. Second, the rapid emergence and adoption of the Web suggested the importance of a Web-based interface to SPARC. To accommodate this need, the core technology of SPARC was rebuilt in Java. Third, having seen what might be possible with the early SPARC system, users proposed three new types of uses: (1) expansion of the data sources to produce a global "field of view" in real time; (2) inclusion in real time of theoretical model output side by side with observational data; and (3) use of the SPARC technology to support distributed, online workshops or conferences.

Figure 7.4 shows a snapshot of the SPARC interface during a recent campaign. The SPARC interface has three main components. First, the SPARC "session manager," shown in the upper left of Figure 7.4, organizes scientific activity by topic into groups called "rooms."[11] Within these rooms, scientists find useful URLs, chat streams specific to that room, and saved configurations for data viewers relevant to that room. Note that each room name is followed by a number in parentheses, which represents the number of scientists currently using that room, and that the names of participants within a selected room are displayed below the session manager. This information provides a crucial form of presence awareness in the virtual setting that would be obtained automatically in a shared physical setting. The chat window, shown in the lower left of Figure 7.4, is a text-based channel for communication among SPARC users. This chat application is persistent, meaning that scientists can join a conversation in progress and scroll back to review earlier comments. Finally, the bulk of the interface

[9] National Research Council, Computer Science and Telecommunications Board (1993). *National Collaboratories: Applying Information Technology for Scientific Research*. Washington, DC: National Academy Press.

[10] Finholt, T.A., and Olson, G.M. (1997). From laboratories to collaboratories: A new organizational form for scientific collaboration. *Psychological Science, 8*, 28-36.

[11] Lee, J.H., Prakash, A., Jaeger, T., and Wu, G. (1996). Supporting multi-user, multi-applet workspaces in CBE. *Proceedings of the ACM 1996 Conference on Computer-supported Cooperative Work* (pp. 344-353). New York: ACM Press.

is devoted to data displays. In this case, Figure 7.4 shows time series plots of electron densities against altitude as observed by five incoherent scatter radars spanning the Northern Hemisphere from the Norwegian Arctic to Puerto Rico. An important feature of the data viewers is the presentation of observations from multiple instruments on a common time axis.

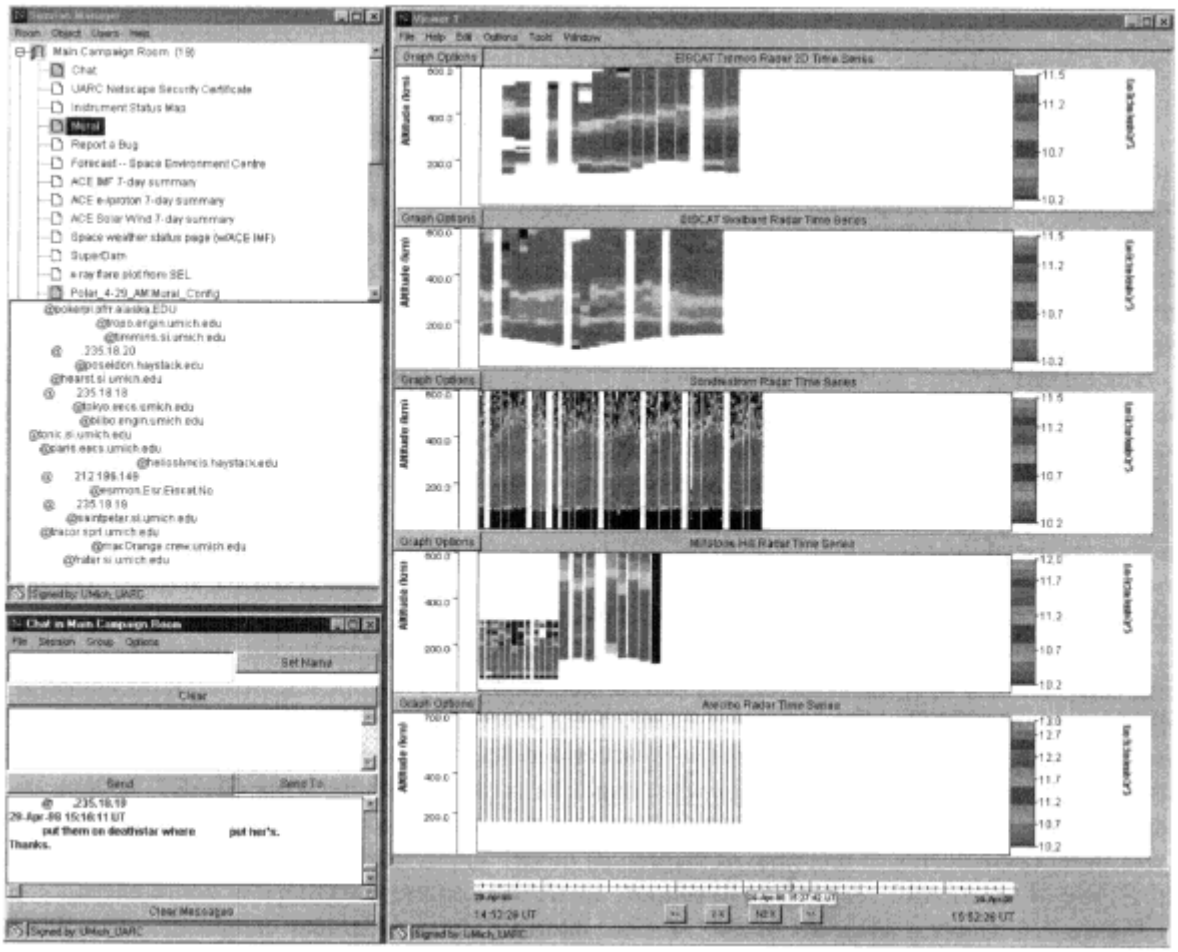


Figure 7.4
The SPARC interface showing data from five incoherent scatter radars during a campaign of April 1998. From top to bottom, the data sources are EISCAT Tromso, EISCAT Svalbard, Sondrestrom, Millstone Hill Observatory, and Arecibo. The session manager is in the upper left comer and the chat window is below it.

While SPARC has had many kinds of impact on the space physics community, two consequences are particularly noteworthy. First, by relaxing constraints of time and place, SPARC makes it possible to carry out collaborative campaigns with more flexibility in scheduling and participation. For example, SPARC makes it much easier to access complementary expertise and to mentor students. In the past, scientists were restricted to expertise available at the remote observatory site. Similarly, students gained the best opportunities to learn about data collection only by traveling to a remote observatory to participate in a campaign. Today, SPARC allows scientists with complementary expertise to work

together, without imposing demanding travel burdens. For instance, Figure 7.5 shows the pattern of communication from a 1995 campaign. In this campaign, a Florida-based space physicist with no incoherent scatter radar experience was guided through a data collection run by a California-based colleague with an extensive background in radar operations (with the added benefit that the Florida physicist could have his students watch as well).

A second notable impact of SPARC is that the collaboratory has accelerated a paradigm shift in the orientation of space physicists to their data. Specifically, upper-atmospheric phenomena reflect a global system in which the atmosphere, the solar wind, and the magnetosphere produce effects over very broad regions. In the past, understanding this system took months of careful integration of multiple data sources. Today, SPARC makes it possible to examine real-time data coordinated on a common time scale from any source on the Internet. The common framework for viewing data means that events at one location are easily correlated with events at other locations. For example, in recent campaigns SPARC has provided simultaneous data from as many as six incoherent scatter radars, from spacecraft, and from unattended instrument arrays across Europe, Asia, and North America. In addition, SPARC provides a mechanism for the simultaneous display of data and model predictions. Traditionally, the substantial computational demands of such models have meant that most of this work was done long after the observational data were collected. Today, improvements in the models and less expensive supercomputing have made it possible to do data/theory evaluation in real time.

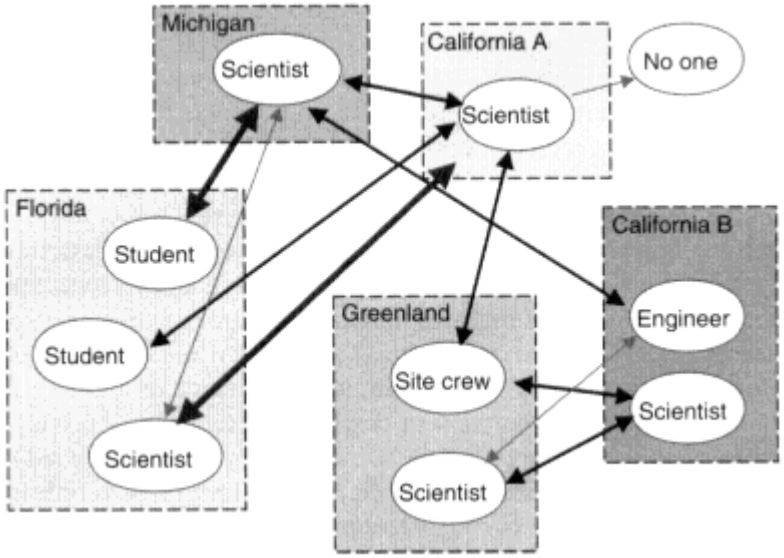


Figure 7.5
The pattern of computer-mediated communication among participants in two interleaved campaigns over a 3-day period in 1995. The participants at Florida, California A, and the site crew member in Greenland constituted one campaign. The participants at California B, the scientist in Greenland, and the site crew member in Greenland constituted the other campaign. The Michigan participant was a scientist/programmer who contributed to scientific conversations while monitoring the SPARC systems. The width of the lines connecting persons indicates the frequency of communication. Lines to the node "No one" indicate questions asked that received no response.

Chemistry on the Internet

The emergence of systems like collaboratories represents an opportunity for chemists but also poses a significant number of challenges. Assuming that many of the technical barriers to Internet use disappear with improved computational and network performance, these challenges can be summarized in terms of potential changes to existing practices. Three changes, in particular, demand attention. First, the introduction of collaboratories—specifically, collaboratories designed to provide remote access to scarce instrumentation—may transform traditional ideas about instrument ownership and control. Second, collaboratories represent new channels for communication; however, additional information flow may be undesirable in many areas of chemistry, particularly if the flow is unregulated (e.g., a threat to proprietary content) or unqualified (e.g., claims that haven't been reviewed or validated). Finally, collaboratories create new arenas for learning by expanding opportunities—specially for students—to join in with experienced scientists in the conduct of research projects. However, this new style of participatory education may require teaching and mentoring skills beyond the demands of familiar lecture and lab-style learning.

The Value of Collaboratories

A key component of the collaboratory concept, at least as realized in the SPARC project described above, is the use of media-rich information technologies to link scientists with each other and with instrument facilities, independent of distance and time. This idea is particularly attractive in fields like space physics, which rely on a limited number of observatories and spacecraft, and where the primary data collection mode is passive. By contrast, in a field like chemistry, experiments often involve direct manipulation of compounds by investigators. That is, while analytic instruments may be viewable and controllable at a distance via network interfaces, many kinds of sample preparation require close proximity between the lab bench and instruments. In these cases, collaboratory technology may not be that useful for chemists. However, there may be productive ways to use the Internet to link chemists to papers, results, or data. For instance, collaboratories may become the mechanism for ongoing electronic workshops where chemists can present and discuss findings while drawing on the tools and literature used to conduct the initial research, such as visualizations or analyses.

An important mechanism in an electronic workshop might be tools for presence awareness. That is, in a physical setting we can know who is present, paying attention, and so forth. In a virtual setting, particularly with participants drawn from multiple time zones, there is a need to more explicitly represent who is doing what and when. Visual Who (<http://judith.www.media.mit.edu/Judith/VisualWho/VisualWho.html>), developed at the MIT Media Lab, is one instance of a device for helping people navigate a virtual space.[12] In Visual Who the user display indicates who is active (those with names shown in the display) as well as the recency of activity (indicated by color, where red is more recent, and blue is in the past). Another feature of Visual Who is that the display groups people by sub-dimensions, which could correspond to specialty, status, organizational affiliation, and so forth. Such a tool would help scientists identify experts in unfamiliar specialties, as well as help find old colleagues at distant sites.

[12] Donath, J. (1995). Visual Who: Animating the affinities and activities of an electronic community. *ACM Multimedia 95—Electronic Proceedings* . November 5-9, 1995. San Francisco, California.

Free Flow of Information

The major impact of the Internet on chemistry so far is probably the use of electronic mail for scientific communication. As described by one chemist at Michigan, ". . . exciting things are happening. I have now published papers with people that I've never met, and in one case, never even talked with aside from electronic communications. Sort of a virtual person as far as I can tell, although real samples did arrive by real courier." It is not difficult to imagine this kind of process accelerating with the adoption of collaboratories, where remote collaborators might jointly analyze a sample, then share their results via an electronic workshop, and ultimately use collaboration tools for editing publications.

A variant of this process exists today in chemistry in the form of online conferences. For example, in chemical education, the series of Confchem "meetings" have all been conducted online (see <http://www.chem.vt.edu/confchem/1998). In these sessions, papers are published on Web sites, and over a designated period other scientists read the papers and comment on them via chat rooms and e-mail distribution lists. The authors respond to these comments and over the period of the conference, authors and readers engage in computer-mediated dialog. If ventures like electronic workshops and conferences are to succeed, chemists need to solve the problem of chemical mark-up languages. Today, equations can be represented on Web pages as graphical elements, such as bitmaps or GIF images, but these equations have no formal mark-up syntax, which means documents can't be searched for compounds or equations. For example, efforts to base chemical mark-up on standards adopted by the World Wide Web Consortium, such as XML, suggest that in the future chemists will have convenient tools for writing and reading equations, and for searching (see < http://www.xml-cml.org>).

Participatory Education

Visible efforts to introduce new technology into the chemistry curriculum include the use of Web pages to present course content and the creation of CD-ROM supplements to textbooks. These innovations have their primary impact on individual learners, and even in this case there is some skepticism. For instance, using a CD-ROM instead of paper is largely a substitution of one medium for another and not a fundamental shift in pedagogical orientation. More exciting is the possibility that the Web, through collaboratories, may open new avenues for participation by a wide variety of students in chemistry research. An illustration of this approach is the Collaboratory for Undergraduate Research and Education experiment conducted at the Environmental Molecular Science Laboratory (EMSL) of the Pacific Northwest National Laboratory.[13] In this setup, a class of honors chemistry students at the University of Washington used the EMSL collaboratory facility to use advanced analytic instruments at PNNL and to interact with expert users of these instruments at PNNL (see <http://www.emsl.pnl.gov:2080/docs/collab/projects/CURE/index.html>). Within SPARC, mentioned above, undergraduates used the collaboratory facility to participate "alongside" senior investigators during a combined optical/incoherent scatter radar campaign. As the diagram in Figure 7.5 shows, through the collaboratory students in Florida viewed live data and discussed it with scientists in Northern California, Michigan, and Greenland. For these students, the opportunity to view phenomena as they occurred brought to life material that previously was only the stuff of lectures and textbook explanations.

[13] Myers, J., Chonacky, N., Dunning, T., and Leber, E. (1997). Collaboratories: Bringing national laboratories into the undergraduate classroom and laboratory via the Internet. *Council of Undergraduate Research Quarterly* , *17*, 116-120.

CONCLUSION

The Internet offers exciting new opportunities for chemists. The collaboratory concept is just one illustration of how Internet-mediated science may affect the relationship of researchers to instruments and data, of colleagues to each other, and of teachers and advisors to students. While Web-based tools may alter much of the current familiar landscape of practice and pedagogy, it is important to recognize what the Web cannot do. Specifically, simply “surfing" for information is not a replacement for learning. Amidst the temptation to browse endlessly among an ever widening array of online resources, students and researchers must still take time to absorb and reflect on ideas in order to master and understand key concepts.

ACKNOWLEDGMENTS

Thanks to James Finholt, Albert Finholt, Peter Murray-Rust, and James Penner-Hahn for feedback from the chemistry perspective. Thanks to Dan Atkins for his ideas on the reality gap in information technology. And finally, thanks to Stephanie Teasley for helpful comments and suggestions on earlier drafts. Requests for reprints should be addressed to (a) Thomas A. Finholt, Collaboratory for Research on Electronic Work, C-2420, 701 Tappan St., Ann Arbor, MI 48109-1234; or (b) <finholt@umich.edu>.

8

A Computer Science Perspective on Computing for the Chemical Sciences

Susan L. Graham
University of California, Berkeley

As a computer scientist whose formal education in chemistry stopped at high school, my goal is not to describe aspects of computational chemistry. Instead, I will try to suggest to you, from a computing perspective, why high-performance computing is difficult. Then from a technical point of view I will summarize some of the issues that we have to contend with if we are really going to take advantage of all the exciting computational opportunities outlined by other participants in this meeting.

Let us first consider performance. If we want to get more out of computing, the way we get that is by using parallelism (Box 8.1). The reasons we use parallelism are (1) to reduce the overall elapsed time in doing a demanding computation; (2) to keep the calculation moving when delays arise in sequential computation; and (3) to overcome fundamental limitations, like the speed of light, that bound the speed of sequential computation. Parallelism has been used in computing for a very long time, and it exists at many, many levels of the computing hierarchy. So, there is a great deal of parallelism in the box that sits on your desk, in a single processor.

Higher-level parallelism than that found in a single processor can be achieved by using multiple

BOX 8.1 PARALLELISM AS SOURCE OF COMPUTING SPEED

- Is a remedy for fundamental limits
- Exists at many different levels
- —Single processors
- —Shared memory multiprocessors
- —Distributed memory multiprocessors
- —Networks of machines
- Doesn't come free

processors in a variety of organizations, some of which share memory, some of which communicate over a network, some of which are tightly coupled, some of which are communicating over very long distances. That parallelism really does enhance our capability, but it doesn't come free.

Let us consider where some of the parallelism comes from (Box 8.2), to try to understand what it is that the programmers and the computer scientists do and also why the world, which is already complicated, is getting even more complicated from a computing point of view.

- Since the early days of computing there has been parallelism that goes on at the bit level. In other words the computer can access multiple bits of information at once and can operate on them in parallel. Fundamental hardware operations such as addition take advantage of that kind of parallelism.
- There is also parallelism at the single-instruction level. In virtually any modern processor it is possible to execute multiple instructions at the same time. In one instant, multiple instructions are both issued and executed.
- There is overlap between computational operations such as addition and data movement such as reading and writing values to memory. Thus, it is possible to write the data that must be saved and to read the data that will be used next at the same time that computation is going on.
- Finally, almost any software system one uses has parallelism in that multiple jobs can be executing at the same time. In particular, when one job stalls because it is waiting for data, some other job takes over and does its thing. That *time-sharing* technology is actually quite old at this point.

Part of the difficulty in exploiting the parallelism available even on a single processor is that the improvements in different aspects of a computer are not occurring uniformly (Figure 8.1). There are many versions of Moore's law, which predicts the rate of improvement in computing technology. It almost doesn't matter which one we use. The basic message in Moore's law is that the speed of processors doubles every 18 months, roughly speaking.

The speed of accessing memory improves as well, but it is improving at a much slower rate. So, over time, the gap between how quickly a processor can execute instructions and how quickly the same processor can read and write data is widening. That means that any time you have a computation that depends on accessing data, you can lose all the performance that you might have gained from the faster MIPS (million-instructions-per-second) execution rate because the processor is waiting to write what it has just computed and read what it needs next.

BOX 8.2 PARALLELISM IN MODERN PROCESSORS

- Bit-level parallelism
- —Within floating point operations, etc.
- Instruction-level parallelism (ILP)
- —Multiple instructions execute per clock cycle
- Memory system parallelism
- —Overlap of memory operations with computation
- Operating system parallelism
- —Multiple jobs run in parallel on commodity SPMs

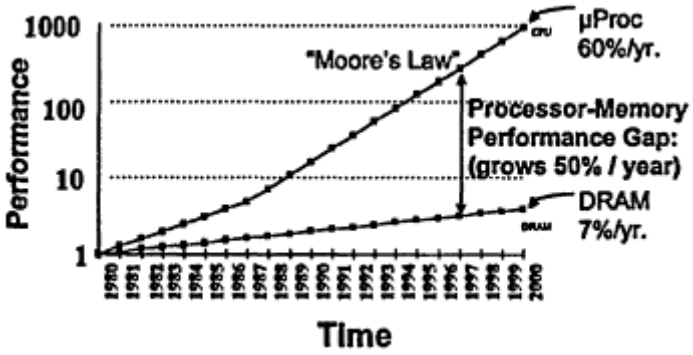


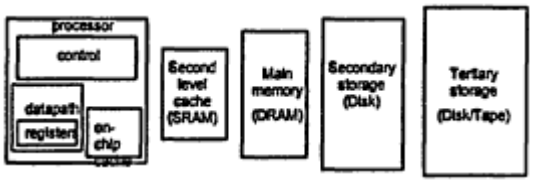
Figure 8.1
Processor-DRAM gap (latency).

Manufacturers aren't stupid. They are aware of that problem, so there are a number of technical devices that computer designers and manufacturers have used to try to mitigate that problem. The result is that storage has become more and more complicated.

First consider the processor. Figure 8.2 shows the memory components and underneath indicates the approximate speeds of access and the sizes. There is very fast so-called main memory available to the processor. The processor also has a limited number of high-speed registers that provide even faster access to data values. Next are the disks attached to the processor directly. Disks and tape available through a network are at the far right of the diagram—they have substantially higher storage capacity, but much slower access.

On or between the processor and main memory are other storage devices called *caches*. Caches were introduced to bridge the gap between the speed of the processor and the speed of the memory. A cache can be thought of as a faster memory that holds a copy of the recently accessed data, or in the case of a program cache, a copy of a recently executed (or about-to-be executed) portion of the program. (In a

Intended to mitigate processor/memory gaps
As storage capacity increases, speed of access decreases
Caches hold values that are spatially and temporally close



Speed (ns):	1s	10s	100s	10s ms	10s sec
Size (bytes):	100s	Ks.	Ms	Gs	Ts

Figure 8.2
Memory hierarchy.

computer, the program is stored in memory, and part of the slow-down in execution can come from accessing the next portion of the program.)

The point about the caches is that they hold chunks. So, when a cache is loaded, it is filled with a sequence of values that happen to be stored close together. What is held in this cache is that which has been used most recently, together with the values stored physically just before and just after. So, if the program jumps around wildly, accessing data that is over here and accessing data that is over there, the data that is over there overwrites what was in the cache. Now the processor has lost its very fast access to what was in the cache before. Consequently, the strategy from a programmer's point of view is to try to cluster all the information that is being used at approximately the same time so that while it is in use, it all lives in this very fast memory called the cache.

If a program is fetching chunks at a time, the only way to keep the needed information close by is to have actually stored it close together, because the hardware is going to grab it close together. That involves a certain amount of strategy in designing a program so that its data is close together and in the right place at the right time.

Now, suppose we consider high-end computing and some of the ideas that lead up to the kinds of systems Paul Messina has described, in which there are multiple processors (Box 8.3). Now this gap between the processor speed and the memory speed gets much more complicated. There are multiple processors executing at the same time, and logically at least they are sharing data. There are data written by a part of the computation executing on one processor and read by other parts of the computation executing on other processors. Maybe the processors are actually physically sharing memory as well; maybe they are not. So the latency—the delay in waiting to get the data—now increases. Furthermore, there is contention for the data because multiple processors may want the same information. Consequently, one needs mechanisms for synchronization to make sure that one processor does not try to read something before another processor has written it, to make sure that two or more processors don't try to write the same data at the same time, and so on. Now bandwidth, which is the volume of information that one can communicate at one time, also becomes more difficult to deal with because, in fact, it becomes less predictable.

The time it takes to get a certain amount of data from here to there is not necessarily precise, and it depends on what is going on with all these other processors. So, the different components that all

BOX 8.3 MULTIPLE PROCESSORS COMPLICATE THE PROBLEM

- Data is shared among processors
- Latency (data access time) increases
- Bandwidth (volume of data communication) varies
- Performance components become more non-uniform
- Processors must be scheduled (assigned a part of the overall computational task)
- Strategies are different depending on architecture
- *End-to-end* performance requires

- —Keeping together all the processors busy doing useful work
- —Having the right data in the right place as soon as it's needed

together constitute the performance of one of these computing systems become very non-uniform, very interdependent on effects that are going on at all these different levels of memory hierarchy, and therefore very hard to understand and to predict.

In some sense the processors are now executing independently, but they are all cooperating to do the same computation. And so, there is an issue of how to use the different processors: What part of the calculation is done by each one of them, and how is that done so that the overall performance is as good as possible?

The real issue is *end-to-end* performance—how long it takes to execute the program and get the results. In a multiprocessor setting, end-to-end performance requires that all of the processors be kept busy all of the time doing useful work. Otherwise it doesn't matter what the theoretical speed is—you are not getting the benefit of it if some of the processors are sitting idle some of the time. In order to keep the processors busy all of the time doing useful work, the data they need must be in the right place in the right time. That means moving information around from one place to another in the memory hierarchy: from one memory to another, from memory to cache to registers and so on.

Some of that is under the control of the programmer, and some of it is not, but all of it affects the end-to-end performance of a calculation. For that reason, sometimes one sees that the peak performance of a system is something wonderful, but the actual performance that a given person is getting on his or her calculation is much worse, and it is very frustrating.

In order to cope with this memory/processor speed disparity and the scaling-up that Paul Messina talked about, different system designers have taken different architectural approaches. One of the other complications is that the strategies that cause computations to be done very efficiently on certain architectures are different from the strategies one would use on other architectures, even though the different architectural approaches all have their merits with benefits in terms of performance.

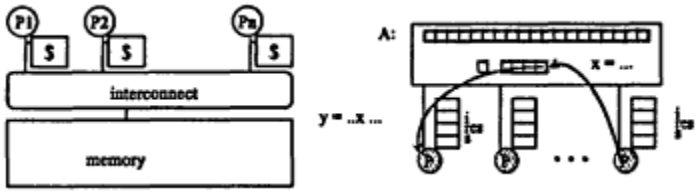
What software designers do to cope with this situation, in part, is to develop programming models—ways of thinking about what is going on with all the varieties of parallelism that attempt to match the architectural organizations. There are a number of these models. I am just going to show you two of them very briefly, just to look at their differences and to give you some sense of why what works well on one platform doesn't necessarily work well on another platform. Where I am going with this is to point out how very vulnerable legacy code is. In other words, code that has had an enormous development investment, that has really done good things for scientists, is aging very quickly. As we move through these architectural changes, it no longer performs very well, even though it might still manage to get some work done.

Without going into all the details, the first architectural paradigm in the multiprocessor world is that there is a collection of processors and they all share the same physical memory (Figure 8.3). There is some hardware in between the processors and the memory called the *interconnect* that allows them all to get to the same memory. The natural programming model that goes along with that architecture is to have the programs on each processor take the view that they are all looking at the same data. For example, if one has an array A, then all of the processors can access that same array. Therefore not only is it the case that one processor can write some information and the other One has it right there to read, but also the way that the different computations running on the different processors can communicate is by one writing data that the other one reads. The conceptual idea is that all the information is shared, and it is just the computational part that is being done in parallel.

In contrast, a lot of people believe that the only way you really scale up is to have a setup in which there is a collection of separate processors, and each one has its own memory (Figure 8.4). When the processors communicate, they communicate over a very high speed network. They each have their own local memories. Therefore the issue of when data is really shared and when a processor is reading or

writing a copy of data—and therefore not necessarily the same copy that some other processor is looking at—now becomes more complicated. Deep down, communication is going on by sending messages back and forth over this high-speed network.

Processors all connected to a large shared memory



Multiple processes (threads) use Shared Address Space

each has a set of private variables
collectively have a set of shared variables
communicates implicitly through shared variables
coordinate explicitly by synchronization operations on shared variables

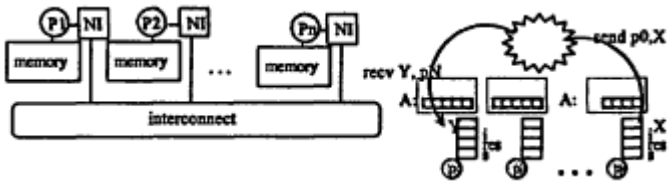
Figure 8.3
Shared memory machine/programming model.

Now if there is an array A, it is actually physically scattered among the memory of the different processors. One might then choose to organize a computation so that each processor works on part of the data that is in its own memory because it can get to it much faster. Ultimately if they are doing a collaborative calculation, there is a point at which the different parts of the computation need to look at data that reside in memory on other processors, so latency kicks in again.

Latency is a problem because a program has to go out on the network to read the data on other processors, get it back, and write it in local memory, and only then can the program use those data. As part of the programming model, one of the big issues is how much the user needs to know about all the message passing. To what extent does that complexity need to be exposed to a computational scientist

Each processor connected to own memory (and cache)
Each "node" has a network interface (NI)

all communication and synchronization done through NI



program consists of a collection of named processes
processes communicate by explicit data transfers (message passing)
logically shared data is partitioned over local processes

Figure 8.4
Distributed memory machine/programming model.

whose real concern is not with moving data from here to there but whose real concern is solving a set of equations and getting some scientific answers?

So, what does a scientist do in order to create a parallel program in this environment (Box 8.4)? One of the hardest problems, and presumably one of the issues that came up repeatedly in other talks, is that in order to take advantage of parallelism one has to identify the work that can be done in parallel. Work can be done in parallel if one part of it isn't dependent on the results of another part.

The goal is to break up your calculations into components that can all be done at the same time, and then they can collaborate on communicating their results. That requires a high-level strategy about the problem, and it involves different problem-solving techniques, and it involves people who understand the domain in which they are solving the problems. It may not be something that can be decided in advance. It may be that where the parallelism is depends on the data, so that as the calculation proceeds it has to reorganize itself in order to maintain parallelism.

If that decomposition of the problem is done wrong, then you have the situation in which some processors sit idly waiting for other processors to complete. That situation sometimes is described as a *load imbalance*. If the computational load isn't equally distributed across the system, then the overall computational speed goes down. The computation is not getting the advantage of some of the processors.

So the user has to figure out where the parallelism is. Then he or she has to figure out how to partition the work across the processors, and how to partition the data so that the data sit close to the clients who want to use it (because of all the communication delays). Finally, at the lower level somewhere in the system, some piece of software or some programmer has to worry about actually describing the communication, describing the synchronization, and scheduling which work gets carried out on which processor.

The limitation in the speed of a parallel computation is determined by the parts that are not parallel, i.e., that are serial (Box 8.5). There is an inequality called *Amdahl's law* that says that the increase in speed you get by parallelism is limited by the portions of the computation that are sequential. That is where the delays are. That is where you have to wait. The best you can possibly do if you have P degrees of parallelism is to get a P-fold increase in speed; the advantage of more processors is going to be diminished by the sequential portions.

So, in figuring out how to find the parallelism in a computation you have to find it all to really gain

BOX 8.4 CREATING A PARALLEL PROGRAM

- Identify work that can be done in parallel
- —Requires high-level strategy
- —Could depend on input data
- —Insufficient parallelism or unequal task sizes cause **load Imbalance**
- Partition work and perhaps data among processes and processors
- —Excess communication and synchronization reduce parallelism
- —Minimize communication by maximizing locality (and doing redundant work)
- Manage the data access, communication, synchronization, and scheduling

the benefit. Finding wonderful parallelism for a while and then having the world stop while the computation reconfigures itself can cause an enormous degradation in end-to-end performance.

BOX 8.5 PERFORMANCE GOAL

Maximize speedup due to parallelism

$$\text{Speedup}_{\text{prob}}\ (\text{P proc}) = \frac{\text{Time to solve problem with "best" sequential solution}}{\text{Time to solve problem in parallel on P processors}}$$

Amdahl's law
Let s be the fraction of total work done sequentially

$$\text{Speedup}\ (P) \leq \frac{1}{S + \frac{1-S}{P}} \leq \frac{1}{s}$$

Even if the parallel part speeds up perfectly, it may be limited by the sequential part.
Problem size is critical!

Finally, let me mention one more complication. I said earlier that one must partition the work across the processors, to achieve parallelism, and also partition the data across processors, to reduce communication latency. Alas, these two strategies can conflict (Box 8.6). If there is more simultaneous computation, using local data, then there can be more data that needs to be pushed through the network to integrate the results of the parallel computations. Moving data around only indirectly advances the computation, and too much slow data movement can undo the benefits of high degrees of parallelism. And again, the trade-offs are different for different architectures.

The research challenges in Box 8.7 deal mostly with performance. Given a new problem that we want to solve, can we, the chemical scientists, find enough parallelism in the problem to be able to

BOX 8.6 LOCALITY AND LOAD BALANCE TRADE-OFF

- Optimizations that increase parallelism decrease locality
- —Subdivide data
- —Migrate computation
- Optimizations to improve locality decrease parallelism
- —Move everything near one processor
- Hard problem is optimizing both
- Non-uniform costs make it even harder!

exploit the platforms that are coming along? That depends on the problem, but it also depends on the strategy for solution.

BOX 8.7 RESEARCH CHALLENGES

- Increased speed comes from greater parallelism—can we find sufficiently parallel problem solutions?
- How much of the details can be hidden from the applications programmer?
- What are natural ways to express computation that also yield efficient code?
- How can platform portability be achieved?
- How much reuse is possible, e.g., via libraries and packages?
- How are debugging and tuning done (how does the user understand what's going on)?

To what extent can the computational science and computer science community hide from the applications programmer the lower-level details? Can we hide details about sending a message, receiving a message, synchronizing, and so on so that it doesn't clutter up the thinking about what is going on at the higher-level strategy? How can one describe these calculations so that they are readable by chemists, but so the description will still enable the generation of efficient code, meaning that there is a path that allows all that lower-level parallelism to be exploited as well? How can we do that so that it transcends changes in platform, so that once you have a strategy, it has some persistence over some of these changes in latency, in bandwidth, and in various aspects of performance?

Given that the strategy shift is a platform shift, how much can we continue to depend on libraries or on reuse of codes that have actually had a great deal of intellectual investment in them? Will they continue to give us the benefits that we want to believe they have, or does the performance simply degrade almost without people noticing until it is too late? Finally, given these complicated systems, such as the ASCI system that Paul Messina has described, how does one get a grasp on what is going on overall? How do we understand what the overall performance is and what contributes to it, and how does one get a mental model of what to do when things seem to be going wrong?

Now, as if that weren't complicated enough, the scientific community is becoming even more ambitious. Not only is there attention to high performance in the sense of computation, but now we also want to build computations, simulations in particular, that use massive amounts of data (Box 8.8).

BOX 8.8 NOW GENERALIZE THE SITUATION

- High performance
- Massive amounts of diverse data at many sites
- Diverse platforms and devices
- Collaborations
- Really big and challenging problems

Those data are gathered in a variety of ways. They live in a variety of places around the world and are represented in very diverse ways. We want to carry out simulations and modeling using components created by other groups, using other platforms and other data representations.

As we contemplate building these large, coupled, data-intensive simulations and models, the platforms that we might want to couple together to solve very hard problems are diverse. It is difficult enough to think about solving one application on one of these complicated parallel computing systems; now we want to reach out and get data from some other computational platform that has different characteristics.

Creating ambitious simulation and modeling systems requires collaboration. Collaboration is not just among people but also between people and devices, people and instruments, people and computational platforms. The problems that we want to solve are getting much bigger as well.

Consider the following situation, illustrated by two figures that come from Andrew Grimshaw at the University of Virginia. We are now in a situation in which, given the ascendancy of networks and the promise that they are going to get even better, we can think of doing an electronic experiment, having an electronic laboratory that can reach out electronically to other sites that have diverse, possibly unique capabilities (Figure 8.5). For example, there can be ways of doing real-lime data collection from observational data obtained from instruments that are attached to the network, and data repositories that exist at various remote sites. Figure 8.5 shows a map of the United States, but the resources might be anywhere in the world. In order to use them there needs to be some way of getting all these very diverse pieces to fit together. If you are a chemist carrying out an electronic experiment, you configure the experiment conceptually, assemble all the pieces, and describe at a very high level what it is you want to do, where you are getting the resources from, and how you are going to use them.

What a number of people are doing is building systems, such as the Legion system pictured in Figure 8.6, which provide so-called *meta-computing* capabilities. Computing at that level consists of staging the experiment: assembling the pieces, getting permission to use them, and getting them to interoperate, i.e., to work together. The vision is that the user sitting at a workstation puts this virtual environment together, identifies all the components, runs the electronic experiment, and looks at the results.

In addition to all the performance issues that are going to get even worse, there are issues about security. There are issues about fault tolerance—if you finally manage to get all the pieces coordinated and get your experiment going, you don't want it to die because one component failed, one processor went down. The way the network remains robust is by finding an alternative path if one path gets

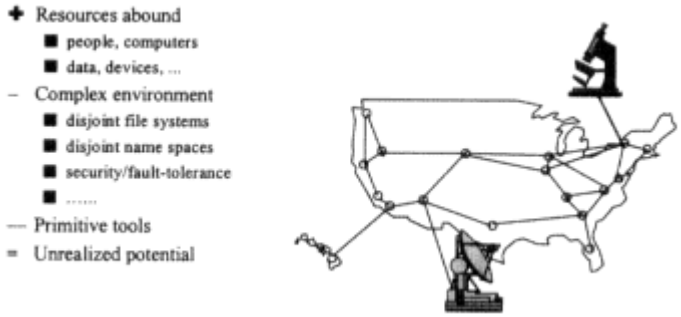


Figure 8.5
The opportunity.

blocked. The goal is to do that in a transparent way, so that the user focuses attention on the experiment and not on the details that are below the surface of what is intellectually important in solving the problem.

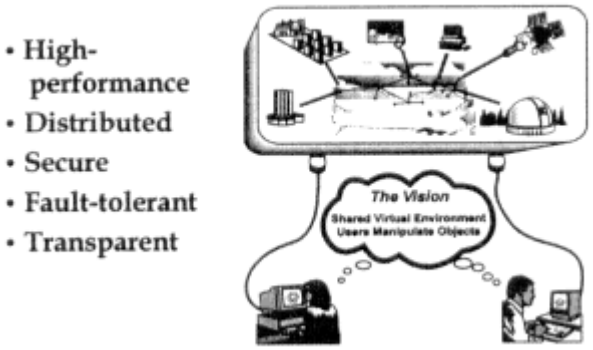


Figure 8.6
The Legion vision—one transparent system.

There are also some non-technical issues of how we live in this emerging world (Box 8.9). There is a serious concern among the community about where the people are going to come from who are going to have the knowledge and skills to allow us to live in this computational world. The strategies one needs to solve scientific problems in the kinds of computing environments I have described require deep knowledge of the domain, in this case chemistry, and also deep knowledge of information technology in order to put things together and make it all work.

Ideally, one would have very expert computational chemists, people who are wonderful chemists and wonderful information technologists and understand how to put those two areas together. An alternative is to have multidisciplinary teams. To make multidisciplinary teams work, there are a lot of sociological issues that have to get solved about mutual respect for what the other side does, about talking the same language, about understanding how to identify what the real scientific and technical problems are. We are making progress there, but we still have a ways to go.

BOX 8.9 MORE ISSUES

- Where will the software developers Come from?
- —Strategies require both deep domain knowledge and deep information technology knowledge.
- —Both expert computational chemists and multi-disciplinary teams are needed.
- Where will the languages, libraries, and tools come from?
- —Some needs are generic and will be met by the marketplace.
- —Scientific computing is not an economic driver nor an important market—don't count on commercial vendors.
- Where will the computing and communications resources come from?

BOX 8.10 PITAC RECOMMENDATIONS: EXECUTIVE SUMMARY

Findings:

- Information technology is vital to our well-being in the 21st century.
- Total federal information technology R&D investment is inadequate.
- Federal IT R&D is excessively focused on near-term problems.

Recommendations:

- Create a strategic initiative in long-term IT R&D.
- Increase the investment for research in software, scalable information infrastructure, high-end computing, and socio-economic and workforce impacts.
- Diversify modes of support.
- Establish an effective structure for managing and coordinating IT.

SOURCE: Summarized from the President's Information Technology Advisory Committee's report to the President, *Information Technology Research: Investing in Our Future*, February 1999 advance copy, National Coordination Office for Computing, Information, and Communications, Arlington, Va.

Where are the languages, the tools, and the libraries going to come from to do computation in this increasingly complicated world? There are needs that aren't peculiar to chemistry, and there are needs that aren't peculiar to science, and a lot of those will come from the marketplace. That is going to give us a lot of the networking technology we need.

There are other issues that are more peculiar to what we do in computational science. Scientific computing does not drive the market, and so we cannot expect that over time the vendors are going to step up and provide the specialized software solutions that are essential to at least part of what we are trying to do. That means that the scientific community is going to have to find a way to develop that software, and we are going to have to find a way to do it that provides high-quality robust software and doesn't consume all of everybody's time.

Finally, there is an issue that I know many people in the chemistry community have struggled with, namely, that even though in principle we can build these very high performance, very powerful systems such as the one Paul Messina described, how is the average bench scientist going to get access to them?

In closing, let me just say a little bit about the committee I am on, the President's Information Technology Advisory Committee (PITAC). When the High Performance Computing and Communication Act was passed in 1991 or so, part of the legislation said that there was to be an advisory committee for high-performance computing. It took until 1997 for the committee to be established and, by the time it was appointed, there were lots of issues on the table besides high-performance computing. The committee was given the additional task of looking at the next-generation Interact program that was then emerging in the government, and then broadened its agenda to look at information technology overall.

The committee is drawn primarily from the computer science and communications areas. It issued an interim report in July and will issue a final report early in 1999. The high-level findings and recommendations are shown in Box 8.10.

BOX 8.11 SOFTWARE RESEARCH

Findings:

- Demand for software far exceeds the nation's ability to produce it.
- The nation depends on fragile software.
- Technologies to build reliable and secure software are inadequate.
- The nation is underinvesting in fundamental software research.

Recommendations:

- Fund more fundamental research in software development methods and component technologies.
- Sponsor a national library of software components.
- Make software research a substantive component of every major IT research initiative.
- Support fundamental research in human-computer interfaces and interaction.

Make fundamental software research an absolute priority.

SOURCE: Summarized from the President's Information Technology Advisory Committee's report to the President, *Information Technology Research: Investing in Our Future*, February 1999 advance copy, National Coordination Office for Computing, Information, and Communications, Arlington, Va.

The investment in research and development in information technology is not keeping up with the growth in the importance of the area. Furthermore, because of the tension caused by the shortage of money, the investment in R&D is increasingly short term. In other words, if we spend our money only to solve today's problems, which have measurable milestones and goals, we shortchange the longer-term investment in developing the new ideas that are going to fuel us in 10 or 15 years.

There are also recommendations concerning how to educate more people and how to give the average taxpayer access to computing and the like. I don't mean to dismiss these recommendations, but they are less relevant to this workshop than the research recommendations.

When this committee first met, everybody, no matter what field they came from, said, “The real problem is software." Our recommendation is that the research investment of the government be increased, especially in software, in scalable information infrastructure, in high-end computing, and in work force and socioeconomic issues. By scalable information infrastructure we mean not only networking, but also all of the information that lives on our networks—all the Web pages, databases, and mail systems; everything that involves using information in a networked environment. Thus there are significant software issues there as well.

In making software research a priority (Box 8.11), as with any grand challenge problem, you have to worry about more than how you are going to solve that problem with any strategy you can figure out. We also seek to develop some underlying solution technology that is reusable and that will allow you to solve next year's problems and the problems of the years after.

The committee work is an ongoing process. We are currently in the process of gathering feedback from all of the communities we talk to, so that the final report can be as strong as we can possibly make it so that we can actually see this exciting computational environment move ahead.

Acknowledgment: My thanks go to my colleagues Kathy Yelick, Jim Demmel, and David Culler for sharing the slides from their Berkeley graduate course in parallel computing, CS267.

DISCUSSION

Barbara Garrison, Pennsylvania State University: I do computational chemistry, and I am inherently limited by Amdahl's law, yet I have problems where I could use parallel computing, but at times I feel like I am wasting cycles because of imperfect scalability. Has there been any effort to design hybrid serial parallel machines or operating systems where there could be sharing of the parallel resources so we are not in fact wasting the cycles?

Susan Graham: That is what the old operating systems notion of multiprocessing is really all about. The system can detect when you are waiting for something, and it can go and run somebody else's program meanwhile.

The complication when you now get these memory hierarchies is that if the system is going to run somebody else's problem while yours is waiting for its data, then the staging is going to fetch their data and flush your data out of the caches. That will affect your performance as well. Thus what you suggest can be done, but the extent to which it is really going to help is something that we don't totally understand. Having the system wait for you is going to be the most efficient thing for your job, but there is a question of throughput and how you share that resource with other people.

Evelyn Goldfield, Wayne State University: I have two questions. I am, also, a computational chemist who uses parallel programming and computing. One of the things that I have been wondering about is the different architectures. I use massively parallel distributed memory computers, but all these shared memory computers are coming along. I wonder to what extent the memory is actually shared, and to what extent we really do have to change our computing paradigms. How important is it for people to have different programs for different computing paradigms for different types of architecture? If we have to have completely different programs for different machines, it is going to be quite dismaying for people.

Susan Graham: This, again, is something that we understand imperfectly at this time. If you have distributed memory solutions, distributed memory strategies, it is not that hard to get them to do well on shared memory systems. It is going in the other direction that is harder, in my experience.

What one really wants, and this may not be something that is completely within one's control, is to have a certain amount of adaptivity. If your program makes it apparent where the parallelism is, where the opportunities are, and if a compiler can figure out where the communication has to be, where the sharing is and so on, then you want the system software to adapt the program communication, synchronization, etc. to the architecture. So, it is at the strategy level that it is really most important that you are not locking yourself in at the high level to particular details of the architectural model.

Evelyn Goldfield: I have one other question that I think is really key and that is, I think, in the minds of a lot of chemists. We are willing to waste computer time if it is a choice between wasting the computer's time and wasting our own time. The only thing we really, really care about, truthfully, is how long it takes us to get the job done as long as we get the cycles. I thought of this when you talked about load imbalance. You can have a lot of load imbalance and still get your job through on time. How much are chemists going to have to really worry about these computer science questions if you don't want us to waste the cycles? It seems to me it is the computer scientists that have to come up with that solution.

Susan Graham: It is not that we don't want you to waste the cycles. Somebody once said to me, "You know, people have telephones, and they don't worry about keeping the telephone busy all the time. It is an appliance, and it is there when they need it." That gets back to the kinds of problems you are trying to solve, and the economics. If you can afford to have a machine that sits on your desk that is powerful enough for what you want to do, then the primary issue is going to be ease of making it do what you want it to do. But there are chemists who have problems that even the ASCI System can barely handle. They cannot afford to have part of the system sitting idle, particularly when their access to it is once every so often for a limited amount of time.

Thom Dunning, Pacific Northwest National Laboratory: I had a couple of comments on your presentation. One is the emphasis on software that came out of the PITAC report. I have always felt that one of the major limitations in realizing the potential of the computing systems that we have had and are going to have in the future is software. So I was absolutely delighted to see this committee recognizing the importance of software, because it is so easy in a system like we operate in to look at the hardware, a physical object, and not recognize the fact that the hardware is absolutely useless without the software. Also, one comment relative to the question that Evelyn Goldfield asked is that we at Pacific Northwest National Laboratory, and people in other places, have developed computing infrastructures. Ours are called global arrays that actually run on distributed memory computer systems, shared memory computer systems, and a collection of workstations, all of which is entirely transparent actually to the applications programmer. This is the case because we have computer scientists who have implemented global arrays on all of these different types of architectures, and the only thing the applications programmer has to do is issue the calls to those particular subroutines. So, there are ways that one can actually write software that performs well on a number of different architectures.

Now, I am not at all clear that that is going to perform well on the types of architectures that Paul Messina described, but clearly many of these problems have been solved, and I have confidence in the creativity of both the chemistry and the computer science community and that we will see it solved for the types of very large systems that Paul described.

Robert Lichter, Camille & Henry Dreyfus Foundation: As a non-computational chemist and a non-computer scientist I want to thank you for an extraordinarily lucid description of how these things work. I think this is the first time I have run into the topic in a way that is comprehensible.

I was also struck by how much chemistry you do know, because the strategy that you described for solving problems—doing isolated computations and then pulling them together—is very much the way a synthetic organic chemist synthesizes a complex molecule. You do little pieces and then glue them together.

When looking at the kind of global picture of marrying hardware and software, we are limited by what exists. I would be very much interested in your wildest vision as a computer scientist of what could exist either in hardware or in software that we haven't even begun to think about now.

For example, one thing that even I am aware of is the concept of DNA computing, which nobody has talked about here. I don't know whether that's because it is not worth talking about, or whether it is not developed enough to talk about, but that is the kind of thing I'm referring to. I'm just curious to see if you could speculate wildly.

Susan Graham: I don't know whether I can do that. I can comment a little about DNA computing and things like that. That is one example of the part of the research agenda that we feel has been neglected of late. In other words, DNA computing is wildly speculative in the sense that the computational model

is totally different, and yet the attempt is to draw it from nature, from something that exists, and it potentially gives huge amounts of parallelism.

The issue from a computer science point of view is figuring out what the algorithms might be that would actually do well in that computational model, and of course, there are issues on how you build such a computer. I think it is really important to explore those directions.

There are people in my field who have wonderful imaginations about these things. I am not one of them and so I don't want to take up your immediate challenge except to say that I think you are right. People are most comfortable thinking sequentially. People are most comfortable thinking in the way I described, take some data, move it from here to there, and so on. We have to break out of that a little bit to at least experiment and see what might happen if we had a very different model.

Gintaris Reklaitis, Purdue University: You described an interesting model of a real-time environment in which you gather data from different sources and different instruments, run the data through a model, and then act upon the results. This is very much along the lines of the supply chain of interest in process operations. In your case, the information is obtained asynchronously, yet the computational models that you describe operate synchronously. Although you are parallelizing the computational tasks, under the synchronous model you are forced to wait for the slowest task to execute and you do not use all of the latest information in executing the tasks. Is there any work in progress on asynchronous parallel computing models?

Susan Graham: I was describing synchrony as a problem. There are times when you really have to worry about the temporal order in which things happen. It is possible, for example, in my shared memory situation that some of the processors are just filling that memory with interesting stuff while other processors are going on and doing their business and not worrying about that until they are ready. Asynchronous models in which one is notified—or the status is posted and whoever cares can look and find out the status—are actually in some ways much more comfortable. They can be easier to build, but they are harder sometimes for people to think about.

Now, it is possible we are not puffing as much attention on that as we might, and that is where the interaction with the application domain is so important. You describe to me what you want to do and then I start thinking about how can I help you do that.

9

Collaboratories: Building Electronic Scientific Communities

Raymond A. Bair
Pacific Northwest National Laboratory

ABSTRACT

High-speed computation now provides the means to examine and simulate systems at unprecedented levels of detail and accuracy. The combination of computation with large-scale databases enables analysis of the prodigious volumes of data coming from today's experiments and simulations. However, when these enabling technologies are coupled with new capabilities in communications, an opportunity is created that can revolutionize not only the scope but also the process of scientific investigation. In physical science research, new distributed computing and communications technologies are being employed that enable researchers to access data, instruments, and expertise independent of their location.

While the term “collaboratory" (or "virtual laboratory") is often used to refer to a set of technologies, perhaps the most significant impact of collaboratories will be the generation of new opportunities to create and sustain active scientific communities. The development and adoption of electronic collaboration capabilities will provide geographically distributed research teams with greater abilities for the organization, close-knit interaction, and rapid response, needed to address increasingly challenging research problems. This paper examines some of the opportunities and challenges presented by scientific collaboratories, and the interplay between emerging collaboration technologies and the research communities they support. Experiences to date point to requirements and success factors for virtual facilities. Examples are drawn from technology development and chemical/ materials pilot collaboratory projects of the U.S. Department of Energy.

INTRODUCTION

One of the scarce resources in chemical research is time—for scientists, instruments, and com

NOTE: Pacific Northwest National Laboratory is a multiprogram national laboratory operated by Battelle Memorial Institute for the U.S. Department of Energy under Contract DE-AC06-76RLO 1830.

puters—to explore and understand complex phenomena and to unlock the principles governing them. Advances in computing and communications systems are having profound impacts on the capabilities that we can bring to bear on research and development problems, providing extraordinary instrument control and data acquisition capabilities, powerful data analysis and visualization capabilities, and simulations capable of ever more detail and scope. This aspect of the computer revolution is rapidly magnifying our science capabilities while reducing the time needed to perform measurements and simulations. However, outside of these areas wholly new impacts of computing advances are emerging, which will dramatically expand our options for using our time to organize and conduct scientific efforts.

The term "collaboratory" (*col*aborate + *laboratory*) is attributed to William Wulf, who envisioned the potential impact of the information age on science, creating a ". . . 'center without walls,' in which the nation's researchers can perform their research without regard to geographical location—interacting with colleagues, accessing instrumentation, sharing data and computational resources, [and] accessing information in digital libraries."[1] Other terms that are used almost interchangeably with collaboratory are "virtual laboratory," “laboratory without walls,” and "collaboratorium." They all encompass the use of information and communication systems to remove barriers of geographic distance and time from research collaborations, not just scientists working remotely, but working *together* regardless of their location. A major emphasis of collaboratories is natural, informal work processes, going beyond text exchange and presentation metaphors, to in-depth, collaborative work.[2]

Collaboratories have potential roles in all stages of the scientific process, from the initial planning and organization of a new project idea and project team, to the design of the experiments and development of software, to the execution of those experiments and simulations and their analysis, to the preparation and dissemination of the results. However, one does not simply *deploy* a collaboratory like a desktop publishing program; one *builds* a collaboratory with scientists, information, and tools. The collaboratory tools required are varied and challenging to develop, requiting both generic capabilities like video conferencing and screen sharing, and domain-specific capabilities to handle the manipulation and display of data types particular to each type of scientific work. Integration is a major component of collaboratory development, spanning groupware, legacy modeling and analysis applications, instrument software, files, and databases. Because of their unique requirements, collaboratories are often leading-edge examples of knitting together new distributed systems technologies.

Collaboratories are an emerging capability that provides new resources for chemical science. This paper provides an overview of collaboratories from the perspective of scientific research, discussing opportunities for collaboratories, examples of the use of collaboratories in chemistry and related disciplines, the kinds of software that are being developed for collaboratories, the impacts that collaboratories are having, and the requirements and prospects for the future.

OPPORTUNITIES FOR COLLABORATORIES

There are a number of arenas that are fertile ground for the development of collaboratories, particularly among scientific user facilities and institutes that provide unique and specialized resources to the scientific community. Although the examples given in this discussion are drawn from the U.S. Department of Energy (DOE) arena, there are many analogous scenarios in our university and industrial

[1] National Research Council, Computer Science and Telecommunications Board, *National Collaboratories: Applying Information Technologies for Scientific Research*, National Academy Press, Washington, D.C., 1993, p. vii.

[2] R.T. Kouzes, J.D. Myers, and W.A. Wulf, 'Collaboratories: Doing Science on the Internet," *IEEE Computer*, 29(8), 40-46, August 1996.

EXAMPLES OF COLLABORATORIES

communities, as well as other government agencies. DOE builds and supports a wide range of national scientific user facilities, ". . . built with the express purpose of being available for the performance of research by a broad community of qualified users."[3] National scientific user facilities make unique research resources available to DOE scientists and researchers from academia, industry and other federal laboratories, and provide opportunities needed to educate and recruit young scientists to meet the demanding challenges of the future.

Figure 9.1 shows 19 of the DOE scientific user facilities most often used for chemical, biological, and materials research. They include four synchrotron radiation light sources, five high-flux neutron sources, four electron beam characterization centers, and five other centers, two specializing in DOE missions of environmental science and combustion. Eighteen of these facilities are operated by DOE's Office of Basic Energy Sciences. The Environmental Molecular Sciences Laboratory in Washington State is operated by the DOE Office of Biological and Environmental Research. The Spallation Neutron Source in Tennessee is under construction. Each of these facilities supports research by individual investigators and collaborative teams from across the country and around the world. Today, scientists often travel to a facility to do their work. Many have limited time and resources to travel and must carefully optimize their experiment time, often limiting their benefit from the facility. The facilities themselves are widely dispersed, and increasingly often scientists need to use more than one facility in the course of a complex investigation.

Collaboratories can increase the effectiveness and value of user facilities like these in many ways. More can be done before the scientific team arrives on site, e.g., detailed planning of the research campaign, and training for the specific equipment to be used there. While the team is on site, communication is enhanced with colleagues at the home institution and collaborators at other institutions. For example, although a professor may not be able to stay long at the user facility to mentor his/her students, collaboratory capabilities facilitate following the detailed progress of the work remotely, and helping with problems as they arise. After scientists leave the user facility, shared analysis and discovery are enhanced through collaboratory capabilities. Thus, collaboratories enhance the productivity of research. More complex problems can also be taken on, as collaboratories support the assembly of interdisciplinary teams. Expertise can be drawn from many more sources, including industry and smaller colleges. This ability to handle more complex problems can also have an impact on what science is done. Thus, by enhancing collaboration, collaboratories enable new scientific processes and new science.

Although there are enough commonalities between collaboratories to speak about them in general, scientific collaboratories are very individual, customized to the style of a research community and to the nature of their research. The essential suite of capabilities needed is particular to the kind of research being performed, as is the relative importance of those capabilities. The processes used to collect and analyze information are also diverse, and the collaboratory needs to reflect that. To date, we've only scratched the surface in building effective chemistry collaboratories; there's still much to learn. The development of a new chemistry collaboratory typically involves adding domain-specific capabilities to a base of collaboratory tools. However, the first step in the process is to learn about how the scientists work, what their information is like, how they need to store and share it, and where the communications and information management problems are that collaboratories might facilitate. Then the computer scientists can team with the chemists to develop and integrate the necessary tools and applications into a working collaboratory.

[3] U.S. Department of Energy, "Office of Energy Research Facilities," in *Pricing of Departmental Materials and Services*, Order DOE 2110.1A, Change 2, May 18, 1992, Chapter III, section 10.

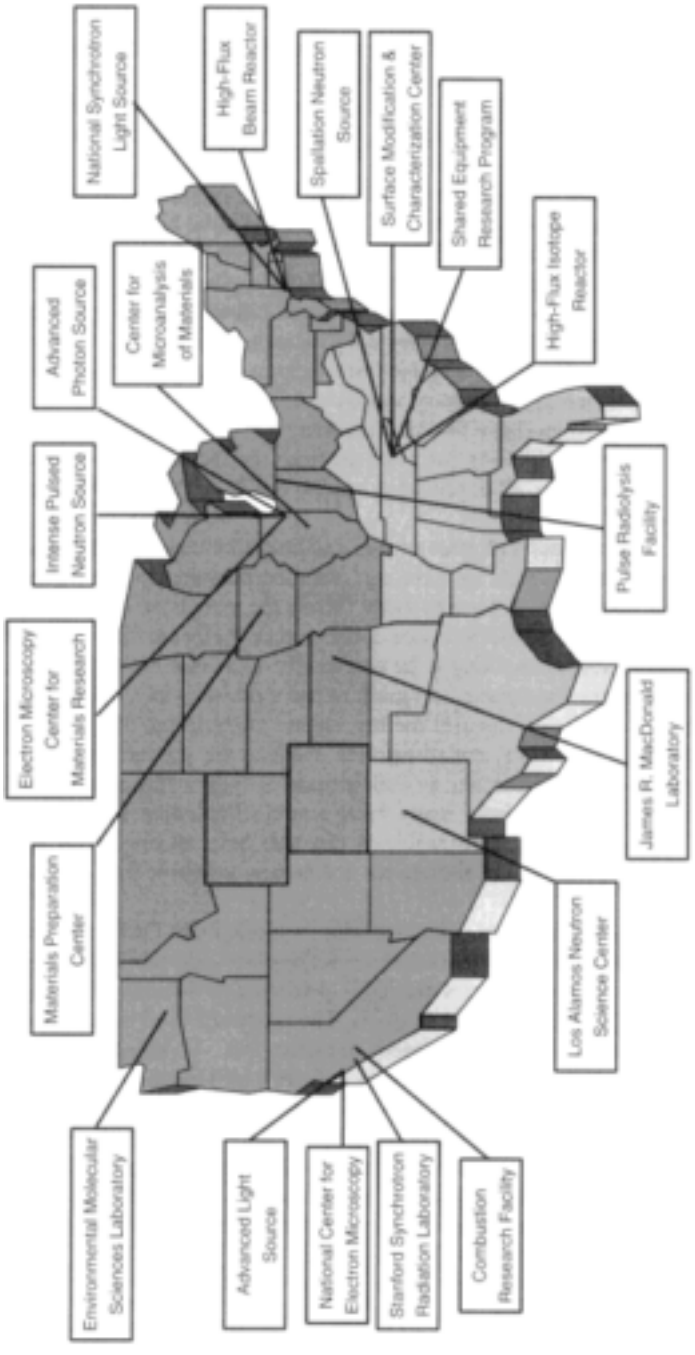


Figure 9.1
Distributed scientific facilities and researchers: selected DOE scientific user facilities often used for chemistry, biology, and materials research.

Experience with many groups indicates that each collaboratory seems to have an overall organizing principle, one of a small number of ways that people create and interact with research results. For example, a couple of pioneering collaboratories had quite different approaches. The Worm Community System[4] created an extensive centralized, shared data repository about a single biological organism, *Caenorhabditis elegans*.[5] Today's human genome databases are likely to evolve into such collaboratories centered on a community effort to create and understand the human genome. Another very successful early effort, the Upper Atmospheric Research Collaboratory (UARC)[6,7] is largely organized around community experimental campaigns, using instruments at a dozen remote experiment sites (including such inhospitable environments as Greenland). UARC continues into the next generation as SPARC.[8] More recently, several pilot collaboratories have been set up by DOE projects. Three of them provide good examples of the diversity and impact of collaboratories on the chemical and materials sciences: the Diesel Combustion Collaboratory,[9] the Materials Microcharacterization Collaboratory,[10] and the Environmental Molecular Sciences Collaboratory.[11]

The Diesel Combustion Collaboratory (DCC) assists the partners of the long-standing Heavy Duty Diesel Combustion CRADA, a collaborative research and development agreement among DOE researchers and diesel engine manufacturers. It involves four national laboratories: Sandia (SNL), Lawrence Berkeley (LBNL), Lawrence Livermore (LLNL), and Los Alamos (LANL), scientists at the University of Wisconsin, and three companies: Caterpillar, Cummins Engine, and Detroit Diesel. DCC is a part of the DOE2000 Collaboratory Pilot Projects program, and gives scientists capabilities that do not exist at any single location.[12] The DCC enhances the flow of information between and among the experimentalists and modelers at the national laboratories and the engine designers at the industrial sites. So, this collaboratory is focused around the information base of results from experiments and modeling runs, providing visualizations that the investigators share. The DCC also provides capabilities for industrial researchers to run chemical and numerical models and simulations remotely on DOE computers. Because proprietary industry research is involved, there is a significant concern about security, which must be addressed by the collaboratory tools. SNL has enabled the collaboratory partners to have a secure encrypted connection, to discuss engine drawings, experimental data, output from a model, or results of a visualization. They may share secure information or applications from a collaborator's computer. DCC also provides a shared workspace via the BSCW product,[13] an image library, a data archive, and shared electronic laboratory notebooks.

The Materials Microcharacterization Collaboratory (MMC) has a different set of requirements.

[4] B.R. Schatz, "Building an Electronic Community System," J. Management & Information Systems, 1992.

[5] The current Web site for C. elegans is <http://elegans.swmed.edu/>.

[6] C.R. Clauer, D.E. Atkins, et al., "A Prototype Upper Atmospheric Research Collaboratory (UARC)," Visualization Techniques in Space and Atmospheric Science, E.P. Szuszczewicz and J.H. Bredekamp (eds.), pp. 105-112, NASA SP-519, NASA, Washington, D.C., 1995.

[7] N. Ross-Flannigan, "The Virtues (and Vices) of Virtual Colleagues," Technology Review, pp. 52-59, March/April 1998.

[8] See the Space Physics and Aeronomy Research Collaboratory (SPARC) Web site at <http://www.crew.umich.edu/UARC/>.

[9] See the Diesel Combustion Collaboratory Web site at <http://www-collab.ca.sandia.gov/>.

[10] See the Materials Microcharacterization Collaboratory Web site at <http://aem005.amc.anl.gov/MMC/>.

[11] See the Environmental Molecular Sciences Collaboratory Web site at <http://www.emsl.pnl.gov:2080/does/collab/>.

[12] The DOE2000 program is described on the Web site at <http://www.mcs.anl.gov/DOE2000/>.

[13] The Basic Support for Cooperative Work (BSCW) product comes from the German National Research Center for Information Technology. See the BSCW Web site at <http://bscw.gmd.de/>.

This collaboratory was constructed by researchers at Argonne National Laboratory (ANL), LBNL, the National Institute of Standards and Technology (NIST), Oak Ridge National Laboratory (ORNL), the University of Illinois, and several instrument/computer manufacturers, including Gatan Inc., R.J. Lee Instruments Ltd., EMiSPEC Systems Inc., Philips Electron Optics, NSA—Hitachi Scientific Instruments, JEOL USA Inc., Sun Microsystems Inc., and Graham Technology Solutions Inc. MMC is also a part of the DOE2000 Collaboratory Pilot Projects program.[14] Each of the research labs had diverse microscopy capabilities when the project began, and several had developed remote microscopy tools.

A major goal of the MMC is to explore and develop a shared electronic virtual environment around a common theme of microscopy and microanalysis, encompassing leading-edge instrumentation and applied to both education and research. MMC has defined a common set of capabilities that are needed for electron microscopy, and is working on data models, graphical user interfaces, and application program interfaces (APIs) for the common architecture. Although essentially all commercial microscopes have computer control, previous generations had some features that were mechanically adjusted, and the instrument control programs were often difficult to drive from another application. MMC has been working with the instrument manufacturers to convey requirements for remote control. Essentially all are now implementing more useful and complete APIs in their products. MMC not only lets one control the microscopes, but also has established tools for sharing analyses among distributed collaborators, and comparing the images from multiple simultaneous experiments at different locations. The results of the work of the MMC are stored in electronic notebooks.

The instrument control architecture that MMC employs (Figure 9.2) is very similar to that for other remote collaborative instruments being developed in the chemical sciences. A commercial microscope may have multiple devices that can be controlled through serial interfaces or network interfaces (TCP/IP protocols). A local server accepts commands, validates them (ensuring the instrument is not operated out of its ranges), issues commands to the appropriate instrument actuators and sensors, and collects data from the spectral and image data acquisition systems. The user interface uses Web technologies so that scientists can interact with the instrument through their Web browser. Note that the user interface (client) is separate from the microscopy server, and the microscopy server is different from the instrument computer. This is key to flexible and extensible designs. Of course, all of the client and server processes can (and sometimes do) run on the same computer.

The Environmental Molecular Sciences Collaboratory is yet another type of collaboratory. The Environmental Molecular Sciences Laboratory (EMSL) is DOE's newest user facility, located at Pacific Northwest National Laboratory (PNNL). Instead of providing a particular kind of capability, EMSL is a collection of many unique capabilities and expertise for a particular mission, environmental molecular science. The focus is on developing a molecular-level understanding of the physical, chemical, and biological processes that underlie remediation of contaminated soils and groundwater, processing and disposal of stored waste materials, and human health and ecological effects of exposure to pollutants. EMSL has three major facilities: the Molecular Science Computing Facility, the High Field Magnetic Resonance Facility, and the High Field Mass Spectrometry Facility. Other specialized capabilities and facilities provide resources targeting research areas in nanostructural materials synthesis, interfacial structures and compositions, reactions at interfaces, and gas-phase monitoring and detection. Consequently many EMSL projects and collaborations cross disciplines.

The preparations for collaboratories at the EMSL began during its construction. EMSL's networks, computer security, shared file systems, and user services were designed to support both individual users

[14] The DOE2000 program is described on the Web site at <http://www.mcs.anl.gov/DOE2000/>.

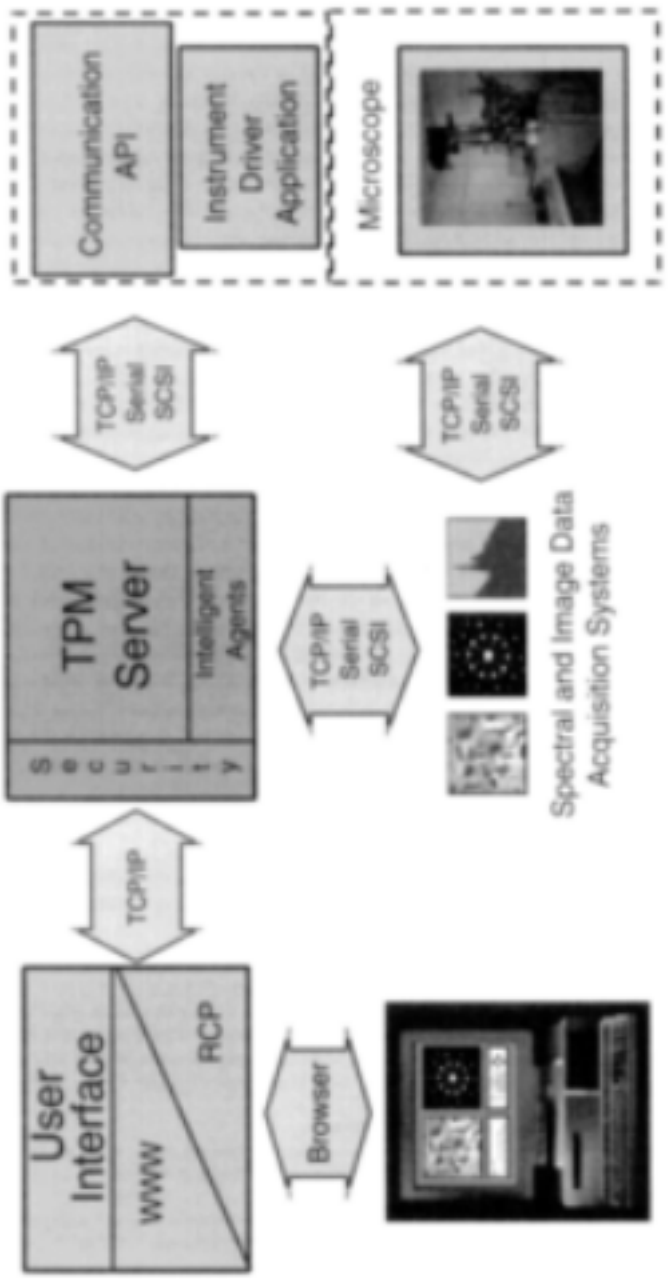


Figure 9.2
Instrument control architecture for the telepresence microscopy instruments at Argonne National Laboratory.

and teams, internal and external.[15] A Scientific Data Management (SDM) system was developed to manage EMSL's 20-terabyte robotic tape archive.[16] SDM captures and stores metadata (information about the data) so that needed files can be located easily, even by scientists in other disciplines. Collaboratory tool development began in 1993 and is currently in its third generation, now part of the DOE2000 National Collaboratories program. User support is provided for new collaboratory tools as they are deployed. This is very important, as collaboratory tools operate in a complex environment and scientists are often not accustomed to using them. EMSL's in-house Instrument Development Lab (IDL) provides custom instrument electronics and software development capabilities. Remote operation, "fly by wire," automatic metadata capture, and automatic data archival capabilities are becoming a routine part of IDL instrument designs and upgrades. Together, these capabilities provide the facility computing infrastructure upon which collaboratories can be built.

Although many collaboratory projects have begun in EMSL, there are three highly developed examples. The Virtual NMR Facility provides a set of extensions to the generic collaboratory tools for NMR spectroscopy. [17] This collaboratory provides secure remote access to operate EMSL NMRs, employing the commercial console software provided by Varian (VNMR). An NMR Spectroscopist's Notebook adds electronic notebook capabilities to view instrument parameters and three-dimensional molecular structures, and to capture NMR spectra from the spectrometers. In the custom instrument arena, EMSL developed an On Line Radio Frequency Ion Trap Mass Spectrometer.[18] All of the primary instrument parameters can be controlled remotely, including sample injection. The instrument server ensures that only reasonable parameters are passed to the spectrometer. Two versions of the software are available, a Visual Basic version that is stand-alone and a Java version that permits shared remote operation as part of the CORE2000 tools, discussed in the next section. Educational uses of collaboratories are also being explored in undergraduate and graduate programs. For example, the RF Ion Trap is regularly used in remote lectures and experiments by undergraduate chemistry classes in the Pacific Northwest. The Collaboratory for Undergraduate Research and Education is a consortium of colleges and universities formed to explore opportunities for collaboratories to have an impact in education through workshops and pilot programs.[19] It is already clear that collaboratories will have a substantial impact on curriculum enhancement, faculty development, and undergraduate and graduate research.

TOOLS FOR COLLABORATION

The major technologies that distinguish a collaboratory from remote computer/instrument access are the collaborative tools that permit scientists to share the scientific process. It's easy to see that tools

[15] The capabilities of the EMSL are described on the Web site at <http://www.emsl.pnl.gov/>.

[16] D.R. Adams, D.M. Hansen, K.G. Walker, and J.D. Gash, "Scientific Data Archive at the Environmental Molecular Sciences Laboratory," Proceedings of the Sixth Goddard Conference on Mass Storage Systems and Technology, GSFC/CP-1998-206850, pp. 409-417, March 1998; and D.M. Hansen and D.R. Adams. "A Database Approach to Data Archive Management," Proceedings of the First IEEE Metadata Conference, IEEE Computer Society Mass Storage Systems and Technology Technical Committee, IEEE Computer Society Press, Los Alamitos, Calif., 1996.

[17] See the Virtual NMR Facility Web site at <http://www.mcs.anl.gov/DOE2000/>.

[18] See the On Line Radio Frequency Ion Trap Mass Spectrometer Web site at <http://eol.emsl.pnl.gov/>, and J.M. Price, M.V. Gorshkov, J.A. Mack, and B. Rex, "An Internet Accessible, On-line Ion Trap Mass Spectrometer for Collaborative Research," Proceedings of the 44th ASMS Conference on Mass Spectrometry and Allied Topics, p. 1175, 1996.

[19] J. Myers, N. Chonacky, T. Dunning, and E. Leber, "Collaboratories: Bringing National Laboratories into the Undergraduate Classroom and Laboratory via the Internet," Council on Undergraduate Research (CUR) Quarterly, 17(3), 116-120, March 1997.

that support real-time interactions will be important in building collaboratories—tools that let scientists carry on the research process when they are geographically apart, with all of the richness of the interactions they can have when they are together in the same room. However, such real-time tools cover only one aspect of scientific work. When we collaborate, we don't always do everything together, yet it is vital that we be able to share the record of what we have been doing (e.g., parameters, data, theories, simulations, analyses), our views on and additions to the contributions of others, and also the record of discussions that have occurred (especially when one or more of the collaborators was not present).

Electronic Laboratory Notebooks

Traditionally, the primary record of the scientific process has been the ubiquitous laboratory notebook. The lab notebook has existed through the ages—Leonardo da Vinci's notebooks are a great example. However, if one looks at notebooks through the ages, an interesting change can be observed. For centuries, the lab notebook was the complete record of a scientist's work. Everything was within the pages of the notebook: theories, proofs, equipment designs, experimental conditions and results, analyses, and conclusions. Today's notebooks often exclude the raw data because it would take up too much space and it's only read by computers anyway. Instead there are references to external data archives: files, tapes, floppies, etc. Tables, charts, and graphs are produced by computers, printed, and pasted in. Increasingly the lab notebook contains less and less of the full scientific record.

On the other hand, computers can easily manage all of those data forms, and more. The concept of an "electronic lab notebook" has been around for a while. However, it has been hard to execute in software. Only in the last couple of years have we begun to have the tools to build an electronic notebook without heroic effort. Web standards provide a ubiquitous interface, and new object languages and distributed object standards support facile interoperability (the ability of different applications to exchange information and work together without custom protocols) and extensibility (the ability to add features to an application without getting into the "guts" of it).

Lab notebooks have many roles:

- Science observations,
- Design notebook,
- Instrument log book,
- Experiment log book,
- Legal record,
- Notepad, and
- Group workspace.

Most notebooks play more than one role. Our electronic lab notebook will need to hold lots of different kinds of data—from instruments, simulations, analysis results, and two-dimensional and three-dimensional visualizations. It also needs to be able to store data files in forms that preserve all of the information (or perhaps a link to the original data), as well as abridged summaries of data in the forms of tables, images, charts, etc. There's also information that individual scientists enter, such as text and sketches, and information that comes from group activities such as presentations, conversations, and planning sessions. An electronic notebook can capture these without a lot of effort. For each of these kinds of data, it's crucial to keep some metadata—information about the data, for example, who made it, when, the chemical system under study, and experimental parameters such as temperature, laser fre

quency, or acceleration potentials. This metadata is the key to a real strength of electronic notebooks. We want the notebook software to be able to retrieve anything we need from the notebook without paging through it, and also organize that information into useful forms. Metadata makes that much easier and more accurate than "full text" searching (which can also be done). The collection of metadata should be as automatic as possible, relieving the scientist from the tedium of recording things like instrument parameters.

The DOE2000 project is building technologies for just such a notebook, and prototypes are in use today.[20] This is a collaboration between three national laboratories: Pacific Northwest National Laboratory (PNNL), Oak Ridge National Laboratory (ORNL), and Lawrence Berkeley National Laboratory (LBNL). Defining shared standards for notebook data has been a challenge, but the DOE2000 notebook has standard APIs for adding editors and viewers to any notebook, and for exporting and importing notebook data. The electronic notebook provides a secure, shared Web-based space, interactive input, and rich media types. It is modular and extensible. The prototypes are exploring additional features ranging from sophisticated querying and searching capabilities, to automated notification of new contents to the collaborators, to mobile, off-network use.

Notebooks are also an essential repository of intellectual property. One of the most pressing issues in electronic notebooks is that of legal defensibility. Technically, signing and witnessing a page of an electronic notebook (or an object in the notebook) is not difficult. Authentication and digital signature technologies being developed for banking and commerce can handle the job nicely. However, there are aspects of the legal defensibility of electronic records that have not been tested in court. CENSA, the Collaborative Electronic Notebook Systems Association, is an industrial consortium promoting the development of commercial electronic notebook systems, with a large fraction of its partners from chemical and pharmaceutical companies. [21] CENSA aims to more rapidly advance the state of the art in electronic record keeping in ways suitable to large-scale deployment and preservation of intellectual property. One of CENSA's programs involves dialog with federal agencies and regulators around the issues of legal defensibility.

When scientists put information into an electronic notebook, they would certainly like to be able to retrieve it later. That's not a problem on the short time scale, but what about 25 years from now? The issue is not the media. The information technology industry is good at managing the process of migrating data from one media to another, as disk and tape capacities grow. The issue is the format of the data. Will we have the programs to read the data in the future? Achieving a reasonable degree of longevity will require planning. Document storage software providers are going to need to guarantee that their software will continue to read today's formats many years from now. We'll also need methods for document translation, methods that preserve the digital signature and legal defensibility of a document. Market forces should drive document technologies toward a solution. Scientists are not the only ones who need these capabilities, so there's a lot of incentive to solve document problems. However, much of the contents of a notebook is data, so scientists have a responsibility, too. We will need to archive data specifications and/or software applications more scrupulously.

Real-Time Interactions

Mapping the many ways that scientists interact face to face into the Internet world is very challenging (Figure 9.3). When they talk about science, many different kinds of media come into play: speaking

[20] See the DOE2000 Electronic Notebook Project Web site at <http://www.epm.ornl.gov/enote/>.
[21] See the Collaborative Electronic Notebook Systems Association (CENSA) Web site at <http://www.censa.org/>.

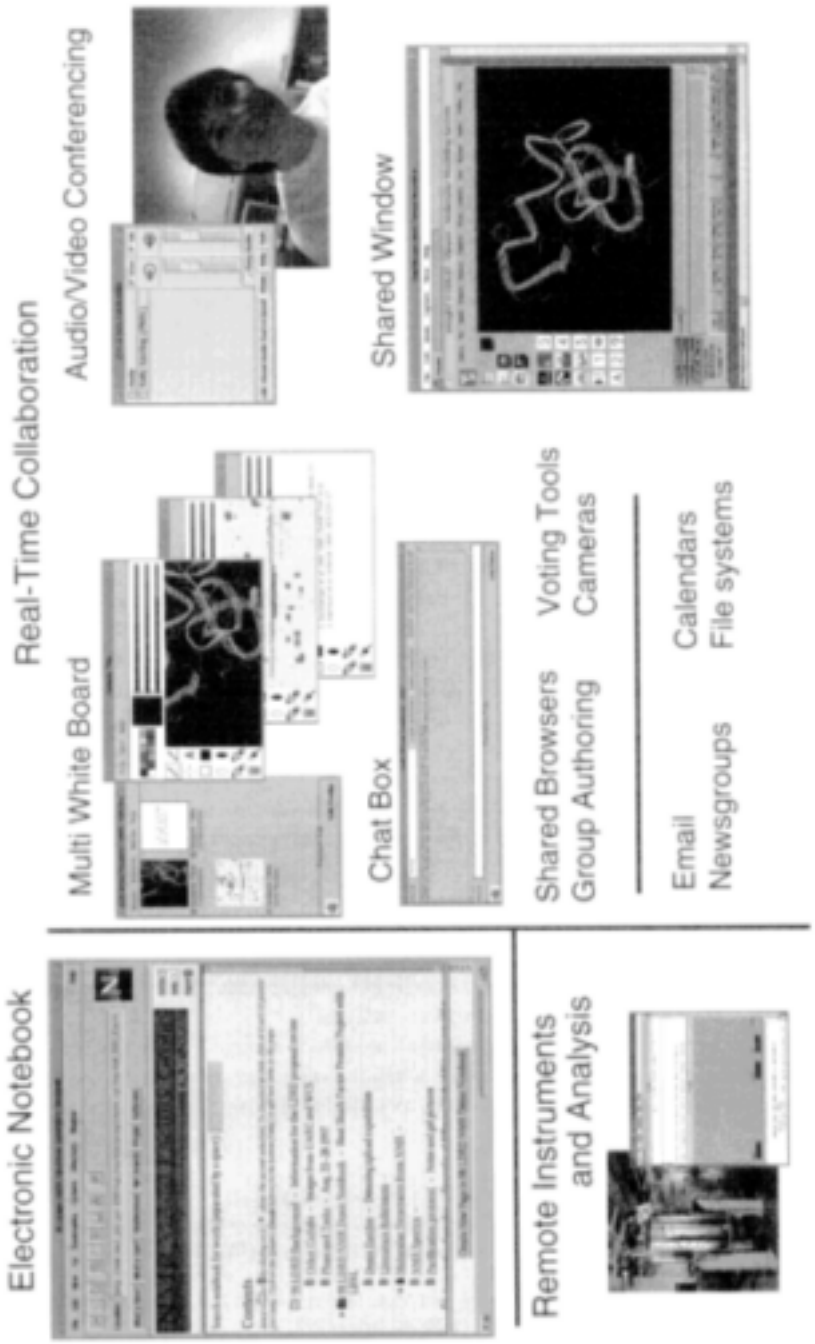


Figure 9.3
Collaborative tools for chemistry research.

and gestures, notebooks, notes, sketches on a white board, physical models, and a myriad of computer applications. People may also come and go during the course of a discussion, according to the need for their expertise or as their availability changes. For people to collaborate electronically, all of these kinds of interactions need to be supported, with natural, fluid changes among them.

Real-time collaboration tools work in a complex arena of computers and the Internet. They inherently involve many users, at many places, having many platforms, with many tools. The scientific computing environment is also very heterogeneous. Although there are "preferred" platforms (Macintosh, PC, or Unix) in some domains of chemistry, it is still seldom that a set of collaborators all have the same type. Hence, real-time collaboration environments need to support multiple platforms. To make matters even more complex, collaboration tools come from many sources, so each tool has its own user interface for connecting to other users. This is a level of complexity that nobody wants to deal with, especially busy scientists. Therefore one would like a common interface that knows how to start up any tool at the click of an icon. Finding your colleagues and discovering active collaborative sessions should also be simplified through user and session directories.

Integrating all of the tools together, knowing who is collaborating, and keeping track of what tools are in use is usually referred to as session management. The session manager used in the EMSL Collaboratory is called CORE2000 (Collaborative Research Environment). It is based on the Habanero framework from the National Center for Supercomputing Applications, NCSA, at the University of Illinois.[22] Habanero is written entirely in Java, a portable object-oriented language. Java Machine Code runs on PCs, Macintoshes, and many versions of Unix. However, there are enough quirks that Java code still needs to be tested on each platform. Habanero is designed to make it easy to add new tools, by providing an event-sharing model for an application to tell a remote copy what it is doing, and vice versa. CORE2000 uses several tools from the basic Habanero tool set, including a chat box, white board, voting tool, and molecule viewer. CORE2000 adds other general capabilities, including screen sharing (via the EMSL TeleViewer[23]) and Internet audio- and video-conferencing (via MBone vic and vat,[24] and CU-SeeMe[25]). These have machine-specific code, because there is currently no way to implement them in portable Java, though that is expected to change.

The EMSL TeleViewer is a general screen-sharing tool with many applications in collaboratories. TeleViewer lets users identify any window on their screen, or define any rectangle on their screen, and share it with anyone else, anywhere. As the contents of that area change, all of the remote copies are updated. Because only the compressed changes are sent, the network bandwidth required is typically much less than for video. With TeleViewer, other scientists can see exactly what is happening remotely, even if they don't have the same kind of computer. The application(s) being run do not need to be collaborative. One can share a spreadsheet, document, instrument console, etc. This is a powerful tool for activities like mentoring, consulting, support, and shared analysis.

In the future, session managers will provide more sophisticated floor control. When just a couple of scientists are working together, it's not too difficult to see when the other person wants to talk or show you something. However, beyond three or four collaborators, additional mechanisms are needed to

[22] See the NCSA Habanero Web site at <http://www.ncsa.uiuc.edu/SDG/Software/Habanero/>.

[23] P. E. Keller and J.D. Myers, "The EMSL TeleViewer: A Collaborative Shared Computer Display," Proceedings of the Fifth Workshop on Enabling Technologies: Infrastructure for Collaboration Enterprises (WET ICE'96), pp. 16-20, IEEE Computer Society, Los Alamitos, Calif., 1996.

[24] See the MBone Web site at <http://www-itg.lbl.gov/mbone/>, and M.R. Macedonia and D.P. Brutzman, "MBone Provides Audio and Video Across the Internet," IEEE Computer, 27(4), 30-36, April 1994.

[25] See the CU-SeeMe Web site at <http://cu-seeme.cornell.edu/>.

mediate discussions and the use of tools, managing how control is passed around for driving the visualization, controlling the instrument, or marking up the document.

There are important software engineering advantages to having a common collaborative tool framework. CORE2000 is a fairly comprehensive set of generic tools for collaboration in science. However, there are specialized tools that need to be constructed to reach "critical mass" in each particular chemistry domain, as described above for the Virtual NMR Facility. Shared, visualization, database access, and instrument control are just a few examples that are usually specific to the particular kind of chemistry one is doing. Common frameworks enable many development groups to contribute to a powerful collaborative toolkit. This approach is necessary to move collaboratories from a cottage industry to broad application. Frameworks available from academic sources include Habanero and Tango, a project from the Northeast Parallel Architectures Center (NPAC) at Syracuse University.[26]

Below the level of collaboration managers, there is the need for other "middleware" to support the distributed applications for collaboratories. One successful framework of this type is the Product Realization Environment from SNL.[27] PRE is a lightweight, Common Object Request Broker Application (CORBA)-based, horizontal integration framework. CORBA is an industry software standard component technology for hardware and language-independent distributed applications. PRE defines how CORBA can be used to connect distributed design tools, databases, files, directory services, and user interfaces. Many collaborative tool developers are moving to CORBA to address interoperability requirements from a standards basis.

IMPACTS OF COLLABORATORIES

At this point, it's reasonable to ask how effective collaboratories are in getting chemistry work done. One study performed in our laboratory followed many groups and looked in detail at how two groups used the tools and what impacts the collaboratories had on their work. [28] One group involved intelligence analysts, and the other NMR spectroscopists. The NMR project is a typical peer-to-peer collaboration aimed at determining the detailed three-dimensional structure of a segment of a heat shock factor protein. The protein was expressed at LBNL and shipped to PNNL. Researchers at LBNL and PNNL then collaborated on experiments using EMSL's high-field NMRs and shared analysis of the three-dimensional protein structure. For each project, work activities were identified and followed, including experiment planning, experiment setup and monitoring, analysis, and reporting. Feedback was obtained through observation, interviews, discussions, and comments.

Across the studied groups, four broad modes of collaboration were observed:

- *Peer-to-peer*, where researchers with a common background and vocabulary work closely together;
- *Mentor-student*, where knowledge and experience are unequal, e.g., one scientist is helping another scientist or a student to understand a new topic—lecture modes are common;
- *Interdisciplinary*, where researchers may share high-level concepts but not a common background, and therefore must translate results into terms each can understand; and
- *Producer-consumer*, where the producer provides information to address a need of the consumer, usually without much common knowledge.

[26] See the Tango Web site at <http://trurl.npac.syr.edu/tango/>.
[27] See the Product Realization Environment (PRE) Web site at <http://daytona.ca.sandia.gov/pre/>.
[28] Anne Schur, Kelly A. Keating, Deborah A. Payne, Tom Valdez, Kenneth R. Yates, and James D. Myers, "Collaborative Suites for Experiment-Oriented Scientific Research," ACM Interactions, 3, 4047, May/June 1998.

The character of the work varied naturally during the collaborative sessions. As collaborative work progressed, scientists changed their mode of interaction to suit the task at hand, often several times during the same collaboration session. Thus, it is important that collaboratory tools support fluid transitions between modes of collaboration, as well as the many types of media used in science. Some feared that local researchers helping to operate the instruments would be relegated to technicians. However, this did not happen. There were ample opportunities for each of the NMR spectroscopists to contribute to the science problem.

Collaboratory tools are designed to manage information to facilitate collaboration. As scientists used the collaboratory over a period of time, they noted a highly desirable shift in the distribution of their effort from data management to analysis. They also benefited from the impromptu and informal forms of interaction that the collaboratory supports. Screen real estate is a valuable commodity. As a team worked together for a while, the center of attention shifted from looking at each other (video conferencing) to concentrating on the data. Often the video was eliminated in favor of devoting more screen space to the data under discussion. Overall, the collaboratory supported non-linear work processes, which increased the productivity of the collaborations.

These and other experiences provide a glimpse of the expected impacts of collaboratories on chemical science and technology. Unlike many technological advances, collaboratories affect both the techniques and the processes of science. This makes their impacts difficult to gauge, and easy to underestimate. At a minimum, collaboratories will change where components of research and analysis are done, and how experts are brought into a project. The ability to better share facilities across a company or a scientific community will change the equations governing their feasibility and viability. The collaboratory's ability to marshal facilities, information, and expertise across disciplines and nations will affect how quickly complex problems can be solved, and thereby what important problems are addressed. The roles of a scientist as researcher, mentor, and educator tend to blur in collaboratories. This creates new and exciting opportunities, and some problems. It would certainly be beneficial to expose more students to "real" science and science students to better mentors. However, time is one of the scientist's most precious resources; the nation's expert on NMR pulse sequences cannot field everyone's questions. However, techniques to help understand the new dynamics and strike the balance are in their infancy. Today, the applications of collaboratories are still within "sight" of current practice in research. The initial focus has been on implementing current work processes, at a distance, and making the logical extensions to them. However, as these techniques become more familiar, one would fully expect to move significantly beyond current practice, to new paradigms for scientific work. This will be very exciting.

THE PROMISE OF COLLABORATORIES

Because of the work being done in universities and government labs, the chemical sciences are one of the first domains benefiting from collaboratories. However, there is a great deal to be done to achieve the promise of collaboratories throughout chemistry and the broader scientific community. Our experience is limited, and there are many technical hurdles to overcome. To succeed, chemical science collaboratories need to be developed by multidisciplinary partnerships of chemists and computer scientists.[29] Collaboratories are not off-the-shelf products; within chemistry, each domain has particular

[29] "National High Field Magnetic Resonance Collaboratorium," a report to the Committee for High Field NMR: A New Millennium Resource, published by the National High Field Magnet Laboratory, Tallahassee, Florida, August 1998.

kinds of data and ways of manipulating and analyzing it, creating the need for specific collaboratory capabilities. To meet the needs of these scientists, collaboratory frameworks must be flexible and extensible. In addition to making tools, we must learn to deploy and support collaboratories in the field, and evaluate collaboratory science in action. This requires "research by doing," the research and development projects that create and support pilot collaboratories in the chemical sciences. The nation's major scientific user facilities are a fertile ground for these initial projects, with the potential for performing new science while improving facility access and efficiency for many scientists.

Advances in computers and networking, coupled with new developments like the World Wide Web and Java, have fueled collaboratory R&D; however, continued progress is needed to support the widespread development and use of collaboratories in the scientific community. There is much to be learned about the representation and use of shared knowledge. Standards and infrastructure for security and authentication are also important for distributed applications like collaboratories. The frameworks that form the foundation for collaboratories to share data, events, and programs are much more complex than normal Internet tools or client-server applications. The architecture and industry/community standards for these frameworks are still an open research issue. As in the past, the nation's research community has much to contribute to the development of the next generation of Internet standards. The needs of science and business communities differ, and so it will be important for the scientific community to be heard against the background of new Internet standards driven by commerce. Networking developments are also primary enabling factors for collaboratories. To scale up the deployment and use of collaboratories, we will need higher-performance networks, more scalable and capable network standards, and better network management capabilities.

CONCLUSION

Overall, collaboratories are an emerging capability that will remove many barriers of distance and time in the sciences. The present confluence of developments in computing, databases, and networking creates a unique opportunity to develop and deploy collaboratories. The impact promises to be great, not only on what science we do, but also for how we accomplish scientific endeavors. There are several important drivers for the development of collaboratories in the chemical sciences. The opportunity to make more progress on a project, the opportunity to employ expertise, data, experiments, or computations that would not otherwise be available, and the opportunity to be first to explore a research question or solve a problem, all represent competitive advantages of collaboratory use. Collaboratories can also affect the complexity and scale of chemical problems that can be considered. Although collaboratories can have value in any size collaboration, collaboratories may be crucial for projects that need large or multidisciplinary research teams. Collaboratories will expand opportunities for timely information exchange between basic and applied R&D efforts. Of course, collaboratories also provide opportunities to manage costs, by optimizing travel, equipment use, and information value.

ACKNOWLEDGMENTS

Many people contributed information and concepts to this paper; however, special thanks are extended to my colleagues lames Myers, Elena Mendoza, Deborah Payne, Kelly Keating, Nestor Zaluzec, Larry Rahn, and Anne Schur. Much of the work described is supported by the DOE2000 program of the Mathematical, Information, and Computational Sciences Division of the Office of Science in the U.S. Department of Energy. A portion of the research was performed at the W.R. Wiley Environmental Molecular Sciences Laboratory, a national scientific user facility sponsored by the

Department of Energy's Office of Biological and Environmental Research and located at Pacific Northwest National Laboratory.

DISCUSSION

Randy Collard, Dow Chemical Company: For industry to use collaboration and extranets effectively, it is critical for us to have security in place and to have flexible security as well as encryption. You spoke about that a little bit with respect to the diesel project. Could you talk a little more about the state of that as you see it and what you see as necessary?

Raymond Bair: The state of security is not as good as I would like it to be. I think some of the capabilities that are coming out of what people classify as the next generation of Internet protocols and capabilities will make this a lot easier.

A number of places have found reasonable success in point-to-point security by using extant tools like SecureShell and virtual private networks, but the virtual private networks are not trivial to set up and administer, and so it is not a technology that I would advocate, except perhaps for cases where the secure interactions are fairly static, as between one institution and another.

William Winter, SUNY-ESF, Syracuse: I have two questions. Many of the applications and instruments that you are using have vendor-controlled software. The first question is, How do you deal with the issue of floating licenses across a network as opposed to just local floating?

Raymond Bair: It depends ultimately on what that license says, and so one cannot predict beforehand. In an architecture with the server independent of the instrument, usually you are interfacing with proprietary instrument software through a meta language that is available for the instrument through serial ports. In that case you are not actually running that instrument's software. But in terms of sharing the screen of the instrument elsewhere, for example, by using a remote X Windows display on an instrument, I am not personally familiar with whether there is a legal intellectual property issue with that. It is a common enough practice, but the terms would, once again, depend on the particular license.

William Winter: My other question is a bit more philosophical. Yesterday the comment was made that although industry is very much involved with multidisciplinary and team approaches, realistically the Ph.D. is going to remain an individual effort. Do you really think that has to be true? Can we talk about having multidisciplinary, integrated, team Ph.D.s where it still is compartmentalized enough that somebody can take credit—"I did this"—as an individual?

Raymond Bair: I think so. I see multidisciplinary teams forming around a number of the environmental research areas where experiments in any one domain aren't going to be sufficient to address the issue at hand and where collections of Ph.D.s have become engaged in addressing a larger and more complex issue by working together in their respective disciplines and sharing information. Given the kinds of problems we are trying to solve, there is definitely encouragement from the funding agencies in the kinds of proposals being solicited in a number of these areas that almost virtually requires working together, and so we will see examples of this happening.

10

The World Wide Laboratory: Remote and Automated Access to Imaging Instrumentation

Bridget Carragher
and
Clinton S. Potter
University of Illinois at Urbana-Champaign

INTRODUCTION

For the past several years we have been involved in the development of remote and automated access to imaging instrumentation for an overall project known as the World Wide Laboratory (WWL).[1] There are a number of advantages to providing remote access to imaging instrumentation. First, it provides access to unique and/or expensive instruments without requiring the user to be physically present at the site of the instrument. In addition it provides the opportunity for collaboration and/or consultation with researchers anywhere in the world, thus providing for a network of distributed expertise. Finally, this technology presents unprecedented opportunities for education and training. These opportunities might otherwise be limited only to those institutions with the means to support expensive and unique instruments.

We propose that there are at least six ways in which remote-access technology can be used in practice:

- *Service*. In the service mode, a local operator can consult with a remotely located principal researcher providing a specimen and who would provide input about the quality of the images and the parameters to be used to acquire data. This mode is extremely important for extending service capabilities at the National Research Resources.
- *Collaboration*. In the collaboration mode, the principal researcher using the instrument can consult with other experts from around the world.
- *Education and training*. Imaging instruments can be made accessible in the classroom for K-12 education and undergraduate and graduate training.
- *Remote research*. The instrument can be used by a remote researcher with minimal local operator intervention.
- *Automated control*. The instrument can be used by a remote researcher, and functions normally performed manually by a local operator are performed automatically by a computer system.

[1] See <http://wwl.itg.uiuc.edu>.

- *Intelligent control*. An intelligent system can perform the same functions as an operator and can learn from the researcher's experience.

In this paper we discuss our experience with the WWL project and some specific examples that we believe demonstrate the ideas behind a successful collaboratory.[2] We address some of the issues involved in providing remote and automated access to instrumentation and its advantages to various categories of users.

WWL: CURRENT IMPLEMENTATIONS FOR SERVICE, COLLABORATION, AND EDUCATION

Instrumentation currently supported by the World Wide Laboratory includes a transmission electron microscope (Philips CM200), nuclear magnetic resonance imaging spectrometers (Surrey Medical Imaging Systems, Varian, TechMag), and a video light microscope. All of these instruments are accessible through Web browser-based user interfaces.

Remote Access to TEM

JavaScope is a Web-based application designed to operate a Philips CM200 transmission electron microscope (TEM) and to view digital images remotely. JavaScope has been written as a client/server application (see Figure 10.1). The JavaScope applet is the client and presents the application interface to the user. JavaScope responds to actions by the user by sending commands to a camera control server and microscope control server that run on a UNIX workstation attached to the TEM and CCD camera. These servers are responsible for controlling the TEM and digital camera using applications and libraries already developed as part of the emScope library.[3] The user interface is shown in Figure 10.2.

As a readily accessible tool for remote consultation and exploratory grid browsing, the basic Javascope implementation has been successful. JavaScope has been used by our collaborators in California (Research Institute at Scripps Clinic) to control the TEM in Illinois and provide advice as to the worth of acquiring data from a particular specimen. It has also benefited students at the microscope by providing them with a means to consult with a remotely located advisor.

Remote Access to MRI

The second example in the World Wide Laboratory is remote control of a nuclear magnetic resonance (NMR) imaging spectrometer by means of a Web browser. This system evolved from our work in developing a distributed control, acquisition, and processing interface for an NMR imaging spectrometer (4T, 31-cm bore) with an acquisition console from Surrey Medical Imaging System.[4] This system,

[2] M. Hamalainen, S. Hashim, C. Holsapple, Y. Suh, and A. Whinston, "Structured Discourse for Scientific Collaboration: A Framework for Scientific Collaboration Based on Structured Discourse Analysis," *Journal of Organizational Computing*, 2(1), 1-26, 1992.

[3] N. Kisseberth, M. Whittaker, D. Weber, C.S. Potter, and B. Carragher, "emScope: A Toolkit for Control and Automation of a Remote Electron Microscope," *J. Struct. Biol*., 120, 309-319, 1997.

called NSCOPE, has been interfaced to a UNIX workstation that provides a real-time processing and control capability. The significant features of this system are real-time control of all acquisition and control aspects of the magnetic resonance imaging (MRI) system, real-time processing of dynamic MRI data, and distributed processing modules for high-performance computing systems.

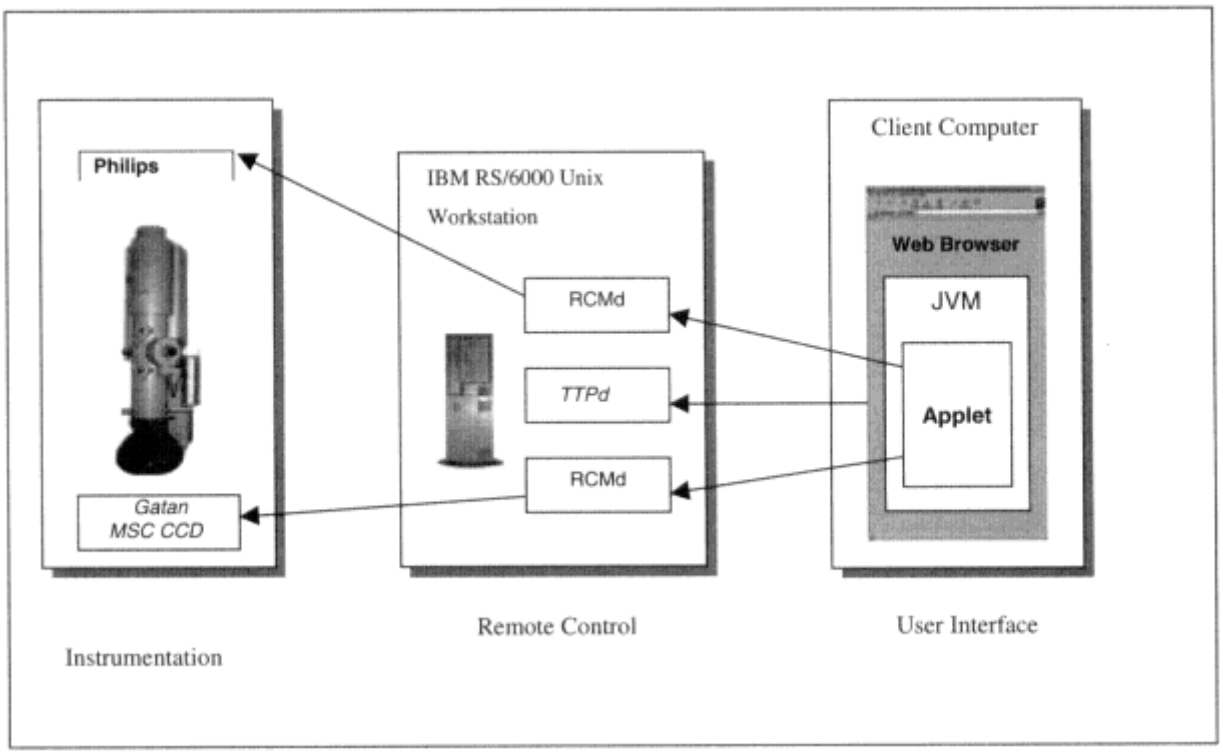


Figure 10.1
Architecture underlying the JavaScope user interface.

The NSCOPE was initially developed to support functional magnetic resonance imaging and dynamic imaging applications. The modular software architecture of the system has allowed us to adapt the system to other user interfaces. For example, the Experimentalist's Virtual Acquisition Console (EVAC) used the NSCOPE system for real-time control and visualization of the MRI system from within an immersive visualization environment.[5] This system used voice commands to control the MRI system from within a room-sized virtual environment called a CAVE. A real-time stereo capability was also developed for EVAC that used the flexible processing capabilities of the NSCOPE.[6]

A Web-based control interface was added to the NSCOPE in early 1996 (Figure 10.3). This allows the MRI system to be controlled from anywhere with Internet access using a standard Web browser. Significant features include real-time acquisition and processing of images from the MRI system;

[4] C.S Potter, Z-P. Liang, C.D. Gregory, H.D. Morris, and P.C. Lauterbur, "Toward a Neuroscope: A Real-time Imaging System for the Evaluation of Brain Function," *Proceedings of First IEEE International Conference on Image Processing*, November 13-16, 1994, Austin, TX, Vol. III, pp. 25-29.

[5] C.S. Potter, R. Brady, P. Moran, C. Gregory, B. Carragher, N. Kisseberth, J. Lyding, and J. Lindquist, "EVAC: A Virtual Environment for Control of Remote Imaging Instrumentation," *Computer Graphics and Applications*, 16(4), 62-66, July 1996.

[6] C.D. Gregory, C.S. Potter, and P.C. Lauterbur, "Interactive Stereoscopic Magnetic Resonance Imaging," U.S. Patent Number 5708359, 1998.

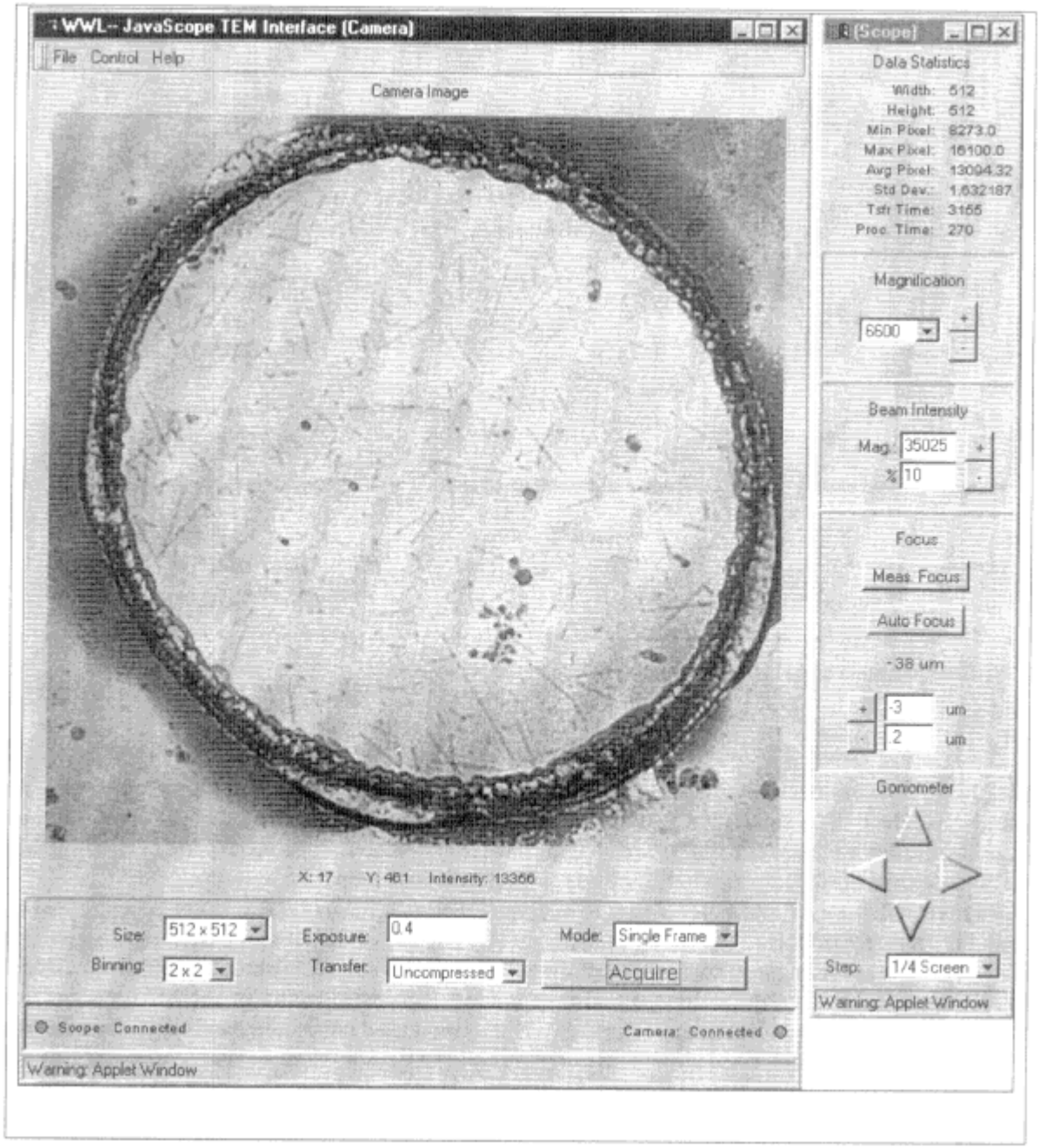


Figure 10.2
Java applets responsible for camera control (left) and microscope control (right).

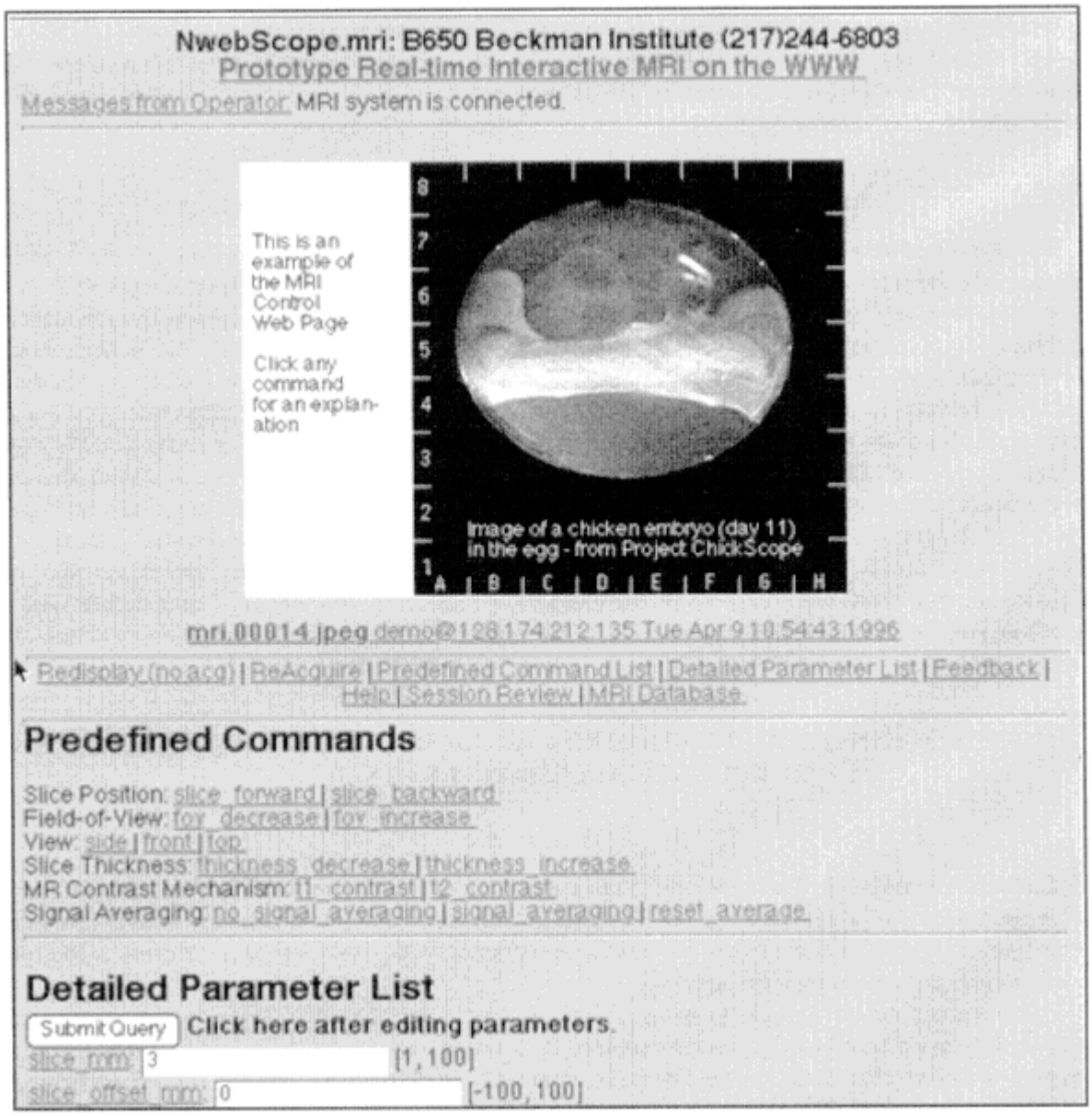


Figure 10.3
Web-based control system provides interactive access to all imaging parameters on the MRI system.

password-protected login system to limit access to authorized users; and scheduling mechanism to limit permission for use of acquisition capabilities to specific users at selected times.

The current system provides complete control of all instrument acquisition parameters from the Web browser. The Web browser interface allows users from various domains and levels of expertise to run the MRI system without the need for extensive system-specific training.

Chickscope: A K-12 Education Project Using Remote MRI

The remote MRI system was used in the spring of 1996 in a project called Chickscope,[7,8] which demonstrated the feasibility of remotely controlling an MRI device through the World Wide Web. The purpose of the project was to enable students and teachers from 10 classrooms ranging from kindergarten through high school to control an MRI system in order to study the maturation of a chicken embryo during its 21-day cycle of development. From classroom computers with access to the Internet, students used Web browsers to control the MRI system and manipulate experimental conditions through a simple online form. Students could generate their own data and then view the resulting images of the chick embryo in real time. The objectives of the Chickscope project were twofold. First, we sought to make extraordinary hardware, software, and human resources available to the classrooms and study the impact of such a system on K-12 education. Second, we set out to "stress-test" interactive, remote control of the MRI system for further development by scientific researchers. Overall the Chickscope project was very successful in that it was well received by the teachers and students, and there has been a great deal of interest and enthusiasm for repeating the project. In addition, it also allowed us to demonstrate that very complex technology could be used effectively by students at all grade levels.

REMOTE INSTRUMENTATION FOR SERVICE, COLLABORATION, AND EDUCATION: LESSONS LEARNED

User Interface

The basic requirements for the World Wide Laboratory in the service, collaboration, and education modes are relatively straightforward. On the user end we need a network connection and a standard Web browser. On the instrument end we need a network connection and interface software to interpret the commands coming in over the network.

There are several advantages to using a Web browser interface. First, because almost everyone knows how to use a Web browser, there is no need for training on a specific user interface. Second, Web browsers are now ubiquitous on all computer systems, and third, there are no special software or hardware requirements. As a result, we can be reasonably sure that a remote user anywhere in the world with access to the Internet will have the tools to run the instruments remotely.

There are other remote user interfaces that use techniques such as remote screen copy (e.g., Timbuktu) or low-level distributed windowing libraries such as X-windows. These systems require specialized software to be installed and maintained on the remote user's computer system.

[7] B.C. Bruce, B.O. Carragher, B.M. Damon, M.J. Dawson, J.A. Eurell, C.D. Gregory, P.C. Lauterbur, M.M. Marjanovic, B. Mason-Fossum, H.D. Morris, C.S. Potter, and U. Thakkar, "Chickscope: an Interactive MRI Classroom Curriculum Innovation for K-12," *Computers and Education Journal Special Issue on Multimedia*, 29(2), 73-87, 1997.

[8] See <http://chickscope.beckman.uiuc.edu>.

Modes of Collaboration

Our experience with the World Wide Laboratory indicates that at least three instrument access modes need to be supported:

- *Single user*. Allows dedicated access to an imaging system by a single user.
- *Multiple non-cooperating users*. Allows several users to access the system simultaneously. The users are not aware of each other. Commands from the users are queued, and the data are returned to the requesting user. This mode is useful in education projects like Chickscope in which several classrooms may be accessing the instrumentation simultaneously.
- *Multiple cooperating users*. Allows several users to use an instrument collaboratively by using mechanisms for passing instrument control among the users.

WWL: Current Implementations for Research

The above examples demonstrate the use of the World Wide Laboratory architecture for service, collaboration, and education. Using this architecture for remote scientific research in which reliance on a local instrument operator is minimal or non-existent poses additional requirements that are discussed below.

Automated/Intelligent Control for Scientific Research

Our group has been extensively involved in a project involving scientific research on remote and automated instrumentation. The goal of the project is to acquire very large numbers of good-quality images from a TEM completely unattended by a human operator. The motivation for developing this automated system arises from the field of electron crystallography, in which a TEM is used to study the structure of proteins at moderate to high resolution (5 to 30 angstroms). The technique most commonly used to preserve the proteins in the TEM is known as CryoEM, in which the protein is preserved in a very thin layer of vitreous ice.[9] The ice is usually suspended over a carbon grid, and the goal of the microscopist is to identify the holes in the grid where the ice is potentially of the right thickness and acquire a high-magnification image of this area (Figure 10.4).

A number of practical problems with the CryoEM procedure tend to make it extremely time consuming and tedious for the operator. The first is that producing ice of precisely the right thickness is not straightforward; as a result the ice is quite often either too thick or too thin, and a lot of searching around the grid is required to find suitable areas. Second, because the electron beam is extremely damaging to the specimen and will destroy it after a very short exposure, the grid can be examined only at very low magnification if a reasonable length of time is desired. The high-magnification image is never examined prior to shooting the micrograph, and this leads, not too surprisingly, to a rather high rejection rate; many of the acquired images are simply thrown away, and only a few turn out to be suitable for further analysis. Finally, because the beam damages the specimen, the micrographs must be acquired with a very small dose of electrons, and the images are very "noisy" as a result. Thus, to determine the protein structure to high resolution properly, the signal-to-noise ratio of the structure must be increased, and this

[9] J. Dubocher, M. Adrian, J.-J. Chang, J.-C. Homo, J. Lepault, A. McDowall, and P. Schultz, "Cryo-Electron microscopy of vitrified specimens," *Quarterly Review of Biophysics*, 21(2) 129-228, 1988.

requires averaging together many images. The end result is that this technique by its nature requires the acquisition of large numbers of micrographs, perhaps thousands, or tens of thousands, to achieve high resolution. Manual methods are clearly impractical, and it was with this in mind that we embarked upon the project of completely automating the acquisition of large numbers cryo-electron micrographs.

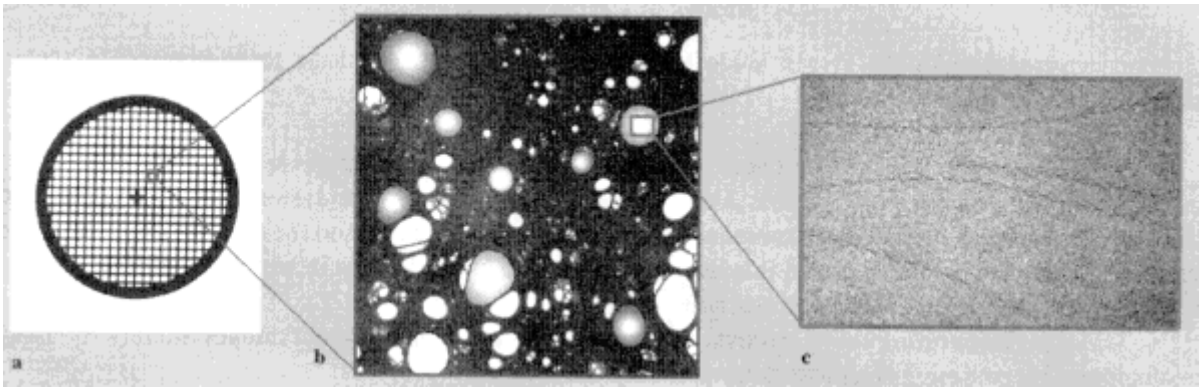


Figure 10.4
Acquisition of cryo-electron micrographs. A copper grid (a) is covered with a carbon-coated perforated plastic mesh (b). A droplet of buffer containing the protein of interest is applied to the grid, blotted to a thin film, and then rapidly plunged into a liquid cryogen. The protein of interest © is reserved in its native form in the vitreous ice.

As a prototype for automated acquisition of TEM images, we have developed a system, called Leginon,[10] to automatically acquire large numbers of acceptable quality images from specimens of negatively stained catalase, a biological protein that forms crystals. Acquiring good-quality images of this specimen is often used as a test for students taking a course in electron microscopy and thus provides an excellent driver for the research methods that must be developed to solve the general problems of automated image acquisition. Furthermore, as catalase is an ordered crystalline structure, assessment of this order provides us with an objective measure of the quality of the automatically acquired images (Figure 10.5). Each low-magnification image (Figure 10.5a) is processed to identify large contiguous areas of density by a template matching method. Image feature metrics (size, mean, variance, centroid) are calculated and stored for each of the identified contiguous regions. These image features are later used in deciding whether a high-magnification image (Figure 10.5b) of the region will be acquired; for example, regions that are too small are rejected. The image quality of each high-magnification image is automatically assessed by calculating the power spectrum (Figure 10.5c), identifying diffraction spots (Figure 10.5d), and measuring the signal-to-noise ratio of each diffraction spot.

Currently, the automated system can acquire approximately 1,000 images in a 24-hour period. In one experiment, we have compared the performance of the automated system to that of a human operator. A total of 288 high-magnification images were acquired manually, and 79 percent of these were acceptable as defined by an analysis of the order of the crystal. In comparison, using the same

[10] C.S. Potter, H. Chu, B. Frey, C. Green, N. Kisseberth, T.J. Madden, K.L. Miller, K. Nahrstedt, J. Pulokas, A. Reilein, D. Tcheng, D. Weber, and B. Carragher, "Leginon: A System for Fully Automated Acquisition of 1000 Micrographs a Day," manuscript submitted to *Ultramicroscopy* , October 1998; see <http://www.itg.uiuc.edu/tech_reports/98-008/>.

specimen, the fully automated image acquisition system was used to acquire 380 images, of which 51 percent were acceptable.

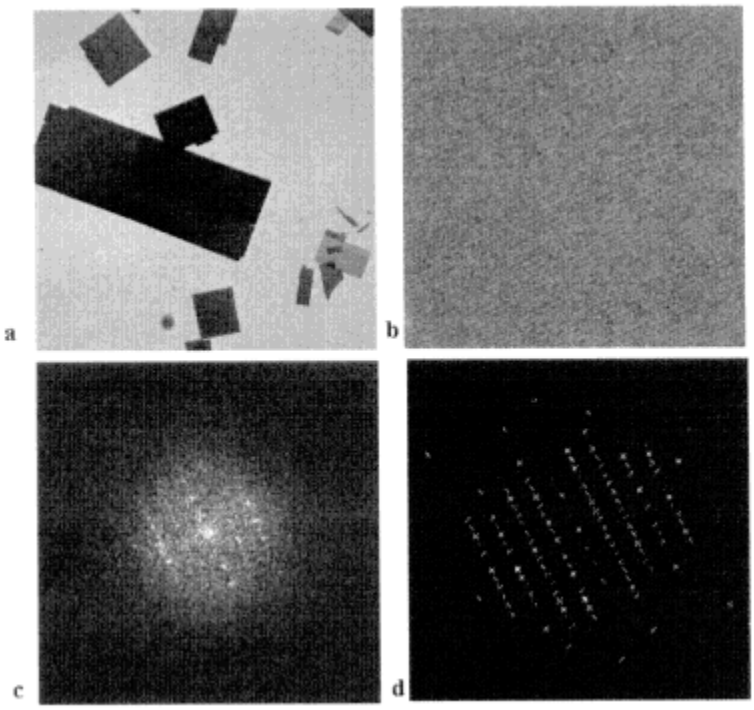


Figure 10.5
Automated acquisition of electron micrographs of negatively stained catalase crystals.

The system can be further improved by adding intelligence to the feature selection criteria. For example, analysis of the results indicated a correlation between average feature intensity and image quality. This feature intensity is related to the thickness of the catalase crystal and indicates that thinner specimens result in more acceptable images. The fully automated target selection criteria were further refined by incorporating an assessment of specimen thickness into the model. By acquiring high-magnification images of only those features that have an average intensity greater than a preset threshold the percentage of acceptable images can be significantly improved. For example, if the threshold, is set to 6,000, the percentage of acceptable images improves to 86 percent from a baseline of 51 percent. Thus, the automated system does as well as or slightly better than a human operator.

REMOTE INSTRUMENTATION FOR SCIENTIFIC RESEARCH: LESSONS LEARNED

Our experience with the development of the TEM project has shown the necessity for incorporating automation and intelligent algorithms into the data-acquisition system. To develop such a system effectively requires a distributed hardware and software environment. The basic architecture of the Leginon system is illustrated in Figure 10.6. The system has these components:

- *Instrument interface*. To develop an instrument interface requires information from the manufacturer, and distributed control is needed for accessing the instrument over a network. This process is

complicated by the lack of open systems and industry standards. Ideally the instrument should require minimal human interaction during an automated experiment. For example, the time between refilling cryogens on the TEM should be extended to support overnight runs.

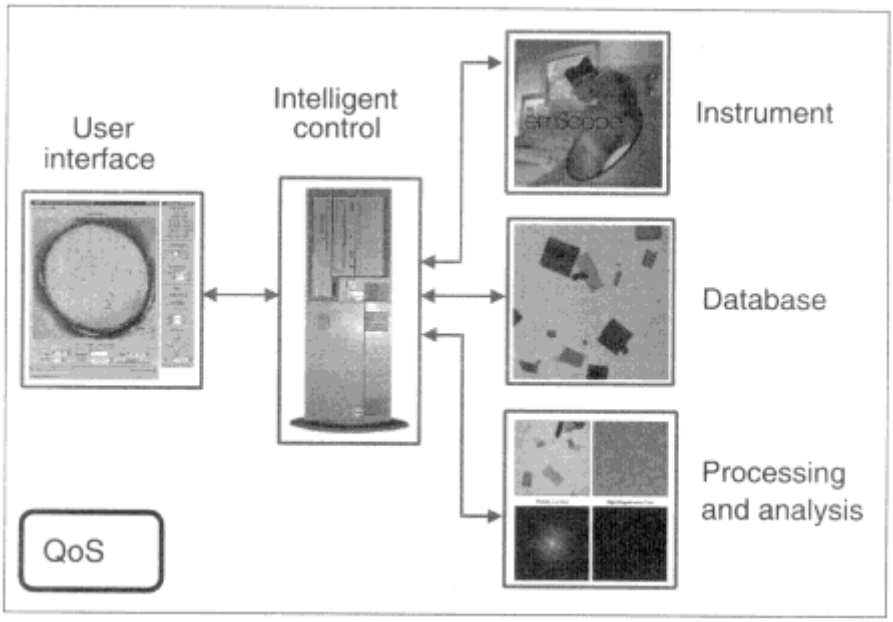


Figure 10.6
Major components of the Leginon system.

- *Database*. It has become clear in the course of the Leginon project that there is a critical need for a database to support the thousands of images that are acquired and the acquisition parameter data that is associated with each image. Incorporation of a database would provide improved data management as well as the ability to track acquisition, control, processing, and modeling parameters.
- *Processing and analysis*. Developing intelligent image-acquisition systems requires that the instrument is closely integrated with processing and analysis software packages. There is a need for integration with commercial and community software packages, and these need to support the interfaces for distributed access.
- *Control*. A distributed control program must effectively synchronize all of the components of the distributed system and needs to be adaptable to each experiment. Ideally, the control program should be portable between systems and extensible by the end user.
- *User interface*. The user interface must be flexible and suit the needs of the user. The system must be flexible enough to support new technologies such as next-generation Web interfaces and virtual reality.

Additional features that would be desirable for use in the World Wide Laboratory include audio and video conferencing and real-time updates of the system status. These extra capabilities enhance the remote researcher's understanding of the current status of the instrument. We also believe that scheduling and security considerations are not only desirable but also essential for turning this technology into a practical reality.

CONCLUSIONS

We have demonstrated that the WWL architecture can be used for service, collaboration, education, and scientific research. The remote instrumentation supported by the WWL is one component of a collaboratory. Although the Chickscope project was not originally intended to do so, we believe in retrospect that this project demonstrated all of the components defined for a working collaboratory. Chickscope provided access to remote instrumentation from the classroom and gave students access to distributed expertise. All of the participants actively contributed to and used an image database that is now used by others, and the project served to develop a community composed of students and researchers from a number of different disciplines.

There is an increased interest in developing the technology to support remote instrumentation. To improve the acceptance of collaboratories in the general scientific community, we need to demonstrate the impact of this technology in the scientific research environment and systematically evaluate collaboratories' contribution to productivity.

ACKNOWLEDGMENTS

The World Wide Laboratory project is a collaboration with several groups at the University of Illinois at Urbana-Champaign, including members of the Beckman Institute for Advanced Science and Technology, the Biomedical Magnetic Resonance Laboratory, and the National Center for Supercomputing Applications. We are grateful for the enthusiastic participation of all the many individuals who have contributed to these projects. Financial support and equipment were provided by the National Science Foundation (Grant No. 9730056, 9871103), IBM shared university research program, Informix Inc., and the Lumpkin Foundation.

DISCUSSION

Peter Taylor, San Diego Super Computer Center: You talked at the end about this issue of cultural adjustment to collaboratories. High-energy physics and both radio and optical astronomy are areas where people have been running consortia and working collaboratively—and I am distinguishing that from a collaboratory—for a number of years. Do we turn to groups like that to find out how to make that cultural adjustment, or do you think the problems there are sufficiently different from chemistry or biology that there is not much to learn?

Bridget Carragher: In truth I don't know, but I think we could look at examples like that to get some experience of what it is like. But more important than that is to get people involved in using these systems. If people are collecting their data and just doing so much better than they were before, that is going to be picked up by 10 other groups immediately, and there will be a groundswell of support. You know—it is a word-of-mouth thing. Just to talk about collaborations isn't it a great thing. I think that in my field, this does no good at all. The only thing to do is to demonstrate that it works. I could say, "It works great in high-energy physics," and people will just shrug. It means nothing to them unless you demonstrate it in their own particular paradigm.

William Winter, SUNY-ESF, Syracuse: I would like to point out to the polymer and materials people in the audience that electron crystallography is a very powerful technique in that area, as well as protein crystallography, and it has been demonstrated by Douglas Dorsett in this country and Henri Chanzy in

France, among others. The question would be, Is it really the best use of very expensive instrumentation to have it dedicated online 24 hours a day for seven days a week for use in elementary school education? I don't know. I am asking you.

Clint Potter: Absolutely not, and I guess one of the reasons we do projects like Chickscope or Bugscope is that it is a technology driver. Putting together the Chickscope project, having second graders bulletproof your system, is a great test of how well your system actually works.

I think it also gets back to this idea of cultural training. The second graders have no qualms about collaboratories even if they don't know what one is. They are doing it, and they are not set in their ways of doing research. Maybe it is that generation that is going to be doing the scientific collaborations 20 years from now, or maybe it will come quicker. We do not have our instruments available 24 hours a day. It is for a specific project.

Bridget Carragher: In Bugscope we will have a period of time once a week, maybe, for a few hours for access, which is a good thing. I think outreach and training are very good things, but of course we wouldn't want to do that all the time. One of the lessons learned from the Chickscope project was that the second graders thought this was entirely natural: Why shouldn't they have an MRI? Why shouldn't they be speaking to researchers from all over the world and accessing thousands of images? There was no surprise for them at all—no "Gee, wow."

Clint Potter: They really didn't feel that this was unusual. That was one of the big things that came out of this project. They were talking to the researchers and the scientists, and it was not just a one-way thing. They did not just suck our energy out of the project. It involved having users at the other end who were really dependent on our system being up for their one shot at it during that day, and having questions coming back. We put together the Chickscope project in about a month and a half from scratch. It was exciting, and there was a lot of energy built out of it because we had these real customers out there.

Sue Fratkin, Southeastern Universities Research Association: Let me also inform all of you, having worked with Chickscope on the Hill, that there is a third element to that project, and that is that the members of Congress could see, touch, taste, and feel that kind of an experiment and so could relate to it in terms of funding. They could see that the young kids were relating to it, were doing very well. The congressmen and congresswomen themselves were playing with it because we were online right then and there, and their reaction was worth all the money because that is where your funding is going to come from. And if they can understand it and relate to it, then you have a much better chance of getting your point across.

Clint Potter: We have never had any funding to do collaboratories. We do it because we think it is actually needed, and even an educational project, we thought, was a good way to prove to our own communities that if second graders could run an NMR spectrometer, then maybe the principal investigator could do it as well.

Allen Bard, University of Texas: In getting back to the issue of the collaboratory in the culture of chemistry, I think the model might be centers for instrumentation. Historically there were NMR centers and computer centers, and I think by and large they weren't terrifically successful. A little aspect of the chemistry culture is that if chemists can get it in their own backyard they are going to do it, but a chemist

is not going to go to a center or send samples to centers, and that is really ingrained in the culture. So, I think where this and the model of high-energy physics and radio astronomy and so on will be most successful is where there are instruments that no one is going to have in his own backyard. I am not so sure about electron microscopes. Maybe very high level electron microscopes might work, but I think that is where you will get a collaboratory. If I can do the experiment in my own backyard or find somebody to finance it, I will probably do that.

Bridget Carragher: Absolutely. And I would always choose to do that myself. I run a big facility with 12 microscopes in it, but if people have it in their own network they should stay there. That is where all their stuff is. That is where their colleagues, their students, and their notebooks are. But that is really also the idea of running these experiments remotely and automatically. You are just sitting at your desk with all your stuff around you, and you are not going to a center. And certainly really decent electron microscopes these days cost $1.5 million to $2 million each. You are not going to have more than a dozen or so of those in the country. You could have thousands of people doing this. The other idea with running these instruments continuously is that if one costs $2 million, you want it up 7 days a week, 24 hours a day. You do not want that machine sitting idle.

11

The Wired Laboratory

David R. McLaughlin
Eastman Kodak Company

This presentation describes how Kodak has used computers and information technology to enhance operations in its research laboratories. This has been an effort to create an electronic or computerized laboratory, and to deliver information at the scientist's fingertips. Some perspective is given on what is meant by the "wired laboratory" and why a commercial enterprise would be interested in having one. It includes areas of impact, examples from our operations, and some speculation about the future.

This discussion is presented from the perspective of an analytical division, within a materials research organization, supporting a commercial business. The business environment requires that a profit be made. This is done by selling more products, by making them at decreasing cost, and by generating new products that sell—more efficiently than the competition. The materials research organization supports the business by developing new materials that can be used to produce new or better products more efficiently. The analytical division contributes to this efficiency by providing measurements and information that are key to understanding and controlling material properties and manufacturing processes. The examples used in this discussion originate from a spectroscopic chemical structure characterization laboratory. The concepts, however, apply equally well for other analytical and chemical information.

The drive to be efficient in all aspects of business is intense. The wired laboratory allows research to be more efficient in the generation and use of information and knowledge. It is these gains in efficiency that have made this work worthwhile in a business producing high-technology chemical-based products.

HOW HAVE ADVANCES IN COMPUTING TECHNOLOGY HELPED WITH EFFICIENCY IN THE ANALYTICAL CHEMISTRY LABORATORY?

Advances in computing technology have helped with efficiency in the analytical chemistry laboratory in four main ways. The first is through automation and simplification of analytical and synthetic tasks. This area includes the use of computer-controlled robots and measurement systems and can

improve repeatability and increased utilization of equipment. There are numerous examples of applications and vendors of combinatorial testing and synthesis; these are not reviewed here.

The second area is in information and knowledge management. Combinatorial methods produce large amounts of data. Managing the data and extracting useful information and knowledge from it has been made feasible through advances in computing technology. Even without the use of combinatorial methods, good information and knowledge management is valuable. Information provides value over time. Previous analyses can help with current problems, and historical information can help with the design of new materials and products. Making this information available in a usable and timely manner is an important benefit of a wired laboratory.

The third area is that of generating and maintaining data in electronic (digital) form. While this may seem like an obvious requirement for modern information management, it is a valuable first step on its own. Having data in electronic form greatly reduces the barriers to its use. Much of the time involved in applying modeling and chemometrics—the analysis of analytical data to extract more information—is consumed in collating and formatting data. Simply collecting and saving the data in electronic form allows more time to be devoted to developing more sophisticated calculations.

The fourth area is data analysis and chemometrics. One of the general efficiency trade-offs in routine analytical measurements is between sample preparation and data analysis and interpretation. Analytical techniques requiring less sample preparation often produce larger, more complicated data sets that increase interpretation time. The phenomenal increases in computing power and capacity have helped to reduce that time. In addition, the chemometrics techniques available today yield information not otherwise obtainable. The direct exponential curve resolution algorithm (DECRA) for separating mixture spectra is an example.[1]

EXAMPLES FROM A WIRED LABORATORY

Quantum—Model of an Integrated Spectroscopy Information System

In the late 1970s, the molecular spectroscopy laboratory at Kodak began to utilize computing technology to improve the efficiency and quality of structure elucidation using nuclear magnetic resonance (NMR) and infrared (IR), mass (MS), and ultraviolet and visible (UV/Vis) spectroscopy data. The ultimate aim was to automate the analysis of routine samples completely. At that time, our expert spectroscopists would receive a number of difficult analysis problems, but would also receive many samples that were routine characterization problems. For example, did the chemist successfully synthesize the material he wanted? We recognized that we could use computers and information systems to make our operation more efficient by automating routine analyses and by providing tools to aid with difficult analyses.

The components of the system that resulted from this project are illustrated in Figure 11.1. The complete system, called QUANTUM, combines spectral and structural analysis software with a sample management system (SoftLog) and a spectral database (SDM).

Historically, the system began with the research and development of analysis tools. As John Pople mentioned earlier, in Order to test our success, we needed to have data. Reference spectra associated with chemical structures were needed to develop and test analysis software for predicting spectra, given the structure, or the structure, given a spectrum. Databases of literature spectra were purchased and put into the system, but they were not adequate. The majority of compounds made at Kodak have never

[1] Willem Windig and Brian Antalek, *Chemometrics and Intelligent Laboratory Systems* **37**, 241-254 (1997).

appeared in the public literature. To test the analysis routines, it was necessary to gather information on Kodak-specific compounds. To gather that knowledge efficiently, the structure and associated spectra need to be readily accessible to the computer. This information was also needed for the analysis software to successfully eliminate routine work. It is not practical to enter that information into the computer just to get a routine analysis answer. It needs to be there for some other reason.

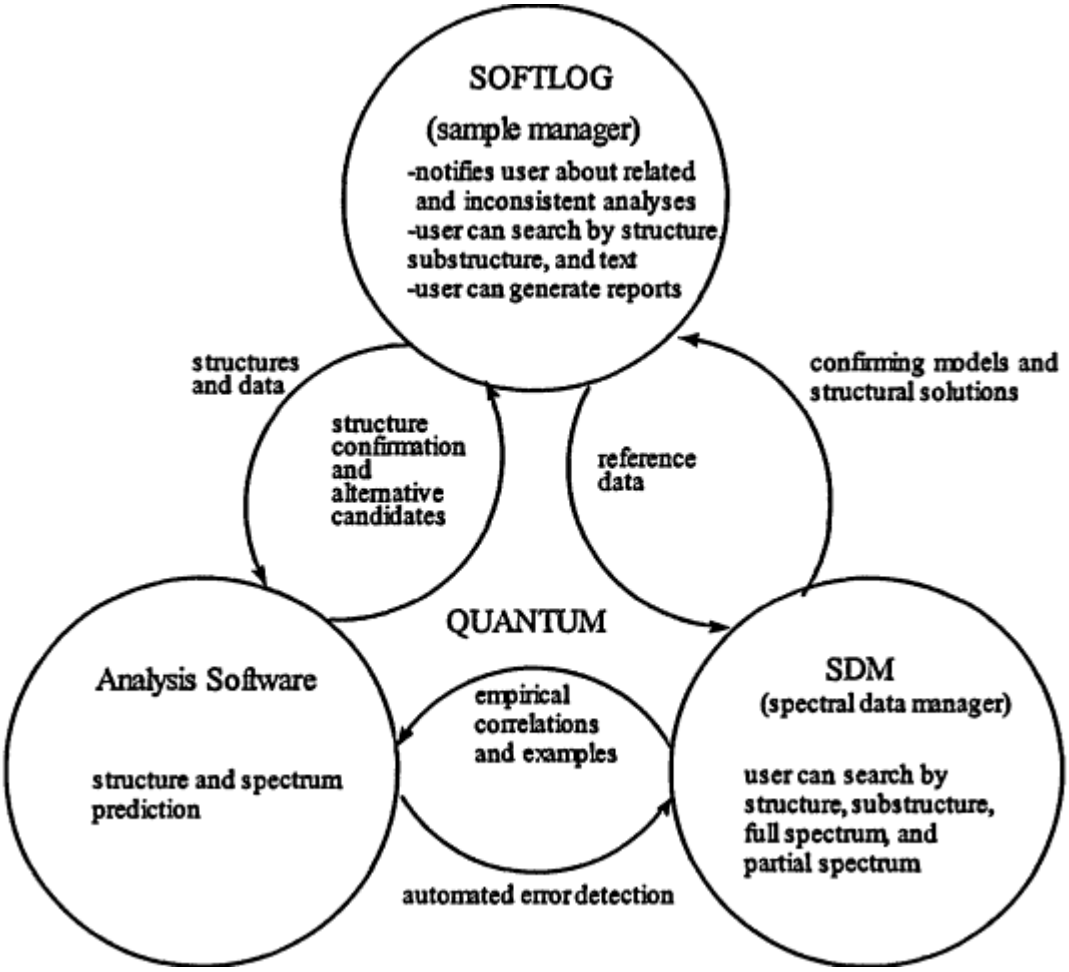


Figure 11.1
Components of an integrated spectroscopy information management environment.

Integrated sample management software (SoftLog) that incorporated chemical structures and spectral data was developed to meet this need. These systems are now commonly known as LIMS (laboratory information management systems).[2] SoftLog incorporated several important features not typically available in LIMS. These include full and substructure searches, spectral display and search, easy

[2] LIMSource, *LIMSource: the best site for LIMS and lab data management system info*, <http://www.limsource.com/> (1998).

incorporation of results into the reference database, and a logical interface to fully automated analysis software. In addition, the user is automatically informed of previous sample or material analyses, inconsistent results, reference data, and related compounds such as impurities, models, by-products, or precursors. The notification of other analyses of the same material included data from spectroscopy laboratories dispersed across the entire corporation. This was particularly useful, because it made analysis information initially obtained in a research environment available to analytical laboratories supporting development.

There is a lot of interaction between the components of QUANTUM. The analysis software is used to check the quality of data going into the reference database and provide predictions for current analyses. The reference database provides models for current analyses and the data necessary to develop the analysis software. SoftLog provided the primary daily interface for users and means for gathering information in electronic form. This is also the part of the system that has changed the most with changes in user interface and desktop technology.

This integrated model of a spectroscopy information system is useful. Even though QUANTUM is 15 years old, the underlying systems are still in use. The greatest use is now through a Netscape browser using the corporate intranet.

Walk-up Spectroscopy Laboratory—Instruments Online

As mentioned above, the initial goal was to make the work of spectroscopists more efficient. Part of that goal was to free them from spending time analyzing simple or routine samples. From the company's perspective of efficiency, the real goal was to reduce the amount of time between the chemist's initial awareness that he needed an analysis and when he got the data or analytical result. The pursuit of this goal has led us to develop walk-up laboratories where the chemists interact directly with automated analytical instruments.

The walk-up laboratories provide rapid access to high-quality, state-of-the-art analytical instrumentation via a one-stop structure characterization area staffed by experts. The laboratory staff maintains the quality of the instrumentation and the integrity of both the methodology and the data. They also work to develop new analytical techniques and methods to the point where they are robust, rapid, and automated enough to function in a walk-up environment.

One last goal of the walk-up laboratory is to provide access to the analytical data and information through a simple, user-friendly, at-your-desk interface. This interface should allow all scientists involved in the chemical commercialization process access to all analytical data generated throughout the company.

The first technique we provided in walk-up mode was NMR. To use the system, chemists enter the laboratory, place their NMR tube in an empty slot for the sample-loading robot, enter some information identifying themselves and their sample on a computer, and select their choice of experiments. A label is then printed, which is placed on a board next to the slot containing their sample. This label is used to identify the samples after the experiments are completed. The NMR spectra are usually available over the network, automatically phased and transformed, by the time the chemists walk back to their office. If needed, the raw data is available for reprocessing.

The NMR facility offers a 300-MHz Varian spectrometer with a 4-nucleus probe and experiments including ^{1}H,^{13}C,^{19}F,^{31}P, DEPT, COSY, and HETCOR. The autosampler will hold 50 samples that are processed in a prioritized order to minimize the turnaround time for most users. The facility is available 24 hour/day, 7 days/ week, with an average analysis time of 10 minutes. All of the data are saved on

servers and are available remotely. Spectral predictions are available using both Kodak and Advanced Chemistry Development software.

These NMR experiments provide information on the chemical environments surrounding individual atoms in a molecule. They can also provide connectivity information such as the number of attached protons or, from the two-dimensional experiments, what protons and carbons are next to each other. This information allows a chemist to confirm the structure of most molecules. Chemists recently hired by Kodak have all used this kind of information themselves in graduate school and are very comfortable with it. They have also remarked, “If only I had had something like this when I was in graduate school, it would have saved me so much time.” That, of course, was our objective.

In addition to NMR, the walk-up laboratory provides access to advanced MS, chromatography, and IR techniques. The MS system provides a 3,000-atomic-mass-unit range, atmospheric pressure chemical ionization (APcI) and electrospray ionization techniques, and loop injection, or short column, separation. Data output includes averaged, background-subtracted spectra for both positive and negative ions and a theoretical isotope pattern display based on the chemist's proposed formula. This provides a molecular weight and formula confirmation in an average of 3 minutes.

The chromatography system produces integrated chromatograms at five wavelengths with area percentages and a diode-array spectrum (160 to 600 nm) for each peak. The average analysis time for this technique is down to 15 minutes. One of these instruments will soon be linked to a mass spectrometer system. When this is complete, a single 10to 15-minute experiment will provide concentration, spectral, and molecular-weight information on mixtures.

The walk-up facility in place at Kodak is a nice example of how efficiency can be improved. Turnaround time for sample analysis has improved by 7 to 10 times, or probably more from the chemist's perspective. This kind of routine analytical analysis is no longer a bottleneck. In fact, chemists often use this facility rather than running thin-layer chromatography plates because it is just as fast and produces more useful information presented in a readily interpretable form. These efficiency gains have been accomplished by using automation to reduce analysis times and by placing the analysis in the hands of the people who need the results. It is successful because it is managed by analytical personnel who maintain quality, provide training and consultation for the users, and are rewarded for improving the efficiency of others. It provides additional value because the information is saved electronically and made available to all who need it.

Improvements in both analytical and computer technology have been critical factors in making the walk-up laboratory possible. The computer and networking technologies provide a robust interface between the users and expensive analytical equipment. They are used for data acquisition, processing, storage, retrieval, and presentation. Analytical and computing technologies have combined to make a system that is robust and automated enough for routine use and that has significant business value.

Electronic Information and Knowledge Management

Information is data in context. Information has value, but only if it can be readily combined with other information. Many science-based companies have been generating information in compartmentalized laboratories in ways that make it difficult to access. Paper- and people-based methods of information management delay a scientist's ability to use this knowledge at a pace consistent with the business need for cycle time improvement and increased efficiency. This is why electronic access to research information and analytical data is necessary.

One area where this efficiency is important is in the movement of new chemicals from research through scale-up to manufacturing. As a chemical moves from inception to final product, fitness-for

use specifications need to be established, and regulatory information must be filed with governments. Government filings require gathering all the compositional and safety information we know about a material. Knowledge of this information for related materials at the time of new research can be used in designing safer materials initially, rather than disqualifying them late in the commercialization process. When fitness-for-use specifications are established, all of the by-products that may be produced are considered. The spectra of many of these are identified initially in research samples. When manufacturing problems occur, quite often the problem is related to a minor component that has been identified somewhere earlier in the commercialization process. Having available a trail of information on that material saves a tremendous amount of time in an environment where time has a significant financial impact.

Information and knowledge management helps scientists learn from the previous experiences of others across the corporation. This becomes increasingly difficult as organizations and their collections of proprietary knowledge grow large and research occurs at worldwide locations. Without electronic access to it, efforts to use information would be very inefficient. An important area where the need is to access the data, rather than the final reports and conclusions, is modeling. Modeling to develop new and better materials is an important part of increasing research efficiency, but it requires electronic access to structures and property data.

Electronic access to analytical data also helps us to perform analyses the least number of times, maximizing the effects of previous results. Multiple scientists working on the same, or different, projects can make use of the results of the same tests. Data collected for one project may be of use years later on a new project. Compounds that were not suitable for one application may be good on another or may be good data points for modeling on other projects.

WIMS—Web-based Information Management System

The types of analytical information needed in an analytical information management system include project and sample information, chemical structures and reactions, reports and conclusions, and spectral and image data. The ideal system would provide an intuitive, easy-to-use graphical user interface that is platform independent (PC, Mac, or UNIX). It would be capable of easily displaying and manipulating images (spectra, structures, and figures) along with all other analytical information, would allow easy downloading of data for local reprocessing, and would easily link or cross-link to existing proprietary and legacy databases. In addition, it should be based on technology that is widely accepted and not unique to our own environment, is dynamic and continually developed for improved capability by many other people, is low in cost for software development and maintenance, and provides worldwide access. About 4 years ago, we realized that "the Web is the way" to meet these needs.

Based on the knowledge gained from the SoftLog sample management system, the new Web-based Information Management System, WIMS, was developed for spectroscopy.[3] WIMS has since been expanded to provide extensive sample tracking and data management across the analytical community. Significant advantages have been gained through data searching and allowing our clients to examine spectral information, reports (as text, Word, Excel, or HTML documents), and other analytical data directly. As a sample manager, WIMS is used to log samples in and out with descriptive fields customized by technologies, to attach structures, reports, and spectra directly to a sample or group of

[3] Stu Borman, *C&E News* **75**(4), 25-27 (1997); Douglas Brown, Antony Williams, and David McLaughlin, *Trends Anal. Chem*. **16**(7), 370-380 (1997).

samples, and to calculate throughput statistics. The database can be searched by any combination of sample fields, technologies, and reports. It also allows automated reports to be generated and e-mailed to clients. The commercial Web-based S3LIMS product from Advanced Chemistry Development was inspired by these original concepts.[4]

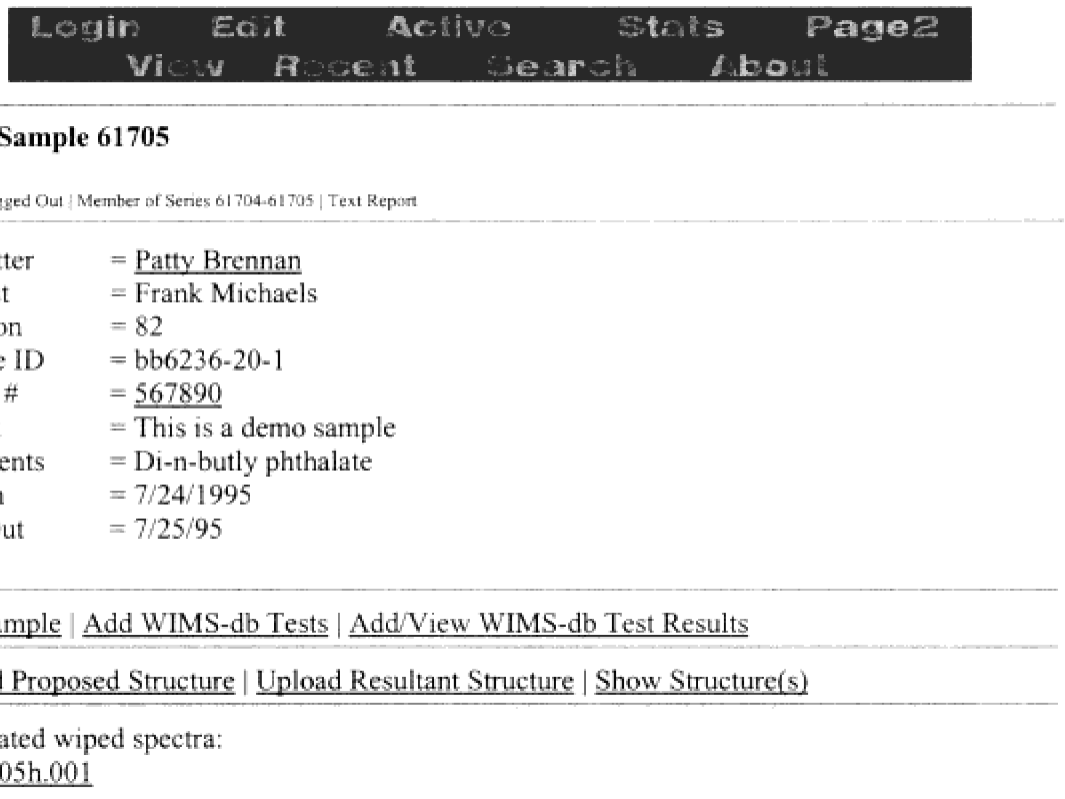


Figure 11.2
Typical WIMS sample view page for NMR.

Figures 11.2, 11.3, and 11.4 provide some examples of WIMS displays. A typical display for a sample in the NMR technology area is shown in Figure 11.2. In this view, the descriptive sample fields for this technology are displayed along with most of the important functional links, like uploading and viewing reports. An important feature of this view worth highlighting is the lack of clutter. An attempt has been made to fill the screen with as much information as is useful without wasting space that causes users to scroll in their browsers. Interesting graphics that do not add utility may be neat initially, but quickly become annoying to users.

Other links on the sample view page include links to display associated spectra (Figure 11.3) and associated structures (Figure 11.4). Following the associated spectral link retrieves the NMR spectrum from an NMR data server and displays it along with the acquisition parameters. The display can be zoomed and printed. The data are also available for local reprocessing if needed.

[4] Advanced Chemistry Development, *S3LIMS: spectral laboratory information management system*, <http://www.acdlabs.com/slims/> (1998).

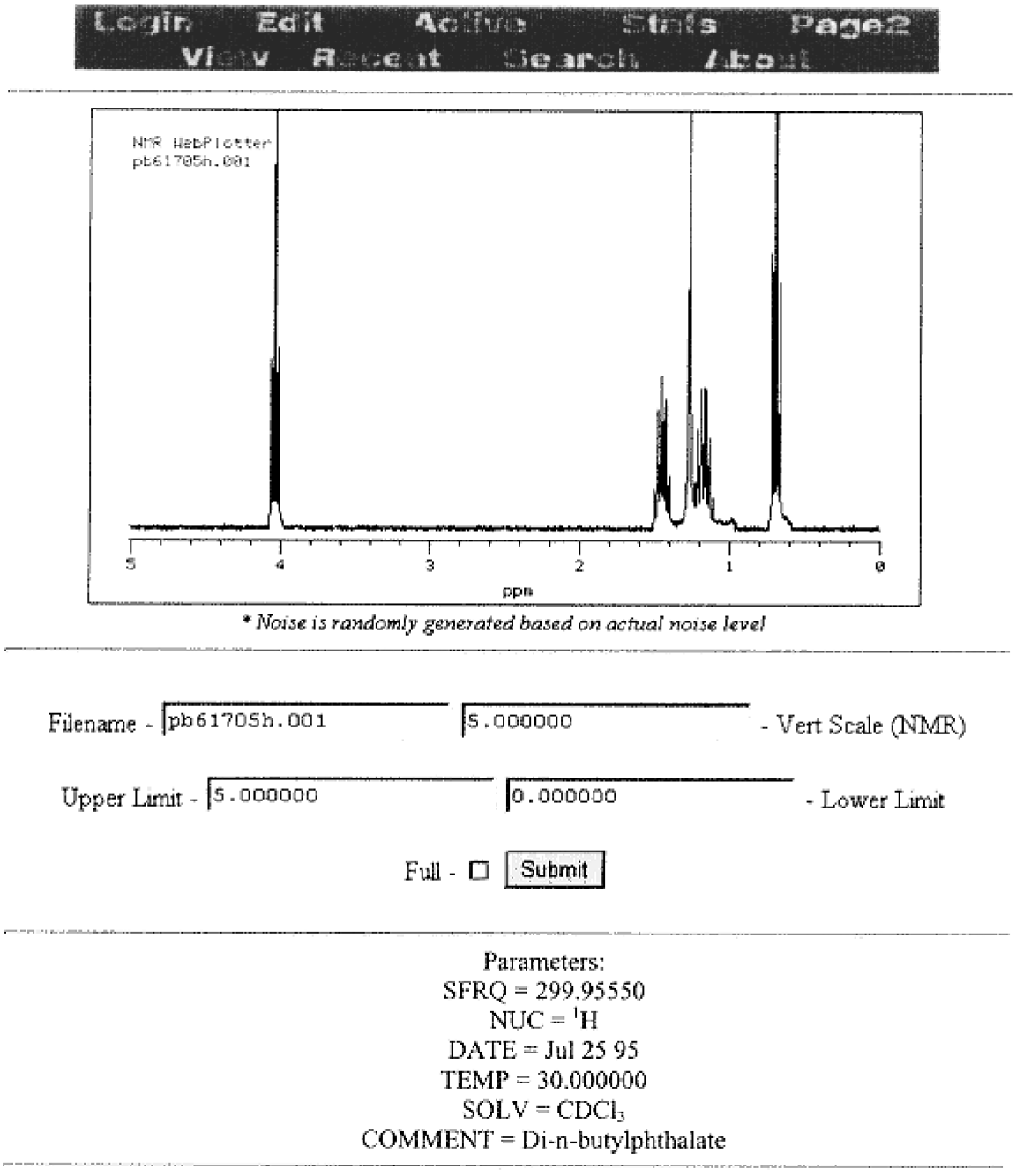
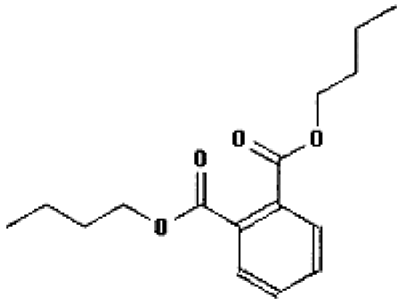


Figure 11.3
Typical WIMS display of a sample NMR spectrum.

Structure Result for Chem # = 567890

Molecular Formula: $C_6H_{22}O_4$
Formula Weight: 278.3483 Molecular Ion Mass: 278.1518012 Nominal Mass: 278



Get Molfile

SMILES: CCCCCC(=O)c1ccccc1C(=O)OCCCC

There is reference data in QUANTUM for the structure. All data , ^{13}C NMR , MS , IR , Other NMR
Predict ^{13}C NMR or ^{1}H NMR
Related structures exist in QUANTUM

LIMS Compound ID: NMR61705RA

Figure 11.4
Typical WIMS display of a chemical structure associated with a WIMS sample, showing links to additional information.

The associated-structures link retrieves the structures from the QUANTUM structure database and displays them with links to other information resources. Typical links allow for display of spectra from our reference databases, predictions of chemical shifts, and links to other WIMS samples. In addition, the structure can be downloaded and used in a local drawing or modeling package. This capability was developed by building a Web wrapper around existing software. It makes this information available to a much larger collection of less frequent users.

Changes in computing technology have had a significant impact on sample management. The most notable one is the advent of Web technology and the realization that the Web is the way for user interfaces. The use of Web interfaces is cascading through all of the chemical property databases in the company. Modeling tools are also becoming available on the Web. Sets of compounds may be submitted to a calculation server to have parameters and estimated properties determined.

Simple interfaces achieve the greatest use. The QUANTUM system illustrated in Figure 11.1 was originally used primarily by experts. When a Web interface was added, usage increased by approxi

mately 50 times. The Web interface has put valuable information within easy access of many more users. This provides real payback to the company.

Another point to learn from this experience is the value of information servers working as peers. The WIMS system involves several servers, each providing a service that is linked together via Web technology to provide the illusion of one system to users. A chemical structure server put up on the Web provides a simple mechanism for other developers to include structures on their Web pages. This is an important point, because it has been very difficult to acquire commercial software that will operate in this fashion. Most information system vendors develop software from the perspective that they are the center of the universe. All interactions happen initially from within their software, which is always in control. Their software will not operate as a peer.

The Electronic Laboratory Notebook

A model of the data-to-information pyramid is shown in Figure 11.5. A major research program may involve multiple projects, each with several experiments that may produce several samples requiring many tests that can produce lots of results. The amount of data present at a given level increases toward the bottom of the pyramid. Consequently, the bottom area has been first to make use of

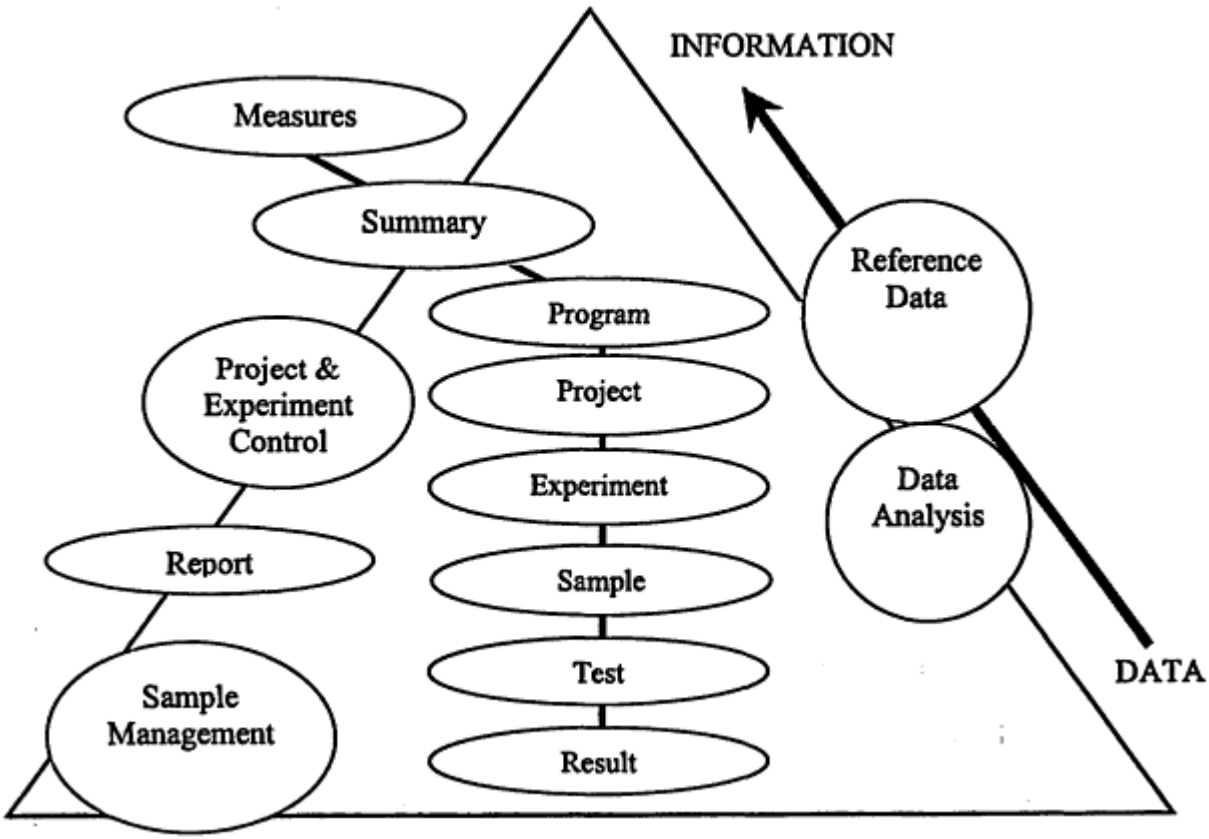


Figure 11.5
The data-to-information pyramid.

computing technology to help with its management. So far, the examples presented here have focused on results, tests, samples, and their associated information management and data analysis tools.

Over the past 2 years, work has progressed at the Kodak research facilities in England to apply computing technology to the next levels up in the diagram. Specifically, an electronic laboratory notebook has been developed to assist with the management of experiments, projects, and programs. Above this level are program measures and summaries useful for research management. These have not yet been addressed, as the traditional methods of providing summary reports and presentations will likely be adequate for several more years.

The electronic laboratory notebook (ELN) at Kodak is implemented as a collection of Lotus Notes databases and applications that enable the electronic storage and retrieval of experimental aims, methods, results, and conclusions. A few years ago, Kodak decided to switch to Lotus Notes for e-mail, making it a reasonable choice for this development. The ELN provides an environment that facilitates the sharing of information across research by means of controlled database access, searches, and hotlinks. It supports the principle of entering new data once and only once.

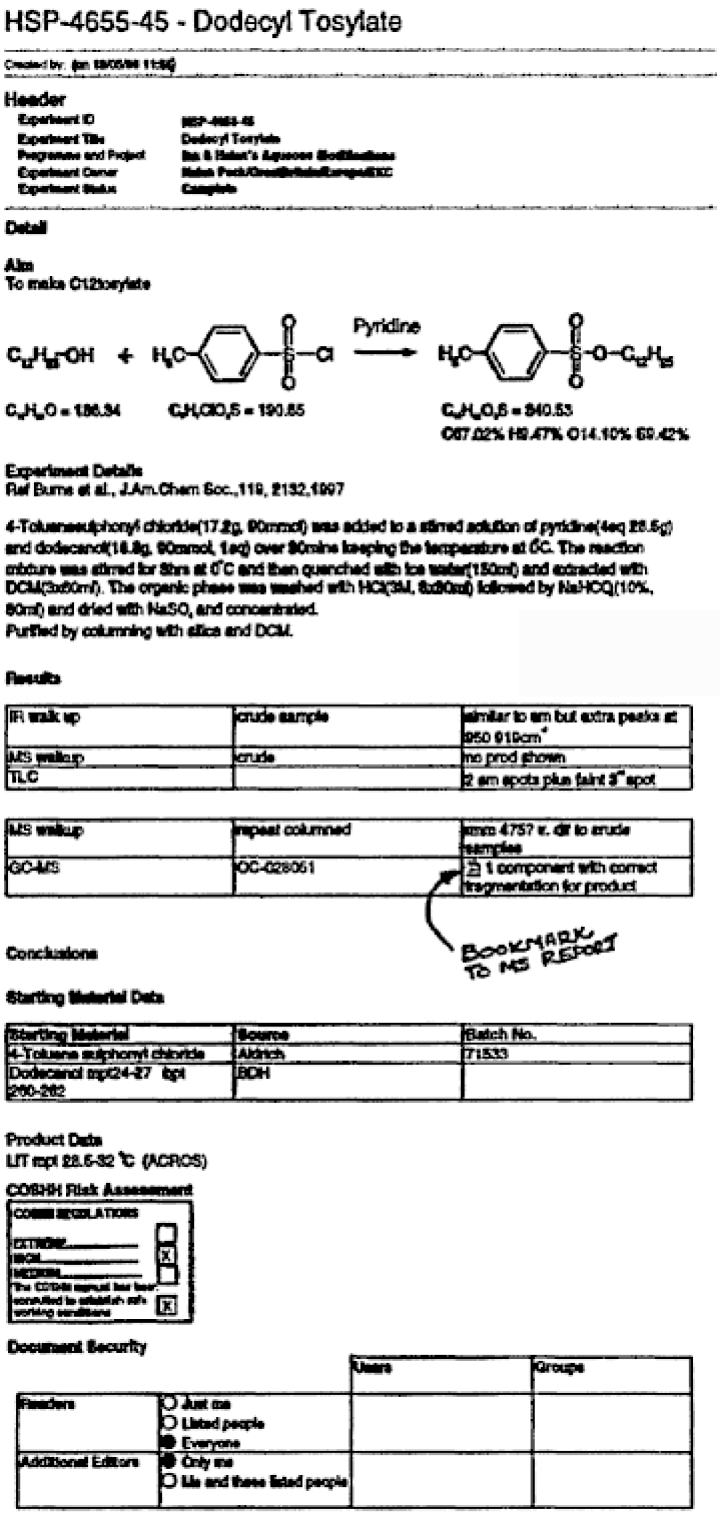
Before there was an electronic sample management system, chemists would submit analysis requests by completing paper forms. With the advent of an electronic system, some chemists would enter the information into the analytical data system and some would still submit paper requests that an analytical technician would enter. The paper request forms for spectroscopic analysis included space for a chemical reaction to be entered. This is useful for identifying impurities in structure characterization problems, but there was never enough added value to enter it into the analytical systems. Now the entire reaction and experimental conditions are maintained in the ELN and are available to the analyst through a hypertext link. The information is captured and maintained at its original source in a way that is useful to the originating scientists. Now that it is in electronic form, the information can be leveraged throughout the corporation.

There are four main Lotus Notes databases underlying the ELN. They are used to store experiments, projects and programs, reports, and summaries. Each database has templates for creating entries. Summaries can be generated automatically by tools in the ELN or entered by the scientists as part of their experiment. It is expected that this database of summaries will provide an important means for searching the ELAN.

An example experiment from the Kodak ELN is shown in Figure 11.6. This is a page from a typical organic chemist's notebook. It shows the aim, chemical reaction, experimental details, results, summaries, and security. The security functions allow the author selective control over who can read and modify the document. There are also areas for entering or attaching any information or file the chemist wishes—in this case she added information about the starting materials. It is probably worth noting that this page closely resembles what the organic chemist would have entered into her hard-copy notebook.

Notice that the chemist performed a number of walk-up tests herself and included the results. There was some question about the mass spectroscopy results, and an expert analysis was requested. The result of that is entered in the ELN as "1 component with correct fragmentation for product" with a bookmark to the complete report, which is shown in Figure 11.7. The full report shows the reaction from the ELN, the spectra, and the comments from the expert analyst who adjusted the spectroscopic experimental conditions to obtain the result.

The electronic laboratory notebook is envisioned to be the tool of choice for scientists to log experimental aims, results, and conclusions. It is expected to enable knowledge sharing while maintaining security, allowing greater collaboration between researchers and increased productivity. There are good reasons to believe that it will be successful. First, computing and computing technology have progressed to a point where a successful ELN can finally be delivered. Second, and most important, the



HSP-4655-45 - Dodecyl Tosylate

Created by: [illegible]

Header

Experiment ID	[illegible]
Experiment Title	Dodecyl Tosylate
Programme and Project	[illegible]
Experiment Owner	[illegible]
Experiment Status	Complete

Detail

Aim
To make C12tosylate

$C_{12}H_{26}O$ = 186.34 $C_7H_7ClO_2S$ = 190.65 $C_{19}H_{32}O_3S$ = 340.53
C67.02% H9.47% O14.10% S9.42%

Experiment Details
Ref Burns et al., J.Am.Chem Soc.,118, 2132,1997

4-Toluenesulphonyl chloride(17.2g, 90mmol) was added to a stirred solution of pyridine(4eq 28.5g) and dodecanol(16.8g, 90mmol, 1eq) over 90mins keeping the temperature at 0C. The reaction mixture was stirred for 8hrs at 0°C and then quenched with ice water(150ml) and extracted with DCM(3x50ml). The organic phase was washed with HCl(3M, 6x90ml) followed by NaHCO(10%, 90ml) and dried with Na_2SO_4 and concentrated.
Purified by columning with silica and DCM.

Results

IR walk up	crude sample	similar to sm but extra peaks at 950 910cm^{-1}
MS walk up	crude	no prod shown
TLC		2 sm spots plus faint 3rd spot

MS walk up	repeat columned	mmm 4757 v. dif to crude samples
GC-MS	OC-028051	1 component with correct fragmentation for product

BOOKMARK TO MS REPORT

Conclusions

Starting Material Data

Starting Material	Source	Batch No.
4-Toluene sulphonyl chloride	Aldrich	71533
Dodecanol mpt24-27 bpt 260-262	BDH	

Product Data
LIT mpt 28.5-32 °C (ACROS)

COSHH Risk Assessment

COSHH REGULATIONS
EXTREME
HIGH [X]
MEDIUM
The COSHH manual has been consulted to establish safe working conditions [X]

Document Security

		Users	Groups
Readers	○ Just me ○ Listed people ● Everyone		
Additional Editors	● Only me ○ Me and these listed people		

Figure 11.6
Display of a page from a chemist's electronic laboratory notebook, showing a bookmark to an MS report.

Sample Name:HSP-4655-45.2 (W14681)

Customer: Helen Peck

Created By: Sandra Shurvell on 01/06/98 at 12:20

JOB ID:

KODAK LIMITED

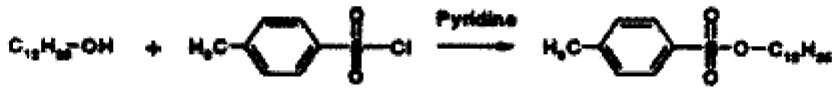
ANALYTICAL LABORATORIES
RESEARCH DIVISION, W92

To . . . : Helen Peck
Research
Harrow

Sample name . . . : HSP-4655-45.2
Sample ID . . . . : OC-028051
Job ID . . . . . : 00000002
Date logged in. . : 29-MAY-1998
Date of report. . : 1-JUNE-1998

Test . . . . . : Mass Spectrometry
Authorised By . . : Sandra Shurvell Tel : 33198 Date : 1-JUNE-1998

Component name	Result
MS Report.....	W014681.RPT
Test Reference Number.....	W014681



MS Report

The GCMS shows a single major component, which although not showing a molecular ion, shows fragmentation consistent with the proposed product.

HSP4655-45.2 GCMS

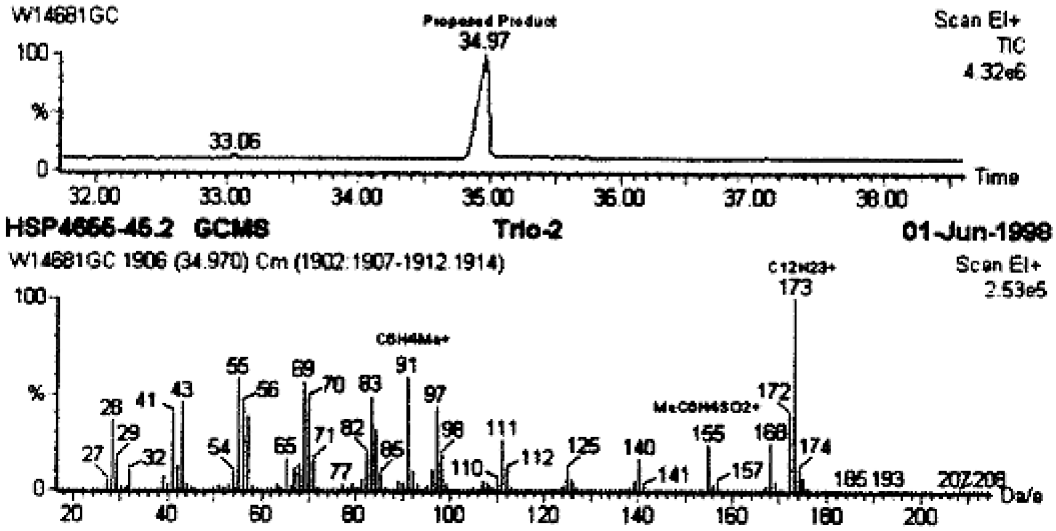


Figure 11.7
The MS report linked to the electronic laboratory notebook page shown in Figure 11.6.

ELN prototype was built at the request of chemists. Once the prototype was in place, 80 to 90 percent felt it was good enough to use and would recommend its use to their colleagues. This desire is widely shared among the research scientists. It is a reflection of the strong drive for continually improving research efficiency and the generation of more knowledge per unit time.

THE FUTURE

Information Management

In the wired laboratory of the future, all scientists will use an intelligent electronic laboratory notebook linked to all data-generating equipment. This will automatically provide all relevant information to the scientist and capture all of the knowledge that is generated. It will allow research to progress using sophisticated experimental design and modeling. Most materials will be modeled before they are made. In the intelligent electronic laboratory notebook, objects will be recognized as they are entered and linked to underlying databases automatically. The environment will be built with integrated links between systems communicating as peers. Much of this vision will be become reality over the next 5 to 10 years. Although electronic laboratory notebooks have been discussed and prototyped in the literature for some 10 years, there is now end-user pull for them, and computer hardware and software technology can now support them. They will drive much of the electronic laboratory of the future.

Success, however, will require the development of better search systems. Knowledge is information in context. Searches through large information resources must provide good contextual searching and answer set refinement. Without such tools, so many false positives or irrelevant answers are returned that the results are useless to the researcher. The summary database in the ELN may help with this problem. It is also an area of active research and development driven by the need for Web search engines.

Information systems will also benefit from better input devices. They will make it easier for scientists to interact with computer systems, particularly when generating information. These devices will be easily carded into the laboratory environment and written or drawn on, like pages in a paper notebook. Interfaces will allow chemists to draw their chemical structures on a sheet of paper and have them upload directly into the computer with a connection table that is searchable. There are already products on the market that are consistent with this direction. One example is the Cross notepad, [5] which allows information to be written on a portable tablet and uploaded as an image. It comes with software that can be trained to convert neatly written words and phrases to text.

The continued move to information systems will result in a nearly paperless laboratory. It will no longer be necessary to print reports so that they can be mailed or archived as they are today. This prediction applies to the laboratory and not to the office environment. People will still print information until another medium is found that is just as convenient and cost-effective. The need to archive information in printed form will be replaced by digital means.

Some obstacles exist for information management. One is the need for the current rate of increase in computer processing and storage capacity to continue or accelerate. The information systems will need to store staggering amounts of data. Perhaps a more limiting obstacle is the lack of good commercial software. Software must be available at reasonable cost and quality that is simple to support and maintain and is not wasteful of hardware resources. Software from multiple sources will need to

[5] A.T. Cross Company, *Cross Pen Computing Group: The Pad*, <http://www.cross-pcg.com/crosspad/> (1998).

cooperate in a peer-to-peer manner (standards may help with this). Better methods for dealing with software development are needed to improve quality and reliability and to deal with disruptive changes in technology. Unfortunately, the realities of the current software market do not appear to reward quality adequately. Differentiation and time to market seem to drive the greatest short-term profits.

Data Analysis and Instrumentation

Analytical technology in combination with data-analysis techniques will continue to advance, reducing the time required for sample preparation and data interpretation. In some cases, current analyses will be replaced by drastically different technologies that, perhaps after more calculation, yield the same or greater informational value. There will be a strong drive toward small-scale automated syntheses and testing. There will also be more applications of on-line and in-line sensing and control, with analytical instrumentation connected to networks and controlled remotely. These applications will extend the concepts of speed and robustness demonstrated in the walk-up analytical laboratory toward smaller size and price.

Success in these areas will likely involve greater use of embedded systems and plug-and-work components. Proposed extensions to Java™ may prove useful in this arena.[6] Increasingly, instrumentation will be controlled through Web interfaces. Several companies are beginning to produce low-cost boxes that can be used to connect almost any device to the Web. One example is advertised as the world's smallest Web server.[7] The complete Web server is less than 4 inches square and is configured to provide real-time weather data from Cambridge, Massachusetts. It is designed to be simple and low cost. These devices may be excellent alternatives to PCs or UNIX boxes for connecting instruments to the network, particularly from a support perspective.

As more information becomes available only in electronic form and computerized processes become a critical part of business processes, dependence on the network increases. A robust, high-bandwidth network becomes a requirement. This may be more a cost issue than a technology issue.

Virtual reality techniques have been studied for decades. They provide excellent mechanisms for people to understand and interact with visual information. Particularly useful applications of virtual reality have occurred for fighter pilots and people with disabilities. Virtual reality is a tool that uses the senses to transmit a lot of information to the brain quickly in the form that it normally processes. It is meant to help unlock the power of the human mind. Interfaces common in the laboratory today are very poor by comparison. Widespread use of virtual reality is limited by our ability to collect appropriate information to display in this mode.

Tandem techniques that produce multidimensional data are now common in analytical chemistry laboratories. They are useful because they generally use a small amount of sample, require minimal sample preparation, and have relatively fast analysis times. At the same time, they generally produce large amounts of data that often require relatively long interpretation times. Techniques that produce information in three-dimensional space, such as 3D NMR, are easily understood when viewed through virtual reality techniques.

At present, it is not clear that virtual reality is necessarily the appropriate tool for understanding much of the multidimensional analytical data produced today. This is unfortunate, because computer

[6] Sun Microsystems, Inc., Jini™ technology lets you network anything, anytime, anywhere, <http://java.sun.com/products/jini/> (1998).

[7] Phar Lap Software Inc., The World's Smallest Web Server, <http://smallest.pharlap.com/> (1998); Dr. Dobb's Journal, October, 40-46 (1998).

monitors are also inadequate for displaying this data. Spectral interpretation often involves the fine detail spread across a wide, high-resolution x-axis. A computer monitor allows only a small portion of the detail to be viewed at one time.

There are some areas where scientists would like to "see" results to understand them. These include seeing how molecules are interacting with each other or flowing through processing equipment. The challenge to the analytical community is to devise real-time measurements that, when displayed with virtual reality, will enable this understanding.

SUMMARY

There is a strong business need to generate new products with ever increasing efficiency. Automation of the data acquisition and information management functions of the laboratory can increase the efficiency with which the knowledge generated by research is applied to new products. This arises through increased ease of use, accuracy, and quality of information and knowledge made possible, in part, by the advancements in computers and computing technology.

As systems are developed, it is important to remember to keep them simple. They must be simple to build, simple to use, and simple to maintain. Software technology changes relatively rapidly. Keeping it simple helps incorporate new technology faster, deal with obsolescence, and deliver functionality to users faster. Rapid delivery of simple systems returns the greatest value.

The Web is the way for user interfaces, as long as the Web belongs to everybody. The peer-to-peer model of the Web is an excellent way to build sophisticated information systems from simple components.

Good-quality, reliable software available at reasonable cost is one of the most critical needs for the future.

CONTRIBUTORS

Many people have worked on the wired laboratory at Kodak, mostly part time. They have contributed to the development, philosophy, recognition, and acceptance of the value of working electronically. They are as follows: Brian Adams, Christine Alvarez, Brian Antalek, Gustav Apai, Todd Beverly, Derek Birch, Caroline D. Bradley, Doug Brown, Don Bushman, Juris L. Ekmanis, Nancy Ferris, Tammi Flannery, John Flynn, Susan M. Geer, Joan M. Hessenauer, J. Michael Hewitt, Peter Horne, Andrew J. Hoteling, Thomas C. Jackson, Emily Jones, Thomas F. Kaltenbach, Philip LaFleur, Mary Lee Lasota, William C. Lenhart, Ilia Levi, Thomas Marchincin, William McKenna, David McLaughlin, Frank M. Michaels, Stephen D. Miller, Wendy F. Miller, Peter Monacelli, Dominic J. Motyl, Vi Neri, Ian Newington, William F. Nichols, Ed Osborne, Alan Payne, Julia Pich, Bob Price, Ted Sears, Craig Shelley, John P. Spoonhower, John Trigg, John Paul Twist, Jon Waterhouse, Luann Weinstein, Antony Williams, Willem Windig, Barry Wythoff, and Agyare Yeboah.

DISCUSSION

David Smith, DuPont: David, I am very interested in how this walk-up lab is located with respect to the community that it is trying to serve. I have worked at Kodak Park, so I have an idea of how large it is. It is as large as the Experimental Station, larger in fact, and frankly, I have a hard time imagining that a chemist is going to walk from one end of the Experimental Station to a walk-up laboratory to get a sample back in 10 minutes.

So, the question is, Do you have multiple instantiations of this laboratory in various parts of Kodak Park, or is it just located in one place with a high-density of chemists in the area?

David McLaughlin: The walk-up facilities are located in areas where there are high densities of people that need to use it. So, for example, in the main research complex where a lot of synthetic organic chemists are located, there is a concentration of analytical tools for structural characterization. In Kodak Park where the scale-up and delivery to manufacturing operation occur, there is, also, a lot of synthetic work that goes on. So, we have a very similar facility, but with techniques that are suited specifically to that environment. In yet another set of buildings there is testing of emulsions and photographic properties.

So, the walk-up facilities are distributed where the need is. When we first put up a walk-up facility there was only an NMR. We put it on the second floor. There was another NMR on the third floor. The chemists on the sixth floor would come and use the one in the walk-up facility on the second floor. The ones on the third floor would generally use the one on the third floor, even though it was an old instrument and gave poorer-quality results. They used it because it was convenient. Convenience is a big part of it. It is similar to the ease of use I described for the Web interface. If you create an easy, simple-to-use interface, then all of a sudden lots of chemists will begin using it. So, you do need to locate the walk-up facilities close to where they are needed.

Session 3

Panel Discussion

Sam Kounaves, Tufts University: I have two questions. First, a comment directed to the people sponsoring this conference. Since it is on computers I think it would be interesting, if it is possible, for the participants to make some of their slides available on a Web site for the rest of us to use. Most of us are going to go back to our departments and to our institutions, and some of us would like to present this information to a wider audience. Having a simple PowerPoint summary, credited appropriately, that we could integrate into our own talks would be very useful for distributing this information more widely.

My question deals with an issue that is going to arise again in this afternoon's talk, that is, archiving. When I did my Ph.D. thesis years ago, I did it on an Apple II and I even kept some of my lab notebook on an Apple H computer. When we wanted to go back to it one day to get some information, it was practically impossible. I had to find an Apple II computer, plug into it, and try to get the information out. Several of our speakers this morning, and I guess all of them in some ways, implied that there was software that they had been using to archive the information, Lab Notebook, Lotus Notes, etc. I am still wondering how this is going to work out in the future.

I know one way my thesis can still be available, for example, is through an institution that is dedicated to archiving. University of Michigan Microfilms, for example, is, in theory, still archiving that information and will eventually switch over to a digital format, and it will still be available. What are your thoughts on archiving information? How can we go about doing this so that the lab notebooks are useful years from now? Are there any ways that you can see this happening, or do you have any thoughts on this process?

Raymond Bair: There are a couple of approaches that people are taking, largely along the lines you might expect. One approach is to require the makers of the notebook or document management system to provide a certain degree of compatibility with future formats, and future versions of their product. This kind of requirement is coming out of some of the commercial electronic notebook efforts. This is going to be mandated by the people that are buying these notebooks, the large companies.

However, that doesn't solve all of the problems; it just addresses the document management

company's formats. There are some interesting challenges ahead. For example, what if you stuck another kind of document into this commercial notebook or document management system? They are not really responsible for all of the different kinds of documents and data you might use. There's also an issue of progressive conversions, for keeping file formats up to date as time passes. That brings in issues of fidelity—if we convert files, how can we assure ourselves that the converted objects are still correct? There are challenges with electronic signature systems, too, since you have signed the original binary file, which has not been translated. So, what does it mean legally when you convert that file in the future? How do you retain that authentication that you had in the past? So, in addition to the format issues there are also issues of authentication.

By the way, my slides are on the Web at <http://www.emsl.pnl.gov:2080/docs/collab/presentations/ppt/csr/>.

Bridget Carragher: I think this is a problem. You cannot even read a Microsoft document that is one version behind the version on your desktop. If you cannot do that with Microsoft—which is probably the most ubiquitous software around—we are in a lot of trouble. But I think we are not the only ones facing this problem, and I think again, the scientific community isn't going to drive this problem. This is a huge problem for the world as the world moves onto the Web, and I think there are going to become tools that do automated updating. But where is your thesis now? Probably a printout somewhere is your real evidence that you wrote it, and I think that in part will continue to be the case.

Clint Potter: In a sense I think you also have to throw stuff away because you cannot keep everything. Perhaps you don't have the original data for your thesis anymore. So, you have got to be smart about what you save and think about what the things you save are—the things that go into libraries or university microfilm services.

Bridget Carragher: You publish things that you want to keep. The publishing record is partly what is there.

David McLaughlin: I think you should try to save the information in a format that you can easily move forward. The more open the format, the better. If you store your Word documents in Rich Text Format, that may give you more forward viability than a binary Word document. We often store spectral data in an ASCII format that is easily readable. It takes more space to store it in ASCII, but we know we can move it forward. An important part of our plan is to convert the format of our information as future versions of software may require.

David Smith, DuPont: We said several times during the course of the meeting that software is important, and perhaps just as a sanity check for myself, I would like to ask Susan about the component-based, object-oriented paradigm for software development. From your viewpoint as a computer scientist, is this really a viable approach for the future development of software or is it just a fad that is going to disappear in the next 5 years?

Susan Graham: I think it is more than a fad, and the reason is that structure and structuring are very, very important, and this provides a structuring mechanism. Object-oriented programming is just a structuring mechanism, and if you get the structure wrong then it is going to be just as bad as any other disorganization. But I think it is an approach that is only now possible because it requires more heavy-duty computing and in particular image sizes. Storage is used much more, and so the ideas are actually

quite old. The ideas of object-oriented programming are from the 1960s, and the technology has now caught up to the point where it is viable.

So, I think it is going to evolve, but I actually think it is a step forward.

David Smith: In the presentations we have heard the word "complexity" used quite often, and yet I haven't heard anyone mention, for example, the work on complexity that is going on in the Santa Fe Institute, such as concepts like autonomous intelligent agents in the software field. Does the panel believe that they will have any real impact on the class of problems that confront us?

Susan Graham: Clearly they have impact. There are opportunities there, and there are risks, and the risk with autonomous agents is that you no longer have control, and so, with the best of intentions the agent may be getting in the way, particularly if it is imperfect.

Those are all ways of managing intellectual complexity, and that is one of our biggest limiting factors—that it is hard for a person to get his head around everything that is going on, and the more some of those issues can be compartmentalized, the better off we are going to be.

Raymond Bair: There is no question that agents are going to have value in doing a number of useful things. However, Tom Finholt's hype curve comes to mind (from his talk last night). There is a considerable gap between the reality of what can be done now with intelligent agents, and some of the talk about them.

Thomas Finholt: The digital library projects are a good illustration of the gap between reality and science fiction, if you will, with regard to intelligent agents. Particularly in the Michigan digital library projects, the strategy has been to use an intelligent agent architecture for organizing bodies of information. I don't do that work, but I have followed those projects closely. Today, there is a huge gap, in my opinion, between the prototype applications that have been demonstrated and systems that will stand up to the rigors of everyday use (i.e., operational production systems). I think we can say that with respect to intelligent agents, we may be where we were with object-oriented coding in the 1960s and 1970s. That is, the software architectures are not there yet to truly implement the idea, but the further development of intelligent agents is definitely something to monitor for the future.

Stephen Heller, National Institute of Standards and Technology: Just a couple of comments about archiving, which I think is more a red herring than anything else. In fact there is no obvious ultimate solution, because of the changes in technology, so I would like to ask a question of the panel. How many of you actually have a real printout on a piece of paper of your bank statement or actually have physical stock certificates as opposed to all this stuff being stored somewhere electronically?

My feeling is that between the stock market and the bank accounts in the world, most people have fairly significant concerns about their resources, and concern about some of these scientific lab notebooks probably pales in comparison to the amount of concern people would have with problems with those financial resources.

I don't think people walk into their banks and ask for proof that their bank accounts are being properly archived, and their dividends and stock certificates are properly recorded.

So, it is a questionable issue to bring up at this point and for the foreseeable future, I think.

Susan Graham: I have a question for the people who were talking about electronic notebooks. One of the purposes of an electronic notebook is to have a historical record that is used, among other things, for

establishing priority and for integrity concerns in science. Once the record is electronic, what are the safeguards that you are using to make sure that you have the benefit of binding and the benefit of the fact that you chemists can analyze the page to see whether it has been altered and things like that?

Raymond Bair: I am not quite clear on what you mean by binding.

Susan Graham: Traditionally the notion was you didn't use a loose-leaf notebook; you used one with a binding so that you knew the order in which the record had been kept.

Raymond Bair: The approach that we have taken in our notebook conforms to the traditional model. If you would like to remove an item in your notebook from view you may delete it, but it doesn't go away. It becomes an icon, and it says, "Deleted," but you can retrieve what was there. There is a genuine need to be able to mark out stuff that was wrong, for example, so you do not get confused in future searches. However, that doesn't solve all the problems scientists have. There is a genuine need for something we haven't fully developed a concept for, a scratch pad of sorts: temporary information that exists for some intermediate time before it is canonized in a notebook. People are still working on concepts like this.

Participant: How do you prevent altering of the notebook?

Raymond Bair: You prevent alteration with the same kinds of technologies that electronic commerce is adopting to prove that you are an owner of a transaction. You can compute a hash code of an object of any size, and use public/private key technology to validate that the document has not been changed. This is your digital signature. You can also use a trusted time authority, along with your document hash and public/private keys, to establish an unchangeable date for the signature.

Susan Graham: But my question was actually prompted by something David said in which he explained how beneficial it was to have links. If you have links and particularly if you have URLs, then how do you know that the document you are referencing hasn't been changed?

Raymond Bair: If you are really going to have this for a record, for example, to determine priority, you cannot put a link to something that is temporary in the notebook.

David McLaughlin: Before devising a solution to this problem, I think you must give some consideration to the amount of effort it will take versus the need to prove the case you propose. For example, I have heard of cases where scientists have published fabricated results. From a scientific perspective, results are not considered valid until somebody else has repeated them.

Patents are used to protect intellectual property. With the exception of the United States, the critical date is when a patent application is filed, not the date the discovery was made.

In the United States, the date of invention is often established using laboratory notebooks. The primary requirement is that the pages be dated, signed, and witnessed. It is fine to keep the notebook pages in a loose-leaf binder. One pragmatic way to deal with the legal issues of electronic laboratory notebooks is to print out each page, including all the links, sign it, date it, witness it, and put it in a binder. Most of the lawyers I have spoken with believe that this is not really necessary. They believe that the Patent Office would accept an electronic lab notebook when a log is kept of every modification that is made. The logs can be written to optical disks with a date and time stamp and, if warranted, a digital signature.

If every change you make to your notebook is written into a log that you do not have access to, then it becomes very hard to fake entries. Deception would probably require a conspiracy, more than one person. I see no need to make an electronic notebook any more tamperproof than a paper system. Restricting unauthorized access to the information is of greater concern.

Stanley Sandler, University of Delaware: I am concerned about remotely operated or Web-operated equipment, because we had problems in our department with, a word I haven't heard here yet, a hacker.

There is great potential for a hacker to unknowingly cause equipment damage or real safety problems. No matter what degree of security we have, there is always a hacker that is going to be able to get through. How does one protect oneself and one's equipment?

Bridget Carragher: You cannot. You can do the best you can, and again, adopt all the tools that are available. We password-protect our instruments and we take various precautions like that, but in truth you cannot guarantee protection. But most of the interfaces we can build using Web browsers are pretty fail-safe. We have had kindergartners using these instruments, and they do not obey the rules. They bang on all the buttons and hit everything at once, and maybe the high schoolers are even worse. They treat an instrument as a video game. So, you can protect your instrument by your user interface and disallow things that would be dangerous. I think that is the more important thing—to build in those safeguards in the software.

Clint Potter: I think you could take the essentially same steps taken in the security world for workstations and computers. As that technology gets better, it can be incorporated into remote instrument technology. I don't think we should be inventing new security mechanisms.

Bridget Carragher: No, we will take advantage of whatever is out there, but we have 100 machines in our facility. They get hacked into all the time. There is nothing we can do about that except deal with the problem when it comes up.

John Pfeiffer, Air Products and Chemicals, Inc.: Let us take that question one step further if we can. One of the things that you at Illinois and at Kodak are doing is making very sophisticated tools and very sophisticated methodologies accessible to knowledgeable but maybe not expert users. So, a risk is that the knowledgeable user will abuse the capability unknowingly. How do you reflect on that? How do you, if you will, put some bounds on that as you provide these tools via Web interfaces or whatever easy-to-use interface?

Bridget Carragher: You mean they can gather data and misinterpret it?

John Pfeiffer: Exactly. One can complete a statistical analysis that is invalid, extending beyond the assumptions built into the technique, and then the scientists may draw incorrect conclusions.

Bridget Carragher: You can do that right now. You can sit in front of an electron microscope and twirl those knobs and get it completely wrong and yet believe the data you are getting. I don't think that is any different whether you put an intelligent or unintelligent user interface in the front of it. We have had this argument many times in my community—you know, that if you make it too easy to use, everybody will come along and misuse it. I think that is not so. I think that if you make it easy to use, you can help people understand what they are doing. You can make it much easier to repeat things using different

parameters. You know, if you are sitting in front of that instrument, and it takes you 3 days to get the data, you are much more inclined to believe the results and not try to repeat them than if you can just say, "Oh, I will just rerun this experiment and check it again with three different parameters."

So, I absolutely don't agree that making things easy to use necessarily lets them be more abused. I think you can abuse data any way that you gather it.

Clint Potter: I think the same thing is shown in molecular simulation packages in the sense that you don't have to write your own code anymore, and you don't have to understand the exact details of all the algorithms, but people are using these things, I guess. I don't know anything about chemistry.

Raymond Bair: Also, I think that distance from the user of the instrument doesn't absolve you from doing some training with that person. You're trying to accomplish the same kinds of things electronically that you would do if you had that person visiting in your lab. The training requirement doesn't go away just because the instrument is remote.

Bridget Carragher: But we are not paternalistic about it. If people want to use the instrument, they should use the instrument. It is ultimately the scientist's responsibility, just as it is now. I don't think we know how remote access changes the way data are gathered except to maybe make it easier.

David McLaughlin: From a walk-up lab perspective, the question could be stated as, Is it appropriate to take an expert analytical chemist out of the loop given that then the end scientist could misinterpret the data? I believe that scientists have a vested interest in not misinterpreting the data that they obtain, because proper interpretation helps them continue with their work and meet their goals. One practice that we follow is to place the walk-up facilities right next to the analytical experts who are working on the more difficult problems. There is always someone generally available during regular working hours to answer questions. We also require training before anybody can use the instruments and offer training classes on how to interpret the data, usually once a year. While the training helps meet some of that need, I think the solution still is having an expert available and approachable. This approach is similar to the examples of collaboratories discussed here that allow you to send e-mail and establish working sessions with an expert. In our case, the expert is physically nearby, making it very easy for a person to ask a question. That is how we try to avoid misinterpretation of data.

William Winter, SUNY-ESF, Syracuse: I wish it was that simple, but I don't really believe that it is. I remember that when the first PC versions of things like MM2 came out, an organic chemist in my department came running up to me with a picture he had drawn on his PC plotter clearly showing a planar *cis*-peptide linkage that he had obtained and claiming this must be right. Actually, it was an N-acetyl glucose linkage but it was the same idea, and it should have been *trans*. My colleague's conclusion was that because it came from a program from a respected person the result had to be right, and that was the end of it as far as he was concerned. Similarly, we have to do something to make people question these things and not think that just because it comes from an instrument on the Web it is right.

David McLaughlin: Really, I think that some people will always believe that if the computer tells you it is so, it must be so. For these people, I agree that the problem is very much an educational issue. It is also an issue of the quality of the software or instrumentation and its use. We attempt to make all of the techniques in the walkup lab environment quite robust.

Bridget Carragher: But not kindergartners. They do not believe anymore. They live on the Web and they don't trust any of it.

Clint Potter: I guess an issue is people using software now that they would never have been able to use before because it wasn't on the PC, and that they know about these mistakes and have learned about these mistakes, so maybe it is just an education issue.

Bridget Carragher: It is an education issue, yes.

12

Chemical Data in the "Internet Age"

W. Gary Mallard
National Institute of Standards and Technology

INTRODUCTION

It is difficult to determine whether discussion of the Internet as a force shaping the way we work is growing faster than the growth of the Internet itself. However, within chemistry and chemical engineering the use of the Internet as a resource for communication is exploding. Scientific publication on the Internet is just beginning. The use of the Internet as an information source in science is also still in its infancy. This paper discusses the changes that are driving the growing use of the Internet and what needs to be done to ensure that the new resources emerging fulfill the needs of the chemical community. Three factors can be identified as the primary drivers:

1. *Reduction in traditional data resources*. The loss of funding for a number of activities that provided information to chemists—cuts in library budgets, reductions in central research laboratories by industries, changing funding priorities at federal agencies—have all led to a reduction in the methods for finding needed data. Many libraries have had to cut out information specialists just when the increasing costs of journals have forced users to spend more time finding data. It is no longer possible in a number of large chemical companies to call on a department that specializes in physical property measurement and estimation. It is difficult to find funding for detailed critical evaluation projects, especially in the area of thermodynamics, thermophysical properties, or kinetics.
2. *Demand for faster access to data*. The need to obtain more information faster is not new, but the ability of computer databases to supply that information in new ways has driven a desire for ever more information. There is a growing sense that information should be available instantly, even if the real need for it is far more long term. In addition, the use of new drag discovery techniques and the development of substructure searching have also fueled demands for more information about a larger set of compounds.
3. *Increase in need for data for modeling and simulation*. The use of modeling has dramatically increased the need for data and, as is discussed below, has changed the nature of the data needed. Simulation of combustion, the atmosphere, urban air-sheds, and chemical reactors has required extensive data on kinetics, transport properties, and photophysics.

DATA NEEDS

The type of data needed in chemistry is changing. The traditional data requirements were for limited sets of data that were used to create correlations, to provide estimates, to test theories. This was a “retail” version of data usage. In industry, government, and academia, this work was typically done by individuals who had a strong background in the underlying physical principles embodied in the data. Errors in transcription were clear, and bad data usually stood out because data were used typically in sets and plotted against other related data. The data correlations were often extended to domains where measurement was either difficult or expensive. Predictions were made from the correlations but again, the background of the practitioners was such that the fundamental physical principles and "reasonableness" of the data were uppermost in their minds. The errors were mostly well appreciated, because the underlying science was closely coupled to the data analysis. The use of the resulting data was related to the confidence that the data were correct or at least that a firm understanding of the bounds of the uncertainty existed.

The use of modeling and simulation has placed new demands on data resources. These result in part from the different and often more complex systems that are being modeled, but also in part from the new requirements for complete data sets. The need for completeness comes about from the very nature of modern modeling programs, which take all aspects of the physics and chemistry into account—at least in principle. Since all physical and chemical processes are included, it is necessary to have data for the parameters that are used in describing the individual subprocesses of the model: diffusion coefficients, heat capacities, heats of formation, rates of reaction, and so on. Because it is essential that some value be placed in the model, there is a need to supply values for parameters for which there are little or no experimental data. This has given rise to a host of estimations and a greater need to determine the role of uncertainty in the modeling process.

For many of the unknown parameters, it is possible to show that any physically reasonable value will be acceptable since the underlying process is not a determinative of the outcome of the model. Thus, if one is in need of a diffusion coefficient for a radical, one can take the limits of the H atom (for which there are experimental data) and some molecule with a molecular weight twice that of the radical. Barring very unusual effects of polarity, the actual value of the diffusion constant will be in that range. By looking at the effect of the high and low values, it becomes possible to set limits on how much of an effect the high level of uncertainty will have on the final result. However, if the same calculation is to be applied to an ion, a completely different set of approximations must be used. The number and scope of the processes modeled in a modern simulation are so large that it is unlikely that anyone has the scientific background to ensure that all of the estimates are "reasonable." This is especially true since the definition of reasonable is a strong function of the problem: what is a small effect in one system may be large in another because of the difference in the process controlling the outcome of the model. For the most part we do not have modeling code that determines the "reasonableness" of the values used as input, nor do we have data resources that can provide physical limits for otherwise unknown data.

TYPES OF DATA RESOURCES

To satisfy the needs discussed above will require changes in the way that data resources are managed. Three broad categories of data resources are discussed to illustrate the problems in meeting these needs.

1. *Archive*. The archive is a set of numeric data of specific properties for specific chemical compounds with full literature references. The data should be clearly identified as to property (heat of

combustion per mole or per gram), including phase (liquid, solid, gas, amorphous), conditions (pressure, temperature, etc.), experimental technique, and ancillary data used to derive the property. Chemical compounds should be identified by structures, Chemical Abstracts Service registry numbers or Beilstein numbers, formulas, and names—including synonyms, common names, and trade names.

In addition, the archive should have removed any obvious errors in data transcription in the original text and adjusted the data for changes in ancillary information (for example, the definition of the calorie, or changes in the heat of formation of a by-product in the reaction).

Wherever possible, automatic comparisons should be made to further detect errors. This may even extend to automatic comparison of the data with estimation programs. The obvious errors revealed by automatic checking should be corrected. However, archive data are not presumed to have been examined in detail as to their accuracy. The uncertainty assigned to each data element in the archive is presumed to be the value assigned by the original author. The archive is not presumed to have extended this definition.

While this represents an ideal minimum, it is never realized fully.

2. *Review*. The review is expected to meet all of the requirements of the archive, but also to have been examined by a qualified scientist. Where appropriate, an attempt must have been made to reconcile data from different experimental methods, as well as from estimations and from high-quality calculations if they exist. Specific experimental and computational results should have been merged to provide an uncertainty assignment that reflects the range of values in which consensus scientific judgment expects the value to fall. For the common case where only a single experimental determination is available, it may be necessary to examine that datum in light of other related compounds. In many cases the experimental data can only be compared to estimated values.

3. *Critical evaluation*. Critically evaluated data should meet all of the criteria set for the review and archive data elements, but the evaluation should also place the data in the context of other related data. Thus, to evaluate the data for reaction of the OH radical with butane critically, it is essential to examine the data not only for the reaction of OH + butane, but also for OH + propane, OH + pentane, and more broadly OH + hydrocarbons. To do such a critical evaluation clearly requires a thorough review to have been made of each of the components. For some experimental data it is possible to use thermodynamic arguments to ensure the overall consistency of the data. Using the kinetics example above, if an independent measure of the free energy of reaction and the reverse rate constant are available, then there is a constraint on the forward rate. The evaluation must then examine the quality of these additional components also. As might be expected, the number of data sets that can be regarded as critically reviewed is very small. There will always be significant fractions of the data that cannot be critically reviewed owing to lack of experiments.

In the "retail" model of data usage, the difference between these types of data resources is not as important as it is when the "user" of the data is a modeling program. Even when there is direct personal use by a scientist, a lack of specific technical background to which the data relate may cause many of the same problems as would occur with direct computer usage. In both of these cases, the absence of an informed user can cause serious problems.

Only when there has been a critical evaluation with a clear indication of the uncertainty of the values reported can data be used in a fully automated fashion. Even in this case there is an obligation by the user to respect the uncertainty values and to assess how they affect the final output of the model. For complex models with high levels of uncertainty in a number of critical parameters, the computational cost of such an assessment can be high, but the data resource has provided the information needed to solve the problem.

Given the limits on critical evaluation, it is fortunate that for many problems a set of values that are of only "review" quality will suffice. In this case there is a single value for the parameter and a single uncertainty. The overall quality cannot be assumed to be as high, but in many cases it is sufficient. Again, the model must make use of the uncertainty.

The use of archive data in automatic systems is problematic. Often there are multiple values for a single parameter, and the reported uncertainties do not encompass all the data. In other cases the archive will contain data that upon examination will be viewed as inaccurate. There is no simple automatic mode to deal with the range of problems that will be encountered here, although it can be sufficient to take data with multiple values and use the average with an appropriately large uncertainty. The success of such an approach will depend on the problem.

PROBLEMS IN PROVIDING DATA

The problem of providing good data for modern computer models can be broken down into three broad classes:

1. *Incomplete data sets*. As noted above, a model must have data for each physical property, rate constant, and thermodynamic parameter within that model. There are broad classes of data for which there are simply no experimental data. For example, there are very few data for any radical diffusion constants, entropy, or heat capacity. Good estimates can be made, but experimental data are very scarce. In many cases the use of values essentially equal to zero will cause the model to fail, so some physically reasonable data must be included. For many properties this requirement is addressed by a combination of data for related properties plus models. This is the approach taken by the Design Institute for Physical Property Data (DIPPR) Committee of the American Institute of Chemical Engineers (AIChE), which has created an extensive set of data for use by the chemical process industry. A similar approach has been taken by NASA in stratospheric modeling. In general, this method has not been used outside very specialized areas.

2. *Uncertain data*. The reported uncertainty in most data is, at best, the experimental variation found. It is rare for any attempt to be made to assess systematic uncertainty in a measurement. Data in older literature can often be "rescued" by a better understanding of some systematic error that was not appreciated by the original investigator. While it is tempting to ignore data with identifiable errors, if the error is systematic and can be corrected for, the data may be useful. In many cases they are the only data available for the property of that compound. By eliminating the systematic error and at the same time recognizing that the correction probably carries its own uncertainty, it is possible to provide data that are useful. This is an important role for reviews of data.

One problem in some data sets is the uncertainty of the chemical identity. As noted above, there is a need for absolute chemical identity. This need is often not met in smaller data collections.

In addition, the failure to account for changes in auxiliary data can lead to serious errors in the reported data that are not present in the experiment. As an example, much of the data on fluorine-containing compounds in the literature before 1970 used an incorrect value for the enthalpy of formation of CF_4 to derive the enthalpy of formation for other compounds. Simply providing the original enthalpy of formation from the literature will give a very incorrect sense of the state of the data. In this case providing the correct number from the original paper is not sufficient.

3. *Errors in data compilation*. Extracting the literature into electronic format is in itself an error-prone project: digits are inverted, signs are ignored, states are not defined. The goal of the compilation into electronic format is to add value to the data; these kinds of errors, for the most part, do the reverse.

DATA RESOURCES CURRENTLY AVAILABLE

Three of the relatively few extensive electronic databases generally available today are discussed. Currently, there are no resources that will meet all the needs pointed out above. Examination of the resources that are available illustrates both the strengths of these resources and the unmet needs.

All resources are available via Internet connections, and one is free. The data sets are the Beilstein database, currently owned by Elsevier; the DIPPR 801 project of the AIChE, currently at Brigham Young University, but during most of its development at Pennsylvania State University; and the National Institute of Standards and Technology (NIST) Chemistry WebBook.[1] These are very different efforts in size and history. Beilstein goes back into the 19th century and until recently was partially funded by the German government. The DIPPR project started in 1980 as a response to the need for high-quality data in the chemical process industry, and is funded by a consortium of members from industry and government. The NIST-funded Chemistry WebBook has been in existence for only 3 years and was developed specifically to deliver data over the Internet.

Beilstein Database

Beilstein is by far the largest of the three databases. Table 12.1 shows several of the types of queries to the Beilstein database and the number of molecules in the hit set. The list illustrates the origin of the Beilstein database as a database of organic chemistry. The properties useful in organic chemistry are well covered; for example, the fraction of molecules for which there are reaction data is quite high. Table 12.1 also gives some sense of the sheer scope of the database: there are more than 7 million distinct chemical species. Beilstein differentiates between optical isomers if there are data on the distinct isomers, so the number is higher than it would otherwise be. This again reflects the organic chemistry origins of the database. For physical property data—for example, the enthalpy of formation—the amount of data is not all that great. However, this may well represent all of the enthalpy of formation data for organic compounds.

The Beilstein database is strictly archival; no attempt is made to do any evaluation and the review literature is not covered. The database is excellent in terms of its chemical identity. In fact, of the three databases discussed here, it is by far the best. Because of its size it will have the most errors—no matter how carefully a database is created, as it becomes larger the number of errors grows.

There are some problems in Beilstein that are unique. As an example, two enthalpy-of-formation values from the same reference are given for 3-oxa-tricyclo[3.2.1.02.4]octane (Figure 12.1) as 53,900 J/mol and 98,000 J/mol. There is no indication that the first value is the enthalpy of formation for the gas phase and the second the value for the liquid phase. In order to determine what the values refer to, it is necessary either to observe that the enthalpy of vaporization is the difference between these values, or to go to the original paper. Given completely electronic access to the data, the information about the enthalpy of vaporization may not have been accessed. In addition, both values are sign reversed. The problem of sign reversal is fairly common in the Beilstein electronic database and probably arises from the convention in much of the thermochemical data literature of giving a table of values as $-\Delta H_{for}$ rather than showing the sign in the table.

[1] For more information on the Beilstein database, see <http://www.beilstein.com/products/xfire/>. A subset of the DIPPR database can be accessed at <http://dippr.byu.edu/>. The NIST WebBook can be found at <http://webbook.nist.gov>.

TABLE 12.1 Molecules in the Beilstein Database, by Query Type

Type of Data As Defined by Beilstein Query	Number of Molecules
Enthalpy of formation (H_{for})	7300
Entropy data (all)	3655
Heat capacity (c_p)	1946
Boiling point at pressure (bp.p)	642000
Viscosity (bulk, kinematic, dynamic)	5000
NMR spectra for ^1H	32000
Reaction (all)	4,600,000
Total number of chemical species	7,300,000

Another example, which in many respects represents a more serious problem, is the entry for hexamethyldisiloxane. This gives two values for its enthalpy of formation, -815,800 and 815,400 J/mol, which are reported to be measured at 25 and 298.2 °C, respectively. These data are referenced to the same authors within a 3-year period. Again, there are a number of mistakes obvious to the expert, and in this case even to the non-expert, but the information is not usable in a system seeking to obtain high-quality information automatically.

The final example of problems in Beilstein is one that is inherent in the way that data are taken for the database. In this case the literature is cited correctly: there are two experimental determinations of the enthalpy of formation of 1,2-difluoro-1,1,2,2-tetrachloroethane: -891.788 kJ/mol and -928 kJ/mol from 1954 and 1982, respectively. However, both determinations are based on different enthalpy of formation data for CF_4. When the experimental values are corrected for the currently accepted CODATA value for the enthalpy of formation data for CF_4, the results are -925.5 kJ/mol and -937 kJ/mol. The agreement between the two experiments is quite good, but without the adjustment this would not be seen.

The first two examples are criticisms of the data quality in Beilstein. The last is not, but it is a warning to anyone who uses the data to proceed cautiously. It is not clear to what extent the problems illustrated above are general in the database. Used as a resource to find available literature data it is invaluable, but it cannot be used uncritically as a direct resource for numerical values.

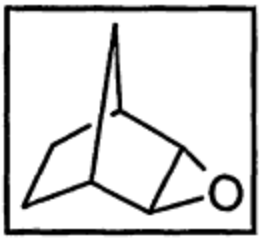


Figure 12.1
3-oxa-tricyclooctane.

DIPPR Project

The DIPPR Project 801 of the AIChE was designed from the start to provide complete physical property data for the molecules in the database. For each of the more than 1,700 molecules, the available experimental data listed in Box 12.1 are collected and evaluated. For properties for which experimental data do not exist, the DIPPR project estimates the data and, if necessary, their temperature dependence.

The data in the DIPPR database is ideal for use in modeling. It is reviewed, and a recommended value or equation as a function of temperature is given for each property for each molecule. There are some problems in that the estimations are not as clearly indicated as might be desired, but this is being corrected. The uncertainty values are all expressed in terms of ranges and not as absolute values. For some properties this is reasonable, but for thermochemical data, it is essential to know the uncertainty directly.

The DIPPR database illustrates the difficulty of providing high-quality complete data. The level and quality of the effort in the DIPPR project have been very high and the project has been going on for more than 17 years with fairly extensive resources, and yet only 1,700 compounds (all stable species) have been added to the database. For the molecules and properties in the database, DIPPR is usually a first choice.

BOX 12.1 DATA EVALUATED IN THE DIPPR PROJECT

Temperature-independent Data

- Critical Temperature (K)
- Heat of Fusion at Melt Pt (J/kmol)
- Critical Pressure (Pa)
- Standard Net Heat of Combustion (J/kmol)
- Critical Volume (m^3/kmol)
- Acentric Factor (unitless)
- Critical Compress Factor (unitless)
- Radius of Gyration (m)
- Melting Point (K)
- Solubility Parameter ($(J/m^3)^{0.5}$)
- Triple Point Temperature (K)
- Dipole Moment (debye)
- Triple Point Pressure (Pa)
- van der Waals Volume (m^3/kmol)
- Normal Boiling Point (K)
- van der Waals Area (m^2)
- Liquid Molar Volume (m^3/kmol)
- Refractive Index (unitless)
- Ideal Gas Heat of Formation (J/kmol × K)
- Flash Point (K)
- IG Gibbs of Formation (J/kmol)
- Lower/Upper Flammability Limit Temperature (K)
- IG Absolute Entropy (J/kmol × K)
- Autoignition Temperature (K)

Temperature-dependent Properties.

- Solid/Liquid Density
- Solid/Liquid Vapor Pressure
- Heat of Vaporization
- Solid/Liquid Heat Capacity
- Ideal Gas Heat Capacity
- Second Virial Coefficient
- Liquid/Vapor Viscosity
- Solid/Liquid/Vapor Thermal Conductivity
- Surface Tension

The NIST Chemistry WebBook

The WebBook is a hybrid database. It is not complete (in the DIPPR sense of having all properties for all molecules in the database), yet is not just an archive (in the sense that reviews from the literature are included, as are reviews and evaluations done just for the WebBook). Table 12.2 gives some sense of the data in the three most recent releases of the WebBook.

Phase-change and thermochemical data have multiple data types. Data are both single points and temperature-dependent equations. As can be seen, the data for any given molecule are likely to be incomplete. The kinds of data are greater than in the case of the DIPPR database (not all data types are shown in Table 12.2) but are not currently as extensive as in the Beilstein database. DIPPR does have more extensive coverage of transport properties.

The WebBook makes use of the review literature in order to allow for later corrections arising from changes in auxiliary values—corrections from the authors and evaluations of the relative uncertainty of the various experimental methods to be incorporated. However, a large portion of the WebBook's data is archival, even if corrected for these changes.

The existence of a large set of data with extensive indexing has allowed the first steps toward evaluation to be made. A list of enthalpies of formation for carbonyl compounds from the WebBook (Figure 12.2) serves as an example of the kinds of problems that are revealed in the data.

In each of these cases, the data are as they appear in the literature and are fully corrected for auxiliary data. The first value for each molecule is from a single author, the remaining values from a number of authors. A pattern of higher stability appears to be measured by one author. How is this to be evaluated? The issue is that there is no simple way to evaluate this son of problem. The level of uncertainty here, which is on the border of what would be resolvable using the best of quantum calculations, may be significant in some applications. Moreover, there are other cases where this author has published values for which there are no other data. How is this to be evaluated as well? While these questions can be answered by expert evaluation of the experimental methods, the differences between experimental methods and high-level calculations, this level of evaluation cannot be done automatically,

TABLE 12.2. Attributes of the NIST Chemistry WebBook

	Version		
Data Type	3	4	5
Gas-Phase Ion-Energetics Data	14,200	14,300	14,300
Gas-Phase Thermochemical Data	2,800	5,800	6,100
Condensed-Phase Thermochemical Data	4,600	5,300	5,500
Phase-Change Thermochemical Data	8,800	9,400	9,500
Reaction Thermodynamic Data	7,400	8,700	9,400
IR Spectra	5,200	5,200	5,200
Mass Spectra	8,300	10,600	10,600
Fluid Property Data Sets	13	16	16
Vibrational/Electronic Spectra & Energy Levels	—	2,600	3,300
Spectroscopic Constants of Diatomic Molecules	—	600	600
Total Species with Data	27,300	31,600	32,400
Release Date	Aug-97	Mar-98	Nov-98

so what data are to be used by a modeling code accessing this data? Averaged values may well be skewed because there is a systematic error in one of the measurements. The answer lies in part with the level of effort indicated by the DIPPR data project. If these data are important, then the effort needs to be made. In part the answer lies in some assessment of how accurate data need to be. The uncertainty given above is still small compared to that for many enthalpy-of-formation values. These may be "good enough," and an automatic average with high uncertainty will be all that is needed. However, the degree to which one can model, predict, and control a system sets many of the economic costs for a system. In general the cost of uncertainty in chemical operations is strongly nonlinear, and small improvements in the prediction and control can yield large improvements in costs.

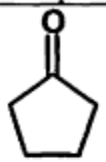
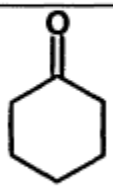
Compound	Enthalpy of Formation
	-197.4 ± 1.3 kJ/mol -194.8 ± 1.7 kJ/mol -193.0 ± 1.8 kJ/mol
	-231.1 ± 0.9 kJ/mol -225.7 kJ/mol -227.7 ± 1.9 kJ/mol -226.8 kJ/mol
	-218.5 ± 0.6 kJ/mol -217.1 ± 0.5 kJ/mol -217.5 ± 0.7 kJ/mol -216.4 kJ/mol

Figure 12.2.
Enthalpies of formation for selected carbonyl compounds from the NIST Chemistry WebBook.

USER DEMAND FOR DATA

One point needs to be made: the need for even less-than-perfect data is very large. The usage of the NIST WebBook in the third release is given in Figure 12.3. Access by a wide variety of users is running at over 5,000 hits per week, with between 40 and 50 percent of the users returning in any given week. In the 220 days that this edition was out, over 120,000 distinct Internet addresses (IP addresses) used the WebBook. Usage clearly tracks the academic calendar of the Northern Hemisphere. Usage over Christmas and the New Year is very low, but even so is more than 1,000 hits per week. Summer usage is lower than at other times, but still is more than 3,000 hits per week.

Comparable data were not available to the author for either the DIPPR or the Beilstein databases. Both have some charges associated with them, whereas the WebBook is currently free.

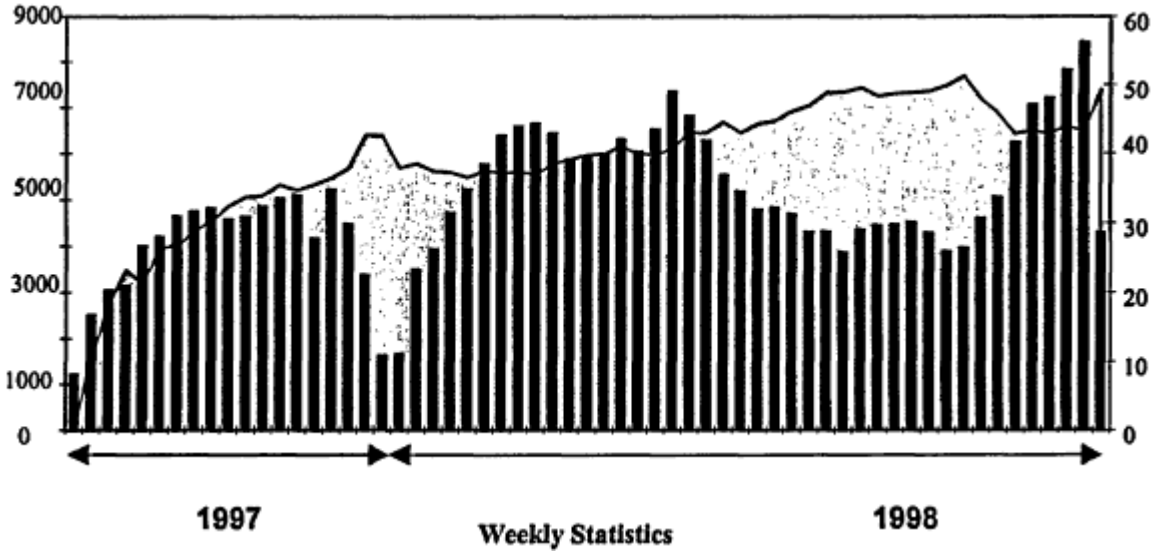


Figure 12.3
Usage of NIST WebBook in the third release. Number of users (left, bars) and percentage of returning users (right).

WHAT IS NEEDED

The demand for data is clearly large, as can be seen by the usage for resources such as the WebBook. The WebBook, DIPPR, and Beilstein are currently not equipped to handle direct requests from modeling programs. The need for communication standards among modeling programs and the databases that they rely on has not been touched on here, but the absence of agreed-upon query structure that would make it reasonable for a database provider to support direct access by modeling programs is not the real limiting step for future use of the data. The limiting factor will be the lack of resources to produce high-quality evaluated data that can be used with confidence.

Building data collections, reviewing and evaluating the data, and distributing the resulting information are not free, but if done well, yield a very large return on investment. A high-quality data collection can save individual researchers thousands of hours collectively. In addition, the evaluation can reduce the uncertainty in data with the corresponding economic benefit of higher-quality modeling and prediction. It is essential that the data provided be subjected to high levels of quality control, that the uncertainty be evident, and that the modeling programs make use of the uncertainty.

Venues such as the WebBook can make some of these data more readily available. The WebBook has been actively seeking researchers who have developed extensive sets of data that they have at least reviewed, if not critically evaluated. These sets are being made available on the WebBook with full credit going to the reviewer. In many cases the archival journals are not interested in this work; often even if it is published, it is lost. Much of the older literature is also being actively evaluated; the authors and copyright holders are being sought for permission to add the data to the WebBook. At a minimum this kind of effort will bring more attention to data that are often valuable but difficult to find. It is hoped that bringing this kind of data to a wider audience will stimulate more such reviews and realization of the need to update some of the older reviews. The ultimate goal is to have electronically accessible all

numerical data for which there are good experimental data and to extend the data with high-quality predictions of known uncertainty where there are no experimental data.

DISCUSSION

Allen Bard, University of Texas: Who pays for this now? The government is paying for this and will continue to do so?

Gary Mallard: At least for the foreseeable future the answer is yes. If you go to the WebBook you will find a little blurb there that says, "NIST reserves the right to charge for this in the future." We have tossed around the idea of having of a $50-a-year usage fee. I don't know whether we are ever going to do that. I think that the usage is general enough that we can justify it as a reasonable use of the taxpayers' money, but whether that continues in eras of tightening budgets is a question I cannot answer. But now we are doing it with internal funds.

Sam Kounaves, Tufts University: Were these actual unique users, or return users, or just hits?

Gary Mallard: No, they are not hits. That is a very deceptive way of determining use. These are unique IP addresses, although we don't know the people behind them. In fact, we know there are more users than this because many commercial suppliers come through single addresses. The big companies all come through gateways. So, we don't know how many total users there are. We just know that there are 110,000 distinct IP addresses, and of those, on any given week about 45 to 50 percent of them have been there before.

Sam Kounaves: Have you ever considered advertising to have people come back on this Web like having an instrument company and some chemical company to support this sort of stuff?

Gary Mallard: It is a little tricky if you are the government. They think they pay taxes, and they do.

David Smith, DuPont: I have a suggestion for some of my friends in the audience. For those of you who are teaching thermodynamics, perhaps it would be a good exercise to have your students calculate the thermodynamic consistency of these articles for some of the compounds that you are interested in. A couple of hundred a year would probably be a worthwhile effort.

Gary Mallard: In the very early days when we had fewer users, there was a sudden spike. Whenever we see a sudden spike we wonder what is going on, and this spike was coming from someplace in Canada. It is pretty easy to trace these things back on the Web, and it turned out that in fact somebody had done just that—had told people to look up a certain number of compounds on the Web and get those data and put them into a report. I don't remember the details anymore. So for about 4 or 5 days, we had a lot of organic chemistry students at one university using this site, and at the time it made a spike. Today you wouldn't even see it.

David Dixon, Pacific Northwest National Laboratory: What is the structure of the database, and what were the manpower requirements to do the electronic version as compared to the manpower already working on the standard printed versions from which much of the electronic versions are derived?

Gary Mallard: The resources needed to do the data evaluation and collection do not really change, whether you are putting data out in an electronic format or in printed format, and so that number is constant and really represents whatever we can find at NIST. In a lot of cases, we have been taking data that we have had for a long time in printed format and just putting it in electronic format. So that is relatively inexpensive and fairly cost effective because it really just takes somebody with good data entry skills to put it in a spreadsheet, and we do a little processing on it.

One person works full time on the database itself. There is a Java applet that displays spectra that can be enlarged. The address is <webbook.nist.gov> and I would urge all of you to go there and try it. A lot of what that one person does is deal with issues on the Web, and we work very hard to make sure that if we display a Greek character it displays on a Sun system, on a Macintosh, and on Windows, and that is a non-trivial exercise. A lot of time is spent in making sure that this is a high-quality product that looks the same on everybody's browser, that the Java applet works the same on everybody's browser, and none of that is easy. So, in that sense there is one person devoted full time just to keeping this thing up on the Web.

The structure of the database is basically a file that is indexed under C-Tree, an available piece of software that is all C code and has been known to compile on more platforms than anything else known to man. We store everything as ASCII, because we feel that when you deal with things like the number of significant figures, you would like to capture that information and not lose it, and so we have some fairly sophisticated algorithms for looking at the number of significant figures. Also, when we convert from kilocalories, which is perhaps what the data was originally entered in, into kilojoules we try to retain all of that information and not have six significant figures of zeroes which aren't significant, but the structure of the database itself is basically just ASCII.

Jack Kay, Drexel University: Are the JANAF thermochemical tables included in this database?

Gary Mallard: Yes, not as tables but as equations in the new format, the Shomate coefficients where there is a 1/T term in the last term.

Robert Cordova, Elf Atochem: I was wondering about the relationship between this and database 19 for structures and properties.

Gary Mallard: What I didn't show you on that form was that in the very beginning we actually had some estimates in the database, and those estimates came out of some of the kind of code that was in the database for structures and properties. We removed all of the estimates. The WebBook really is a kind of evolutionary extension of the structures and properties database. There are a lot more thermochemical data in the WebBook, but the estimation tool that was a part of structures and properties is not there. I think eventually we will put it back, but we just haven't had the resources to do it yet.

13

The Digital Library: An Integrated System for Scholarly Communication

Richard E. Lucier
University of California

Scientific journals have served scientists well for many decades. They have provided a viable means for scientists to communicate their findings to their peers and have served as well as an archival record of scientific progress. Now, however, we are seeing the beginnings of a significant evolution away from what we know as the traditional scientific research journal.

This article is divided into three parts. The first part provides context by outlining what is driving some of the issues that are discussed. Second, the notion of what could be meant by a digital library is discussed. In doing so, I describe what we are doing in California—the California Digital Library—as well as the digital library in general. What we are doing in California is an example of what might occur in other places, so that it is useful and instructive to talk about some of the specifics of that project. Finally, I discuss alternative forms of scholarly communication, i.e., alternatives to the tradiltional research journals.

CRISIS IN SCHOLARLY COMMUNICATION

We have been hearing for the past decade that academic research libraries are in crisis. The fact of the matter is that these libraries, as we know them, including the services that we have come to expect from them, are simply no longer sustainable in their current form. Projected costs, for both the acquisition and the storage of information, are significantly higher than the universities can possibly sustain. While current data indicates that university fees are going up 8 percent, library acquisition costs (in the sciences) have been rising at a significantly higher rate, namely 15 to 20 percent annually for the past several years.

It is convenient to blame publishers for these increases and this crisis. Certainly some publishers, particularly commercial publishers, are part of the cause of these difficulties. In many respects, however, the productivity of scientists is a more significant causal factor. Scientists are producing more and more information, e.g., the American Chemical Society is publishing 10 percent more pages each year, and many other publishers report annual increases of 10 to 20 percent. If this increase in the rate at which information is published continues, and that seems likely, university libraries will be unable to

provide scientists with access to that information if the traditional process of scientific communication is maintained.

At the same time, we are now at an early evolutionary stage in the use of digital technology in scholarly communication. What will happen as this evolution continues—how scholars and scientists will eventually integrate digital information technologies into their work—is not yet clear. A number of issues like cost, ease of use, and academic culture are going to have a major impact on the future application of digital technology in this area. There are many who believe that the application of digital technologies to scholarly communication is as revolutionary as the use of the printing press, and indeed I think that is the case. But it is going to take several years to see how this evolves and what the implications are.

SOLUTIONS AND STRATEGIES

A number of universities are exploring optimal strategies for dealing with this transition in the management of scholarly information and optimizing the opportunities presented by digital technologies. The University of California (UC) has recently completed a 4-year planning process examining all of these issues. The following conclusions form the basis for strategic action:

- Comprehensive access to information is going to replace comprehensive ownership of information. Remember when you expected to be able to get your favorite journal from your university or corporate library? In the future your library will provide you access to that information in some reasonable period of time.
- Solutions will unfold organically. A traditional plan is not desirable.
- The digital library is an agent of change.
- UC should build one digital library.

THE CALIFORNIA DIGITAL LIBRARY

As part of our planning for libraries and scholarly information at the University of California, we developed a shared vision of the library appropriate for the university: *A World-class Research Library for the 21st Century Consisting of Complementary Paper and Digital Libraries Comprising a University-wide Knowledge Network With Services Delivered at the Point of Need*.

The digital component of this library has been named the California Digital Library (CDL), to reflect its potential to serve all of California, not just UC. Created in October 1997 by the University of California Board of Regents and the president of the university, the CDL will open its “digital doors” in January 1999.

The digital library can be viewed as an integrated system for the management of scholarly information. This moves the library beyond its traditional roles of storage, preservation, and access to an active player in scholarly communication through support for alternative forms of publication, which exploit digital technology.

In this context, the digital library will have the following components:

- High-quality digital resources;
- Consistent network interface;
- Distributed services integrated with digital resources at the point of service;

- Alternative forms of publication and the digital dissemination of scholarship; and
- Underlying business and economic models that are sustainable.

It is important to note that the digital library is not going to replace the paper library for the foreseeable future. Rather, the two are going to coexist, complementing one another. How we implement this complementarity is going to be critical to the continued smooth functioning of the scholarly enterprise.

Content is the heart of all libraries, and that is true for the CDL, as well. Already, we see that there are many different kinds of digital collections. One is the traditional published journal literature in digital format. While this is important at these early stages of building a digital library, it will likely become less important in time. Even at the outset, we have a second kind of content, "digital at birth" content, which has always been in digital form only. Third is primary source data. For libraries, that includes special collections, but our faculty also have a lot of primary source data that we are trying to build into our collections. A fourth important area of content for digital libraries is museum collections. Stronger relationships between libraries and museums are developing in the digital era. Last is what we refer to as alternative forms of scholarly communication. Over the next 15 years the division between journals on the one hand and these alternative communication forms on the other is expected to change significantly. By the year 2015, we expect that less than half of the kind of information that we will be providing in the CDL will be in the form of traditional journals in digital format.

It is important to review the changes taking place with respect to the acquisition of digital content. In many instances, libraries are not buying digital journals and databases; instead, they are licensing them. This change has several important implications. Licenses have very different terms and conditions, depending on the publisher. Of critical importance to the library and user communities is the notion of perpetual access. The approach of many publishers has been to license information for the year in which you pay for it. Thus, if you paid for a 1998 subscription in 1998 but not in 1999, you would lose access to 1998 content in 1999. This is in marked contrast to the past. Traditionally, if the library bought a paper journal, it always had that journal. That is not the case with electronic material. These and many other issues surrounding licensing mark a new relationship between publishers and libraries, one full of challenges and opportunities, requiting great vigilance and care on the part of the library community if it is to ensure access for research and education.

Licensing has allowed us to develop large collections of digital material in a short time frame. When we open the California Digital Library, we will have more than 3,000 electronic journals in scientific and technical areas. That, along with the other databases, is a significant amount of content. Licensing has also provided a platform for cost control. By forming and joining large consortia or groups of libraries to license electronic material, we are able to leverage our collective buying power. This has allowed us to buy more materials than we would have, if we had individually tried to license this material. It also sets the stage for how we are going to work with the publishing community in the future. One of the first digital journal licenses we signed at UC was with the American Chemical Society. It is important to note that ACS has been willing to develop a model that is more beneficial to users than the models developed by many other publishers, particularly commercial ones. We have appreciated ACS's willingness to listen and respond to our concerns.

Licensing is a useful strategy to aid in the transition from paper to digital journals, in moving from an ownership model to a service and access model. In 10 years, however, its various components may not be as significant as they are now. Hopefully, licensing at some level will evolve into a standard operating procedure, with much less emphasis placed on negotiating individual terms with each content provider.

ALTERNATIVE FORMS OF SCHOLARLY COMMUNICATION

We are currently using technology in its first phase of adoption: modernization; that is, we are replacing traditional paper journals with traditional digital journals. What we really need to do is to lay the foundation for a future in which how scholars communicate and access the results of research around the world will truly be transformed. We need to invest a significant component of our resources into developing innovations that will facilitate this transformation. It is my belief that the digital library provides the appropriate infrastructure to develop and leverage these innovations, to make collective investments across universities to do self-publishing, digital publishing, and to directly compete with the existing model of scholarly communication.

There are a number of activities to pursue in developing alternative forms of scholarly communication. One is to develop prototype projects with faculty, based on their needs for better ways to disseminate and access information. The most important innovations will come from the scholarly community itself, not from administrators or librarians. A second is policy development. In this area, it is important to examine current copyright policies and behaviors such as the assignment of rights to publishers. Third, the development of new forms of scholarly communication is best pursued in concert with colleagues on national and international levels.

There are a number of potential scenarios for alternative forms of scholarly communication that have been put forward in the last few years. One is called NEAR, the National Electronic Article Repository. What a university provost has proposed is that authors would retain certain rights when they publish a journal article. The rights would permit, within 90 days of the appearance of the paper publication, placing the article on a national server, which is a depository for scholarly articles. Everybody would then have free, perpetual access to the article.

A second scenario, put forward by the Association of American University Presidents, recommends decoupling certification from publication. This is based on the belief that it is the coupling of the promotion and tenure process with publication that is the cause of current financial problems in the publication of scholarly materials. In this scenario, universities would pay professional societies to review the work of their faculty and to certify the work for promotion and tenure purposes. The universities would then place the work on servers so that it would be available to the external community at no or low cost. Only a small portion of this material would then make its way through the regular publication process. So, instead of publications continually increasing, actual paper publications (expensive publications) would decrease, to be replaced by electronic publications on readily accessed servers.

A third scenario calls for universities to assume responsibility for publishing the work of their faculty. A fourth identifies the notion of peer-reviewed servers as the viable alternative. What we have seen in the physics community at Los Alamos, for example, is the establishment of a physics preprint server where not only the physics community, but also the mathematics community as well as others place articles prior to publication. Adding peer review to that preprint process would ensure technical merit, and it would not be necessary to go through the entire publication process.

What is absolutely key to further progress in identifying and implementing sustainable alternatives for scholarly communication is the development of business models for all scenarios. While some of these ideas may sound desirable, their financial feasibility must be demonstrated. Replacement models are not necessarily less expensive, and appropriate due diligence must be rigorous.

A 1998 essay, "To Publish and Perish,"[1] recommends five actions as we move toward new alternatives:

- “Turn down the volume"; i.e., we need to concentrate more on the quality of publication rather than quantity in the promotion and tenure process.
- Librarians must be smarter shoppers, just as we are trying to be through consortia and licensing.
- Get a handle on copyright and property rights issues.
- Universities must invest in electronic forms of scholarly communication and should support new efforts that faculty are putting forward in this area.
- The decoupling of publication and faculty evaluation should be seriously investigated.

National efforts are under way, which should lead to some breakthroughs in the coming years. One is SPARC, the Scholarly Publishing and Academic Resources Coalition, originally developed by the Association of Research Libraries. It currently consists of more than 100 research libraries from around the country that have joined together for the following purposes:

- To create a more competitive marketplace,
- To reduce journal prices,
- To ensure fair use of electronic materials, and
- To apply new technologies to information creation and storage.

SPARC:

- Solicits high-quality, fairly priced publications and guarantees a subscription base;
- Provides start-up capital for new projects; and
- Generates support from important groups like our own faculty and administrators and provosts.

The American Chemical Society and the Royal Society of Chemistry are initial SPARC partners.

CONCLUSION

The challenges that we face in trying to make our way through this evolutionary change are very significant. They range from the political to the technical and financial. As we move toward the realization of digital libraries, basic research is absolutely critical in this area. There are no models; we are in uncharted territory, and we need the research community to inform the way. It is critical that faculty participate in this research. The solutions must reflect your "way of doing business."

DISCUSSION

Stephen Heller, National Institute of Standards and Technology: I have a couple of questions. First, are you working at all with Highwire Press, and do you have any comments about what they have been doing?

[1] "To Publish and Perish," based on a roundtable hosted by Johns Hopkins University and convened jointly by the Association of Research Libraries, the Association of American Universities, and the Pew Higher Education Roundtable, March 1998; published in *Policy Perspectives* , Knight Higher Education Collaborative, Philadelphia (1998).

Richard Lucier: Highwire Press is a very interesting operation. It has taken over the production for a number of societies that would not have had the money individually to invest in digital technology, and so has allowed those societies to remain competitive as we move into this digital environment. Highwire Press has done great work. We work with them in the sense that we pay for all of those publications. We have had discussions about more substantive cooperation, but nothing yet has come out between Stanford and UC on that.

Stephen Heller: The second question is, Should libraries be responsible for the actual archiving when a reasonable solution is found? Right now it is a sort of random process in which the publishers have decided to go into the new business of archiving, which is providing information on a long-term basis, not just selling a subscription and washing their hands of any responsibility to provide anything after the subscription expires.

Richard Lucier: I think it depends on the publisher. We need to have someone of repute take responsibility for archiving the world's knowledge base. I am very leery about saying that commercial entities should take that responsibility.

Commercial entities will take that responsibility only as long as it is profitable for them, and as information gets older it may no longer be profitable.

I don't know if it should be libraries. I don't know if it should be universities. I don't know if it should be the federal government or some other organization. I think there has to be a national strategy initially, and there will have to be an international strategy. No one—not the library, the community, the academic disciplines, societies, or publishers—has yet really tackled that problem very well.

A couple of years ago a commission came out with a report that rather scared everyone, and so no one has touched the issue in the last couple of years, and I hope the dialogue can continue again soon. But there are no good answers, and I am not sure who ought to do it, but I guess I would trust universities before I would trust commercial organizations.

I believe, and Lorrin might correct me, that even our license with the American Chemical Society only guarantees access for 5 years, and I am assuming that some of that literature is still important to you after it is 5 years old.

Stanley Sandler, University of Delaware: I guess I am overwhelmed with the amount of information that is available and being generated, particularly the number of journal pages and such. So, maybe this is more an appeal to my colleagues. I think there is an increasing difference between, let us say, the CPU and the LPU. We know that the CPU is increasing speed and power and likewise the rate at which we can generate experimental data. The LPU, the least publishable unit, I would say hasn't changed anywhere nearly as fast as the rate at which we generate simulation and experimental data. So consequently a paper today may contain about the same information as a paper years ago. Years ago it may have required years of intellectual effort. Maybe 2 years ago it required a month of intellectual effort. Maybe this year it requires a week of intellectual effort, and I submit that we as reviewers are not doing a careful enough job of keeping the LPU together with the CPU, and that is why we are overwhelmed with so much published literature.

Richard Lucier: I think that one of my frustrations in this uphill battle is that I cannot solve these problems. I am willing to support you in any way that you want, but it is the disciplines and the academic communities that have to come to some solution. That last point was one of the reasons that we are very concerned about making sure that we are providing only high-quality information within the digital

library, not just everything that is out there, and it is, also, when different groups like the American Association of Universities look at this decoupling process. It is a recognition that everything that is going into print probably ought not to go into print and that somewhere along the line we have to make some qualitative judgments. You have to make some qualitative judgments. I am not trying to impose anything.

Allen Bard, University of Texas: It seems to me that this idea of decoupling publication and certification is a game, because as everybody will know if you decouple it, you say, "Yes, we certify this as great work, but it is not worth publishing; we certify that other work, but it is worth publishing." Everybody will know that game.

Richard Lucier: I think there are other ways to look at it. One can say that we certify this work, but we don't need to publish it in its complete form in the way that we did in the past, and we might only publish in a formal publication a certain excerpt but maintain on file servers that would be a lot cheaper to do over time, the ability to be able to get access to that information.

I think what decoupling does is allow us to look at the publication process differently so that we can find a cost-effective way, and one that exploits technology in a way that makes the data more useful to you as well.

Tom Edgar, University of Texas: What is your business model as you are constructing it at the University of California? If you look down the road, say 10 to 15 years, do you see any changes in human resources needs for the collective libraries of the University of California system?

Richard Lucier: How many years into the future?

Tom Edgar: Ten to 20, let us say.

Richard Lucier: I cannot look 10 to 20. The most I can look is 5, and yes, we do see changes.

Tom Edgar: I guess the gradient is what I am interested in; is it positive or negative in terms of the number of people it is going to take to provide the California Digital Library services compared to the number you have today?

Richard Lucier: I think that what we have projected is that it is going to take probably an equal number of resources, but ones more focused on providing quality access to information than they currently are. What I can tell you with respect to saving money is that in the first year we can document that we have saved the campuses, in licensing costs alone, about $2.5 million for access to information. If they had gone about it separately and bought this information themselves, it would have cost them that much more.

The other thing that we are able to do is to provide access independent of location. It doesn't matter any longer if you are a chemist at Berkeley or if you are a chemist at Santa Cruz; you can get access to the same kind of information. We feel that it is really important for our faculty and students to be able to have that kind of access irrespective of their particular physical location.

Tom Edgar: The second question is one I will ask and then head for cover. You said that the physicists and mathematicians have agreed to go toward putting publications on Web servers. My impression is that the chemists have really not agreed to do that in the same way. I am curious. What are the differences between chemists and the other group that make chemists behave differently?

Richard Lucier: I think you could answer that question better than I, and I would be really interested in the answer.

Steven Heller: It is a cultural thing. Actually the story with the Los Alamos pre-print server is that they had been doing pre-print exchanges for decades before computers, and when computers came they just put the pre-prints on the computer.

Richard Lucier: Is that true with the mathematics community as well?

Steven Heller: Yes.

Richard Lucier: There is a new biological sciences server that you may or may not know about, a preprint server that has begun as well. Having spent most of my career in the biomedical sciences, I was very surprised about that because the exchange of pre-prints has not been traditional in that field, but they see what is happening in physics and mathematics and have moved to that.

Evelyn Goldfield, Wayne State University: First I would like to say something about the pre-print servers. One of the problems that chemists feel, at least the ones I have talked to—and I think this is a problem that you are going to see—is the question of peer review or multiple versions or error corrections, because from what I understand, things can go on to pre-print servers without any review at all. As a physicist friend of mine explained, "Oh, we will just correct it as we go along," which is fine if you are in that community and you know. But a student could easily be getting incorrect information, and I think there is a resistance on the part of a lot of people to risk that.

Steven Heller: That is not true. There is a link between the versions.

Evelyn Goldfield: I believe that many chemists are wary about non-journal Web-based publishing on account of quality control issues, and how it will impact the review process. They are worried about having a lot of non-refereed papers and multiple versions of papers out there.

My question is that if libraries can no longer afford to purchase commercial academic publications or books, then what do you think the future holds? Are academic commercial publishers going to remain viable? What is the future of paper and books? Do you think there is any future at all and if so, what is it? How do you see that?

Richard Lucier: As I mentioned, I don't see electronic versions replacing paper wholesale at this time. I think it is going to be a long evolution, that there are problems such as archiving that have to be solved before one can replace the other.

We are going through a period now, I think, of trying to understand how our faculty and the research community will use the electronic versions, what they prefer about them.

What we are seeing with things like Highwire Press, for example, is that the print version and the electronic version are getting further and further apart, and the electronic technology is being exploited to provide products that are much more beneficial to you than the paper might have been.

So, there is an evolution going on. I hope at UC that we will be able to cancel some print publications in the year 2000. Right now we have as many as, if not more than, nine copies of a particular journal, one at each of our campuses. We could potentially in 2000, if we provide good electronic access

to some of these titles, cancel all the paper except for two and save one in the north and one in the south for archiving purposes.

Evelyn Goldfield: That will cost the publishers money.

Richard Lucier: Right, it will cost the publishers money. That is correct.

Robert de Levie, Georgetown University: You have talked about journals. How about books?

Richard Lucier: It depends on what kind of books you are talking about. I think that digital technology can be very useful for reference books and reference databases. If you are talking about certain kinds of scholarly treatises in the humanities, I don't think we are going to see widespread replacement there at all in the immediate future. I think the digital technologies are going to take much greater hold in the sciences early on. The humanities, and less so the social sciences, are probably 5 to 10 years behind.

Robert de Levie: Even though those books nowadays are produced mostly in digital form?

Richard Lucier: Yes.

Robert de Levie: You mentioned $2.5 million gained. Is that because you reduced the number of subscriptions from nine to one, and what is the offsetting cost of not knowing whether 5 years from now you will have to buy the paper copies anyway?

Richard Lucier: We won't because we won't have gotten rid of all of the paper. We are making certain that we maintain in storage facilities—we have a storage facility in the north and the south—paper copies should we need to do that.

The $2 million plus was gained by expanded access. So, for example, you might have had 30 ACS subscriptions at Berkeley and at LA but only 5 at Riverside and 8 at Santa Cruz, and now everybody at all nine campuses is getting access to all as well as the fact that the access for, let us say, Berkeley alone, which may have subscribed to all of them, costs less because we went as part of a consortium. So, there are savings in those two areas.

Gintaris Reklaitis, Purdue University: One of the most important and underappreciated resources in the entire publications review process is the reviewers. Clearly as the publications process continues to expand, the demands on the reviewers will also. Do any of the business models that you are examining for scientific publication take into account this important resource and how we might stimulate it to handle this expansion?

Richard Lucier: The model where the university moves into publishing very much takes advantage of that resource, which is part of the university already. Essentially what we do now for the most part is give it away the commercial publishers at no cost so that they can then add a huge mark up to it when we buy back that information that has been peer reviewed by our faculty, and so it makes perfect sense for the university or federations of universities to do that together.

My problem with the Highwire model is that it is one university, and science and scholarship cut across universities too much, and it makes much more sense in my opinion to try to federate this in some way across groups of universities rather than try to go solo, and that is why UC isn't pursuing that particular strategy.

14

Electronic Journal Publishing at the American Chemical Society

Lorrin R. Garson
American Chemical Society

Scientific, technical, and medical (STM) publishing is a unique enterprise with the following characteristics:

- Small number of subscribers: 1,000 to 10,000 subscribers per title.
- High costs for quality control, which includes peer review and technical editing.
- High production costs because of complex information such as mathematics and high-quality graphics.
- Ever-increasing pressure from authors to publish more material.
- Static or decreasing funds for purchasing publications.
- Strong competition among publishers for high-quality content, good editors, and subscribers.
- Dominant STM title publication by commercial publishers.
- Largely price-independent competition for subscription sales. Each journal is a "limited monopoly"; that is, an article published in one journal does not appear in any other. There is a long-standing tradition against duplicate publication.
- Steady decline in subscriptions to STM journals in the past 15 to 20 years, and the trend is expected to continue.

GROWTH OF SCIENTIFIC LITERATURE

Although all of these factors contribute to the growing crisis in STM information transfer, the pressure to publish an increasing amount of material is arguably the greatest single factor in the growth of the scientific literature. Figure 14.1 shows the growth of chemical papers (excluding patents, monographs, and books) during the decades starting when *Chemical Abstracts* began publication in 1907.[1] Except during the periods of World Wars I and II, the increase has followed an exponential growth

[1] Data from "CAS Statistical Summary 1907-1997," Chemical Abstracts Service, Columbus, Ohio.

pattern through the years, including the most recent decade, 1987 to 1996. Chemistry, a relatively mature science, is probably reasonably representative of the growth of STM publishing in general.

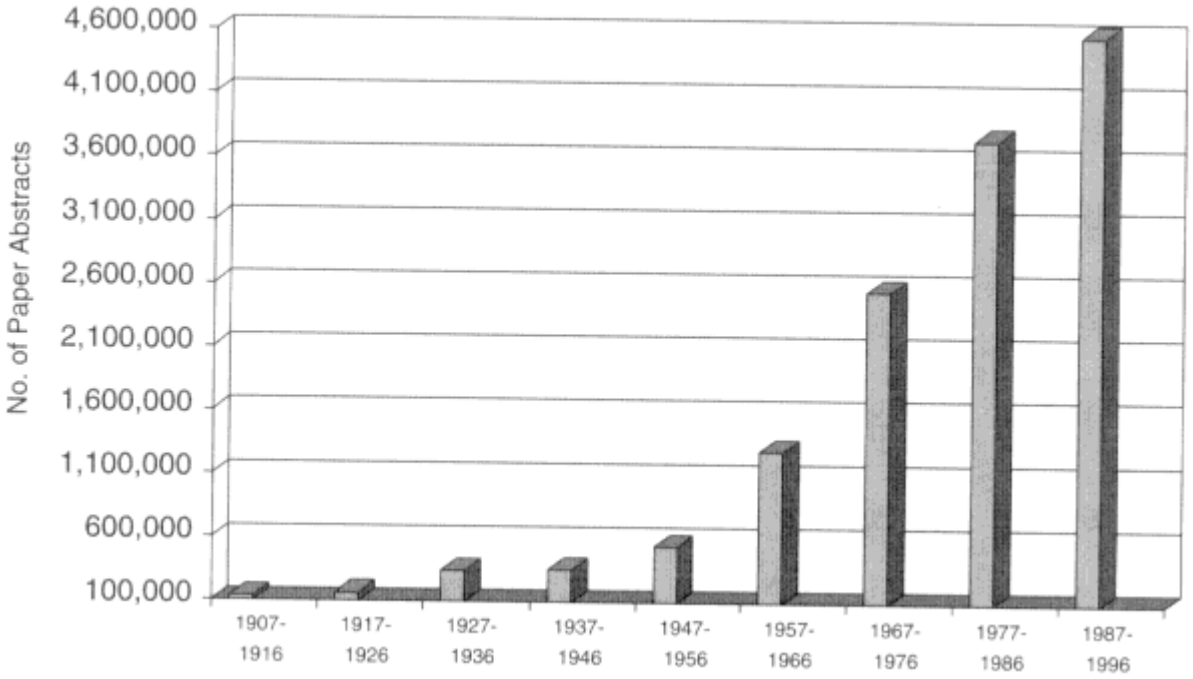


Figure 14.1
Number of abstracts of papers in Chemical Abstracts, 1907 to 1996 by decade.

To remain competitive, publishers publish more material by introducing new journal titles in response to emerging fields and publish more papers in existing titles.[2] Figure 14.2 shows the growth in the ACS journal-publishing program from 1980 through 1997, in terms of both articles and pages published per year. During this 17-year period, the number of articles grew 114 percent and the number of pages published increased by 229 percent; this is an average annual growth rate of 6.71 percent and 13.5 percent, respectively.

This exponential growth of STM literature is exacerbated by the increase in article length. For ACS journals, the average article has grown from 5.39 pages/article in 1980 to 7.30 pages/article in 1997 (see Figure 14.3). The decrease in article length from 7.34 pages/article in 1995 to 7.26 pages/article in 1996 is the result of a concerted effort made by ACS editors to encourage authors to reduce the length of their manuscripts. Unfortunately, at this rime there is no indication that the exponential growth of the STM literature is slowing.

Many subscribers object to subscription prices rising faster than the rate of monetary inflation, ignoring "page inflation" in most titles brought about by the increasing number of manuscript submissions and growth in manuscript length. This situation provides a significant marketing challenge for all

[2] The development of new scientific fields and increasing specialization also contribute to the development of new journal titles.

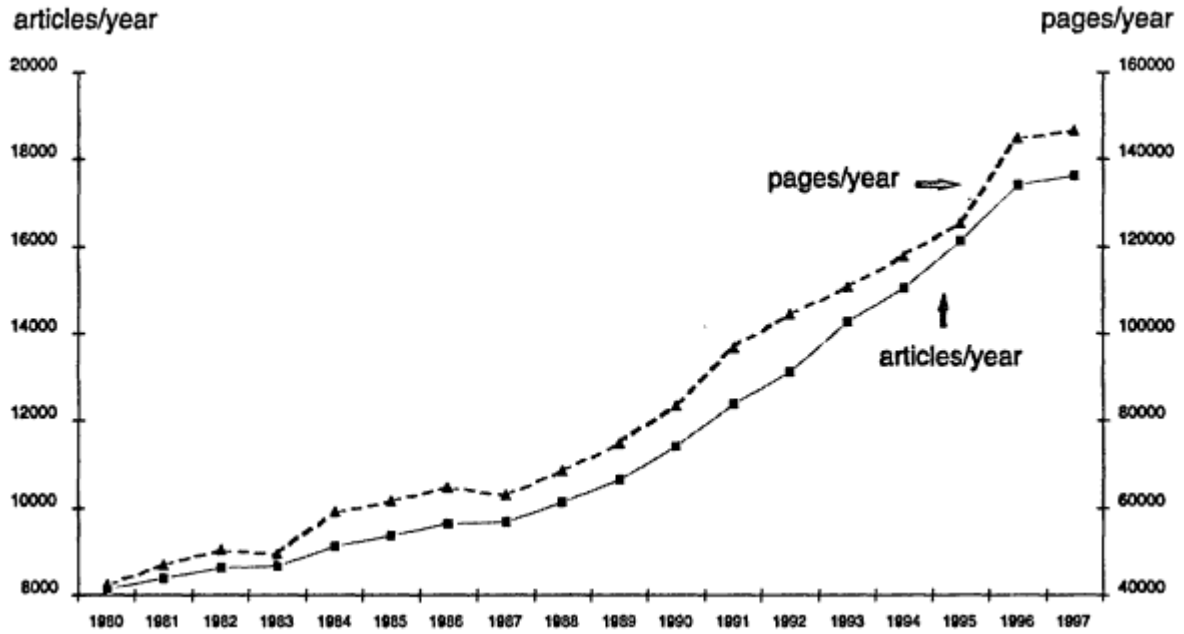


Figure 14.2
Growth in ACS journal publishing, 1980 to 1997.

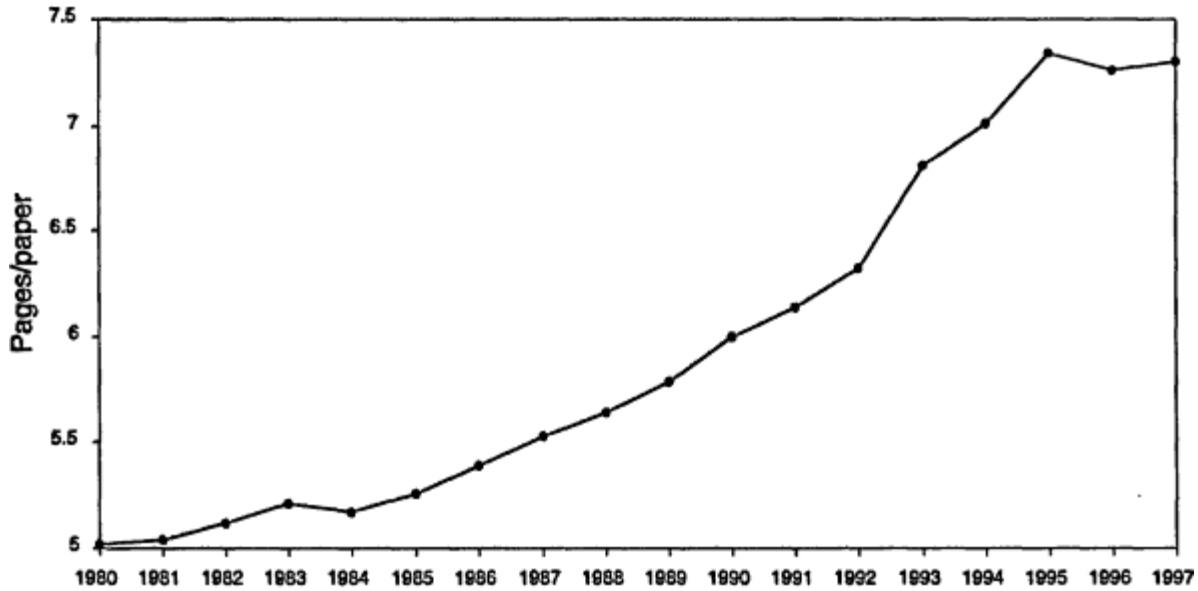


Figure 14.3
Growth in length of average article in ACS journals, 1981 to 1997.

publishers. Also, many subscribers do not differentiate between an 8 percent annual subscription price increase for a $1,000 subscription for a journal from a not-for-profit publisher and an 8 percent increase for a $5,000 journal from a commercial publisher. From the customer's perspective, they are both an 8 percent increase. Yet in terms of dollars, one has increased by $80 and the other $400—a difference of $320. This perception presents an additional challenge in marketing for not-for-profit publishers.

FINANCIAL CONSIDERATIONS

Any journal-publishing endeavor must generate adequate revenue to meet expenses or that endeavor will collapse. In addition, publishers in the commercial sector must also distribute dividends to stockholders and pay corporate taxes. Not-for-profit publishers, which are largely professional societies and a few universities, do not distribute dividends to shareholders or pay corporate taxes, but typically use the small amount of excess revenues over expenses to support their core objectives and programs. The ACS is a not-for-profit publisher (as well as a "not-for-loss publisher") and has had a successful journal-publishing program since 1879. Surpluses from its journal publishing operations, which are targeted at less than 10 percent of gross revenues, are used to invest in future publishing activities as well as to support a variety of ACS scientific and educational programs.

Historically, the great majority of revenue has come from subscription sales. Author page charges and advertising revenue have also been minor sources of income, but subscription sales have been and continue to be the major source of revenues for STM publishers. Tables 14.1 and 14.2 show the sources of revenues and expenses, respectively, in 1996 for ACS journal publishing operations.[3]

Subscription sales constitute over 80 percent of revenues and of these revenues, 90 percent are from institutional sales and 10 percent from sales to ACS members. Sales of journals to ACS members are financially neutral; that is, members obtain journals at "run off" costs. Although sales to members are of little financial consequence, distribution to members is an important part of the ACS's mission to disseminate scientific information broadly, and one could argue that without sales of subscriptions to ACS members, pressure on libraries to purchase more subscriptions would increase. Reprint revenues are rapidly declining with the availability of the journals in electronic form.

Expenses for journal production fall into two categories: first-copy costs and distribution costs. First-copy costs include the first five items in Table 14.2 and constitute 84.3 percent of all expenses. Distribution costs include the last three items in Table 14.2 and constitute 15.7 percent of all costs. The STM publishing industry average for first-copy costs is about 80 percent. Data for 1996 were selected because this was the last year in which there were no production expenses and revenues associated with electronic delivery of journal products, although there were significant R&D costs for these activities. Thus, as electronic delivery of information increases, costs associated with that mode of distribution will increase while costs associated with paper, printing, and shipping will decrease, assuming fewer copies of print journals are produced.

The notion of first-copy cost is important when considering electronic publishing. Regardless of whether distribution of information is via the traditional, printed journal or by electronic means, production of the first copy for the ACS constitutes 84.3 percent of all expenses. Database building and composition are the single largest expense. This includes assignment of journal-article components into appropriate data elements; appropriate rendering of tabular and mathematical material; handling of graphic data, including half-tones and color; page lay-out; and generation of SGML and HTML.

[3] From "Economics of Scientific Publishing," L.R. Garson, 214th ACS National Meeting, Las Vegas, Nevada, September 8, 1997.

TABLE 14.1 ACS Journal Publishing Revenues, 1996

Revenue Source	Percentage
Subscriptions	81.3
Reprints & page charges	10.5
Microfilm & back issues	3.4
Copyright royalties	1.9
Other	2.9
Total	100

TABLE 14.2 ACS Journal Publishing Expenses, 1996

Expense	Percentage
Peer review and external editors	19.3
Technical editing	12.8
Database building and composition	43.4
Marketing and sales	7.3
R&D	1.5
Paper, printing, and distribution	8.3
Reprint and microfilm reproduction	5.3
Miscellaneous	2.1
Total	100

Publishers are often admonished, "If you would only modify your production methods to accommodate an electronic environment rather than a print-oriented world, you could save 70 to 80 percent of production costs." This admonition is usually based on the assumption that publishers are not using computer-based manufacturing systems, or are applying these technologies inappropriately. This presupposition is generally false and is certainly untrue for the ACS's journal publishing program.

CHRONOLOGY OF ACS ELECTRONIC JOURNAL DEVELOPMENT

Like most "overnight successes" in the entertainment world, the success of electronic journals did not occur overnight. For the ACS, the ability to make its journals available on the World Wide Web in a cost-effective manner is the consequence of 25 years of investment in computer-based systems and staff training. Below are several milestones that led to the first ACS journal being made available on the Web in April 1996—*The Journal of Physical Chemistry*—on the occasion of its 100th anniversary.

1975

Journal production initiated in-house using a database approach.

1980

1,000 articles from the *Journal of Medicinal Chemistry* loaded on Bibliographic Retrieval Systems (BRS) as an experimental prototype electronic journal.

1981

Experimental file of 16 ACS journals loaded on BRS.

1982

Full-text file of ACS journals becomes a commercial product on BRS. The file remained active until 1985.

1986

CJACS (Chemical Journals of the American Chemical Society) file made available worldwide on STN International.[4] This file remained active until September 1998.

1993

Supporting Information[5] for the *Journal of the American Chemical Society* made available worldwide via Gopher.

1995

Supporting Information for all ACS journals made available via Gopher.

1996

The *Journal of Physical Chemistry* released on the Web in June followed by *Biochemistry* and *Environmental Science & Technology* in August.

1997

The *Journal of the American Chemical Society* and the *Journal of Organic Chemistry* released on the Web in April.

1997

Remaining 20 ACS journals released on the Web at 5:15 PM on September 7th.

1998

ASAP articles[6] in production on January 1st.

WHAT CUSTOMERS WANT IN ELECTRONIC JOURNALS

The producer of any product or service ignores its customers at great peril. Customers for scientific journals are its suppliers of manuscripts (authors) and purchasers of journal subscriptions (subscribers). Subscribers are both institutions, such as libraries, and individual ACS members. ACS members may be authors as well as subscribers. Our market research has shown that subscribers want the following attributes in electronic journals:

1. Useful, accurate information,
2. Fast delivery,
3. Low cost,
4. Access in perpetuity, and
5. Seamless access across publishers and databases.

Usefulness is difficult to characterize, although each individual has an intuitive sense of what is useful at the moment. Often what is not useful or interesting today may become so in the future. Accurate information seems to be best attained by vigorous peer review and author integrity. Fast delivery is largely controlled by the efficiency of the publisher and the speed of peer review. Low cost is also dependent on the efficiency of the publisher as well as a wide variety of business considerations. Access in perpetuity and seamless access across publishers and databases have not yet been realized and are the focal points of a variety of experiments and studies. Perpetual access is of greater concern to institutional rather than individual subscribers.

[4] Subsequently several other STM publishers also made their full-text, scientific journals available on STN International: WHY from John Wiley & Sons, CJELSEVIER from Elsevier Science Publishers, and CJRSC from the Royal Society of Chemistry in England.

[5] "Supporting Information" is an important component of an article, but is not essential to the major thrust of the paper or critical to its readability. Supporting Information contains information such as experimental details, spectra, x-ray crystallographic data, and various other types of numeric data. Supporting Information is not printed with the corresponding journal article but has been distributed on microfilm and microfiche. Historically, the ACS has published about 80,000 pages per year of this information. Starting in 1999, Supporting Information will only be available electronically on the Web, not on micro forms.

[6] ASAP (as soon as publishable) articles are papers that have been through the peer review process, revision, and author approval of page proofs, and then made available on the Web within 48 hours of final corrections being made. These articles are identical to the corresponding article printed in the journal and made available on the Web, except that ASAP articles do not contain page numbers.

What do authors want in electronic journals?

1. To publish in prestigious journals,
2. Peer review,
3. Rapid publication,
4. Wide dissemination, and
5. Attractive presentation.

In any particular field, experienced practitioners are aware of the prestigious journals in that discipline, and publishing in those journals is more desirable and generally more rewarding to one's career than publishing in journals of lesser prestige. Although an author may feel peer review has treated his work badly at one time or another, peer review is widely held by the scientific community to be valuable if not essential. A well-run peer review operation undoubtedly contributes to the prestige of a journal. Speedy publication is an important consideration for authors in choosing a journal to which to submit their manuscripts. Authors are also aware of the circulation of journals and use this as another consideration in selecting where to publish. *Science* and *Nature* are examples of high-circulation publications to which many authors wish to submit their work. Although there are important publications that use camera-ready material, and which consequently have less than optimal appearance, an attractive presentation of a journal is important to many authors and of even greater importance to readers.

THE ADVENT OF ELECTRONIC PUBLISHING

To borrow from one of Winston Churchill's famous speeches,[7] 1997 marks "the end of the beginning" in the development and deployment of electronic journals. This is undoubtedly the consequence of improvements made in computer and telecommunications technology over the years and in particular the development of the World Wide Web. Web information systems are becoming readily available among the scientific community, increasingly powerful, and easier to use. Researchers and librarians are generally accepting the inevitability of this great transformation. Figure 14.4 shows the increase in use of the ACS Publications Division Web server from February, 1995 to September, 1998—from 647 to 7,976,036 pages. It is interesting to note that the volume of information delivered via the Web to paying subscribers in September 1998 was equal to that delivered in March 1998 when the ACS journals were available without charge and open to the public. As expected when free access to the journals on the Web was turned off, the amount of information delivered dropped, but paying customers are dramatically increasing their use of Web journals—a most gratifying trend.

As one might expect, the use of Web information is international. During the week of January 28th to February 3rd, 1998, the geographic distribution of customers accessing the ACS's Publications Web site was measured. The results are shown in Figure 14.5.

Although 53 percent of the traffic came from the United States, 47 percent came from outside the United States. Japan was the largest consumer of Web information, followed by Germany, the United Kingdom, Canada, France, Italy, South Korea, Spain, The Netherlands, Switzerland, and 69 other countries.

[7] On November 10, 1942, Churchill made a speech at the Lord Mayor's Luncheon at Mansion House on the occasion of the defeat of Erwin Rommel's forces at El Alamein in Egypt. In that speech he made the now famous comment: "Now this is not the end. It is not even the beginning of the end, but it is perhaps, the end of the beginning."

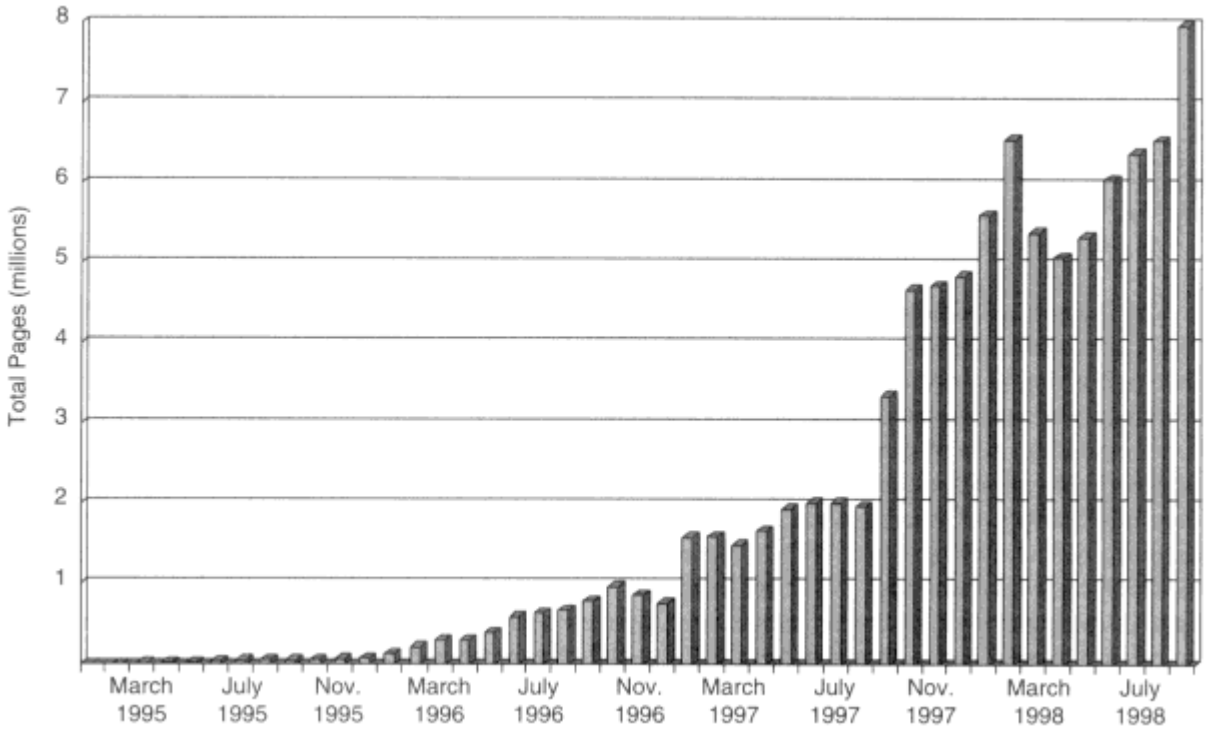


Figure 14.4
Total pages (in millions) transmitted from the ACS Publications Division World Wide Web server, 1995 to 1998.

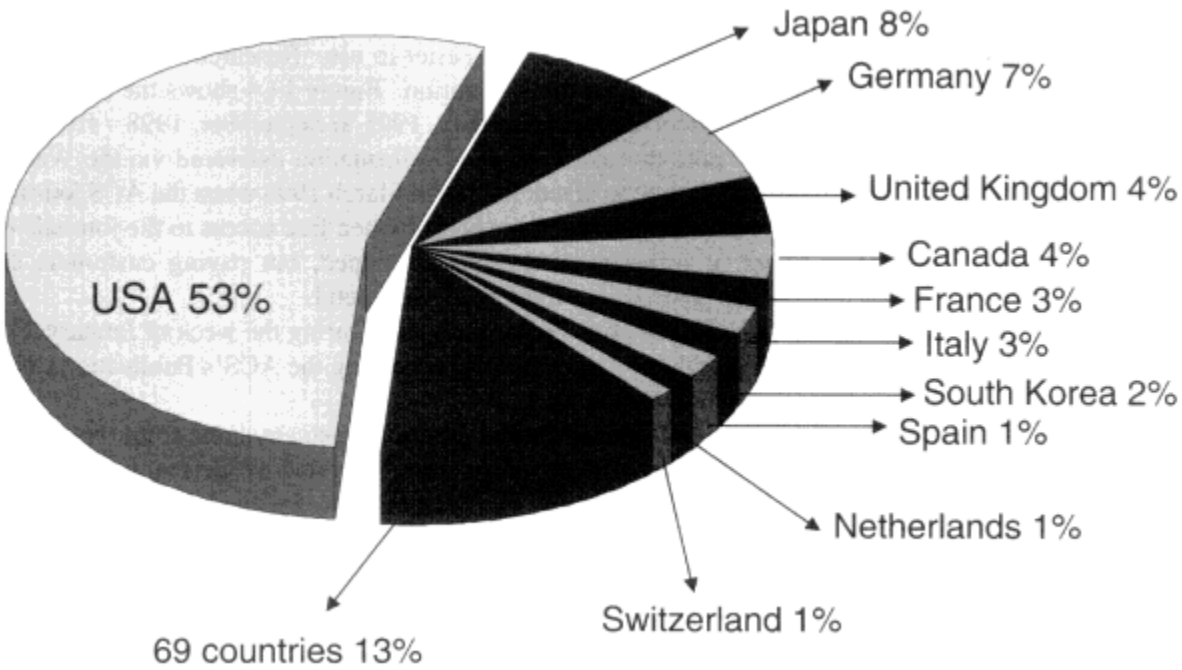


Figure 14.5
Geographic distribution of customers, January 28 to February 3, 1998.

TRANSFORMATION FROM PRINT TO ELECTRONIC PUBLISHING

While the use and sales of electronic journals are promising, there are many unanswered questions facing the scientific community and publishers in the near future. Among these are the following:

1. How can current Web systems become more comprehensive in content?
2. How can access and use be made simpler and "seamless" across databases?
3. How will researchers, librarians, and publishers fund necessary investments in technology?
4. What are fair, reasonable, and acceptable prices for subscriptions to Web journals?
5. What terms and constraints for use of Web journals are appropriate?
6. How much should be charged for an individual article online?
7. Who is responsible for archiving Web journals, and what is the commitment?
8. What are the function and status of journals that are available online only?
9. Will researchers and librarians value and pay for costly enhancements to Web journals?
10. Where will the money come from to pay for both print *and* Web journals?
11. What is the outlook for individual subscriptions—print and online?

Among these challenges, two merit special comment. Significant progress is being made in simplifying access across databases. The Digital Object Identifier (DOI) holds promise for providing persistent, seamless linking. The DOI affords a mechanism for a persistent link to digital objects, such as Web articles or their components, and is inherently a lookup mechanism.[8] It is likely that a minimal set of metadata will be made available with each DOI that will provide for identification of digital objects. It is also likely that abstracting and indexing services will include DOIs for items they cover and thus provide a much richer set of metadata for locating digital information. Other services, such as the National Library of Medicine's PubMed[9] in the medical sciences, and the nascent PubRef service also offer promise for improving linking between databases.

The digital archive is a particularly vexing issue, which is critically important to the long-term preservation of the scientific record. Broad acceptance by institutional subscribers of electronic journals as a replacement for print will not likely take place until this issue is resolved. Traditionally, libraries have served as the institution for archiving information in print. Remember the discussions on acid-free paper? However, libraries are not well positioned to serve this role for electronic journals. Publishers have not served in an archival capacity and have depended on libraries for this service. Commercial publishers are unlikely to serve as a repository for archival data unless there is an adequate market to support the expense of maintaining such an archive. It has been suggested that "trusted third parties" serve as institutional repositories of digital archives, but such institutions are not yet forthcoming, despite the claims of a few organizations purporting to serve this role. The ACS has made a public commitment to preserve its digital journals, as have other society publishers, but such commitments are not adequate by themselves.

Expenses associated With maintaining a digital archive are highly uncertain. Costs associated with migration from one hardware system to another are more predictable than conversion of current file formats to unknown file formats in the future. The diversity of file formats exacerbates the Complexity and expense of future conversion. New file formats for various types of scientific data are also likely to

[8] For more information on the DOI, see the Web site of the International DOI Foundation at <http://www.doi.org/>.
[9] For more information on PubMed, see <http://www.ncbi.nlm.nih.gov/PubMed/>.

be adopted and included with digital journals. An uncertain market for the archive, coupled with an uncertain but probably high cost for its maintenance, is working against a solution to this problem.

DISCUSSION

Stephen Heller, National Institute of Standards and Technology: A couple of things, Lorrin. First of all, it was a very nice presentation. You mentioned 50 free hits to articles posted on the Web, but what happens if 50 Web crawlers come looking through your site, and they are the first 50 that come in? What happens to the first real chemist?

Lorrin Garson: We think we can outsmart the Web crawlers in this instance. We will put them in a protected partition, and I think we can keep Web crawlers out.

Stephen Heller: A second item for e-journals. I think there is a Chemical Abstracts Web site home page indicating that there are a couple of dozen journals that Chemical Abstracts right now does process. There are all of the electronic journals that it does, in fact, abstract and index. So, it is not just three or four journals. Chemical Abstracts has been very, very good about looking for journals that in fact are there for more than a millisecond. I also want to mention the Internet Journal of Chemistry, which I have been involved with as a member of the editorial board, and suggest that people take a look at it. Right now it is free until the year 2000. One thing about it—and we were talking with Richard (Lucier) and a little bit with Lorrin (Garson) about the pricing and the cost—is that the way we put it together, it is not a big major organization operation. The cost of putting it together is infinitely less, and the breakdowns that you had there are quite different. It seems to me that a major problem with the economics that you are talking about for all existing print journals is the overhead and baggage that exists right now. The fact that you are trying to carry two things at one time makes it extremely difficult to reduce costs, and if you just went electronic on a separate operation even the ACS could do it for less than it costs the way you are doing it now.

Lorrin Garson: We think we can, indeed, lower the cost of some operations, but I am afraid we are going to just basically disagree on the general picture that producing electronically would be significantly less expensive.

Allen Bard, University of Texas: Isn't it, though, a question of taking a camera-ready copy from what is now prepared for presentation? I think if you use camera-ready copy, that is, you use a disk the author supplies with very little massaging or fixing up, it is going to be a lot cheaper than what ACS does in a fair amount of production.

Lorrin Garson: That is true. However, we get 95 to 97 percent of our manuscripts in soft copy, and we do a lot of manipulation to get that data into the database. If it were camera-ready copy and we could just print it as we got it, the cost would be dramatically lower.

Allen Bard: Right.

Stephen Heller: Lorrin, that is one of the things this particular journal does. Everything comes in already formatted in HTML. So, it is essentially the camera-ready copy that Al is talking about. It is not that it is cheaper; it is just that the burden of the cost goes to the author as opposed to the publisher.

Allen Bard: But the author, if I can interject, from our experience doesn't do that great a job. Most authors don't want to be publishers, and so they will do something, but it is not what the publisher does. It may be still legible, and it depends on what you want to get out of it.

Robert Lichter, Camille & Henry Dreyfus Foundation: I have a series of questions. From discussions I have had over the last couple of years with a number of folks—librarians, publishers, ACS people—I do want to commend ACS for really taking the lead on this and looking out for the interests of its membership.

One librarian, in a discussion about the societies versus commercial publishers, said with considerable heat and no small amount of vigor that ACS is a business, and one should not distinguish between ACS and commercial publishers. I tried to argue that, but I would like to hear from other people about some of the arguments for making that distinction.

Another question I have regards your comment about scientists preferring to publish in prestigious journals. What actually defines a prestigious journal when you really get down to it?

Third, there have been increasing calls for sort of self-published journals—the journal that publishes perhaps 20 papers a year on a super-specialized topic—arguing that because the costs can be so low, it is possible to have any number of these small highly specialized journals, more specialized than the current specialized journals, and that that is a good thing for communication of scientific results. You have probably heard this, too.

I would like to hear your views and others' views on that assertion.

Lorrin Garson: On the issue that ACS is a business, the ACS does run in a businesslike manner, no question about that. If we didn't, we would be out of business. The fact is the ACS is a legitimate not-for-profit organization, and as I said earlier, we are also not-for-loss. A librarian or any other individual certainly could look at the ACS as a commercial business because it is run in a businesslike manner, but the reality is that the ACS is it not a commercial publisher. I doubt if there will be any resolution of this debate.

Robert Lichter: The next question was that of the very small, specialized journals that can be published so easily. The argument is whether that is good for scientific unity.

Lorrin Garson: I think that the biggest single thing missing in that model is marketing. That notion implies that one doesn't have to market, that you can put up a journal on the Web, send a few e-mail messages to your friends, and everybody will know about it. It doesn't work that way. Marketing is a very important part of publishing.

Session 4

Panel Discussion

Sam Kounaves, Tufts University: There were some questions earlier about the increasing volume of our publications. Most of us are both producers and consumers and at least in my opinion, and I don't know if it is my colleagues', the volume will not ever decrease. In fact, it is probably going to get worse, similar to grade inflation. Nobody is going to take the fast step to turn the volume down, because we use publications in tenure and so on. So, it's never going to happen.

We have some small societies that seem to have been successful in publishing their journals cheaply. I don't know if anybody has heard of the Society of Entomology—its *Journal of Entomology* is basically free to the members, a small group. I also belong to a small society, with about 500 or 600 members, for which I have set up a Web site and put the society newsletter on it. So I have had some experience with trying to start one, and have seen some of the problems involved with it, but on the other hand, it seems to me we may be starting afresh. We are not bound by tradition, as some of the larger societies might be, as to what we can or cannot do. Some of my colleagues don't agree with this, but I feel it is possible to publish a refereed, high-quality journal.

One of a society's missions many years ago was to gather information about members' research and disseminate it among the members. Thus, it seems it is very appropriate that the societies take a leading role in journal publishing. Libraries can be an archiving location for certain things. The societies could be an archiving source for the society publications. For, example, the ACS could be an archiving location for ACS journals. One hundred years from now the journals would still be there. Granted they might undergo several transitions in archiving media, but they would still be there and would still be archived in some format.

So, one question is, Is it not feasible for a society like ACS to take on archiving and to look at such things as page charges for authors and look at these smaller societies that seem to be doing successful publishing, like the Society of Entomology? Is ACS looking at this and is it possible to do this sort of thing?

Lorrin Garson: We are committed to archiving the ACS journals, no question about that. The issue of page charges brings back memories. The ACS has page charges in a few of its journals. These are

vestiges of a time when there were page charges in most of our journals. However, in competing with the commercial journals, which do not charge page charges, we felt we could no longer charge page charges.

There are, also, tax implications. The IRS said that if page charges are mandatory then scientific journals are advertising, and you must mark every piece of paper published in that journal, "This is an advertisement." We didn't want to do that. I don't know if the IRS has changed its position on this.

The ACS Board of Directors has said that we should eliminate all page charges, which even now are voluntary, as soon as we can. That decision was made about 15 years ago, and we are still tagging along with a small number of journals with page charges. Page charges don't seem to be a way to go as best as we can judge.

Richard Lucier: I would like to comment on the notion that small societies handle a large amount of the information that needs to be managed. I agree with you that that seems to be a reasonable way to go, but my concern is how that scales over time. What I have observed over the years with respect to the development of scientific databases is that they often start out as a "cottage industry" product. Developing the database was an interesting and innovative activity; an individual researcher/member of a society took it on for a while, but as the need for access, the importance of reliability, and changes in technology occurred, an infrastructure more robust than the "cottage industry" could provide was needed. Societies like ACS can provide some of that infrastructure. It is my belief that a federation of universities could logically provide that to groups of small societies as well. In some respects, HighWire Press is doing that in the biomedical sciences for a number of societies, providing an infrastructure that does scale. Each of those societies would have a great deal of difficulty building that infrastructure on its own.

Gary Mallard: One thing you said, Lorrin, that I would have to disagree with, concerns marketing. You actually don't have to market things. The WebBook, which gets a lot of usage, has never spent a nickel on advertisement, has never done anything, and it is a little bit like the baseball field in Iowa. If you build it, they do come, and if you build it with a reasonably high-quality product, I think they will come in droves.

Lorrin Garson: Especially if it is free.

Gary Mallard: Of course.

Robert Lichter, Camille & Henry Dreyfus Foundation: That gets back to the question of prestige.

Allen Bard, University of Texas: I can tell you my view. I think there are quantitative measures of it for whatever you want to look at, like impact factors, and although I don't fully agree with that, that goes along pretty much with a kind of community opinion. I know that certain departments, when they make tenure decisions, look at the list of publications, have a numerical multiplication factor, a division factor for different journals, you know, *Science, Nature, JACS*, and so on. So, there is, I think, a culture that believes this.

Lorrin Garson: Another factor that might play into prestige is rejection rate. Prestigious journals, shall we say, tend to have a much higher rejection rate, which probably leads to a higher quality. They publish the more high-quality material, and I think the community, certainly people who have published for any period of time, have a pretty good sense as to which journals have a high rejection rate and which don't.

Now, after you get beyond a certain point, if you are rejecting 90 percent of the papers, obviously you are rejecting a lot of good papers. You can identify the trash and the truly outstanding, but 95 percent of the material in the middle is much more difficult to judge. But yes, I think rejection rate is a factor in prestige.

Christos Georgakis, Lehigh University: I wanted to ask a question about copyright in this electronic age. Let me run you through the steps that one might take. I write a paper. I put it on my Web site and I label it "submitted for publication." I get the reviews back. I do some editing, and of course I assign the copyright to the journal that is going to publish it, but I still leave the original copy of the paper on my Web page for people to download. Am I in conflict of the copyright?

Lorrin Garson: This is good question, but I dislike answering questions about copyright because I know just enough to be dangerous. The issue is that you owned copyright when you first created that article. The question is whether you signed over those rights to anybody. If you haven't you are not in violation of copyright. You own the copyright.

To publish in ACS journals a requirement is that you transfer copyright. At that point you would indeed be in violation of copyright because you no longer own the article. So, you would be asked to take the article down. We have a policy that if authors want to post ACS articles on their intranet, that is, for their institution, that is quite all right, but it is not allowed to post them for general access.

Others of our journals have a policy with regard to what has been posted on the Web. Some of the editors feel rather strongly that if something is on the Web it is already published. It is open to the public, and they would not consider it for publication. Other ACS editors feel that posting it on the Web is equivalent to speaking at a conference. It is not true publication. It is something in between, and they would consider it for publication. So, there are many factors that come into your question.

Gary Mallard: I would like to ask you a question, Lorrin. Do you think that that mode of thinking serves the scientific community, or does it just serve the ACS?

Lorrin Garson: Here is the problem from the publisher's standpoint. If we do not ask for copyright, for example, then let us say 5 years from now we want to do something with the collection of journals that we cannot anticipate, that means we would have to go back to all 30,000 authors per year to get permission. This is impractical, which limits the opportunities we would have.

Gary Mallard: Like what?

Lorrin Garson: I cannot imagine what we might want to do. What we have done in the past, for example, is make the journals available on the Web. Had we not owned copyright we could not have done that.

Gary Mallard: You could have been assigned copyright for that purpose though but not necessarily have owned it—copyright can be held jointly. You don't have to have exclusive copyright and as a matter of fact the ACS does not have copyright to any work by government employees. So, things get published by the ACS without the ACS holding the copyright, and it doesn't in any way restrict you from using that material. That would, I think, answer this question about whether authors can put articles up on their own Web sites, which I think a lot of people would like to do,

Jack Kay, Drexel University: I was curious about your statement that the length of the articles being published is increasing regularly. What is the explanation for that? Do you know?

Lorrin Garson: I am not sure. We haven't done a formal study. In looking at articles I see two things. One is an increasing number of references. There is more literature to reference, and authors seem to be referencing it. So, the number of references has gone up.

I think there are in some cases a lot more data to be published than there have been before just because of modern instrumentation, but these are only casual observations on my part.

Al, do you have any sense of this issue?

Allen Bard: I sense that papers now are more complex and more data intensive, and the effort we have been making in what I consider an evolutionary period is to get more of that put on the Web, in other words at least in this intermediate period to let the printed version reflect the core of an article but try to get more and more of this other stuff and supporting information onto the Web. Different communities within chemistry agree with this to different extents. That would be my best guess without any formal study.

David Smith, DuPont: As I listen to this conversation I believe there is an issue here. I look at three important transformations in this process; one is data to information, then information to knowledge, and finally knowledge to understanding. It seems to me that understanding usually ends up in textbooks, and where we are hung up in the publication game is in whether we should be publishing information or publishing knowledge.

It seems to me that it usually requires a lot of work to transform information into useful knowledge, and I submit that in some of the journals that I read, most of what gets published is information. A lot of it is probably useless information, and we are not doing a good enough job of reviewing papers to weed them out.

Allen Bard: Anybody want to address that? David, do you want to specify some journals? I tend to agree, and I think the librarians should be stronger in gauging the quality of the journals they are taking and in finding and probably not subscribing to journals that are largely of that nature.

Richard Lucier: Let me add briefly that librarians subscribe to the journals that are mandated by their faculty, and we do not feel that it is our responsibility to do that quality control, that it is really faculty's responsibility to do that.

Stephen Heller, National Institute of Standards and Technology: Full copyright is not necessary for a publisher. A license to publish, with the commercial rights to use it, is really sufficient for a publisher to earn income, and full copyright isn't needed. A committee of which I am a member wrote a policy article on this issue of copyright versus a license to publish. This group was sponsored, in pan, by the Dreyfus Foundation, and the committee, The Transition From Paper, put something in *Science* about 2 months ago (September 4, 1998, pages 1459-1460), and I was wondering if there are any comments from the panel members about the notion of having a license to publish and what would be insufficient (for publishers) about that?

Lorrin Garson: I think I have said enough about copyright and done all the damage I care to do! I will let the other panelists try to address this.

Gary Mallard: I am not going to touch it because we are at the other end of it. We actually are going out and getting data out of the literature. I think the risk that you run when creating an electronic database is that people may feel like you are infringing upon their copyright when you take the data out of the literature and put it into an electronic format. The whole issue of fair use, which I don't think has been addressed in any very realistic way when it comes to electronic format, is still something that may well end up in the courts. I don't know whether it is going to end up in the courts for scientific data because there is just not enough money sloshing around in scientific data, but it may well end up there for other reasons. There is this historic notion of fair use and what you can and cannot do, and I don't know where we are on that.

Richard Lucier: I think you as a community have to decide what you want to do about copyright. If you as a community decide that you are going to do what was specified in that editorial I cannot imagine that the ACS, as well as other publishers, won't continue to publish your materials. I think it is truly up to you as a community, and I would encourage you as a community to look very carefully at the recommendations that were in that editorial as being potentially desirable ones to act upon.

Randy Collard, Dow Chemical Company: Just a quick question for Lorrin. You identified in your cost analysis the database as the largest factor. As you look at the advances in technology and expertise, what do you see that is a breakthrough necessary in that area?

Lorrin Garson: I think it is likely there will be a number of small things that accumulate to provide greater efficiency, not necessarily one large piece of technology. For example, within our own environment we are getting more and more clever at writing parsers so that we can take information in various word processing packages and parse them and identify elements algorithmically. That particular piece in itself I think will lead to significant savings.

As far as technology coming down the road, it is very difficult to predict what may have a dramatic effect. Certainly as I look back over the last 20 to 25 years the issue of database publishing in itself undoubtedly has contributed very significantly to our cost containment and the ability to produce electronic products. The number of people involved with producing the electronic journals and doing quality control is two or three people. To manage all that data is a tribute to the database approach to publishing.

Robert Lichter: I would like the views from the remaining panelists about the following: Recently there has been a lot of legislative heat and not too much light about the issue of copyright, which seems to have been temporarily resolved. Do you think this issue is going to emerge again? How do you think the concept of copyright will change in light of electronic publishing?

Gary Mallard: From the point of view of someone who tries to pull information out, the European view of copyright is distinctly dangerous, I think, because in effect where the European Community is heading is basically that you can copyright the telephone book. If that becomes the case, it seems to me that you are a very small step away from copyrighting the boiling point of methane, and certainly American law has said that you cannot copyright the boiling point of methane. I don't know where this issue is going—you can argue that the current law that was just passed by Congress and signed comes very close. If you have put the boiling point of methane into a compilation, then it has the potential to be copyrighted. I think it is extraordinarily dangerous for our ability to deliver scientific information economically.

Appendixes

Appendix A

List of Workshop Participants

Madeline Adamczeski, American University
Richard C. Alkire, University of Illinois at Urbana-Champaign
Herman L. Ammon, University of Maryland
Raymond A. Bair, Pacific Northwest National Laboratory
Allen J. Bard, University of Texas
Rodney Bartlett, University of Florida
John S. Binkley, Sandia National Laboratories
Donald M. Burland, National Science Foundation
Bridget Carragher, University of Illinois at Urbana-Champaign
Randy S. Collard, The Dow Chemical Company
Robert Cordova, Elf Atochem, Inc.
Peter T. Cummings, University of Tennessee
Lawrence A. Curtiss, Argonne National Laboratory
Robert de Levie, Georgetown University
David A. Dixon, Pacific Northwest National Laboratory
Douglas J. Doren, University of Delaware
Richard Dubois, National Institutes of Health
Thom H. Dunning, Jr., Pacific Northwest National Laboratory
Thomas F. Edgar, University of Texas
Karolyn K. Eisenstein, National Science Foundation
Stephen Elbert, National Science Foundation
Thomas A. Finholt, University of Michigan
Farley Fisher, National Science Foundation
Joseph Francisco, Purdue University
Susan Fratkin, Southeastern Universities Research Association
Jean H. Futrell, Pacific Northwest National Laboratory
Barbara J. Garrison, Pennsylvania State University

Lorrin R. Garson, American Chemical Society
Christos Georgakis, Lehigh University
Evelyn Goldfield, Wayne State University
Susan L. Graham, University of California, Berkeley
Jane Griffith, National Research Council
Stephen R. Heller, National Institute of Standards and Technology
Judith Hempel, University of California, San Francisco
Richard L. Hilderbrandt, National Science Foundation
Richard Hirsh, National Science Foundation
Daniel Hitchcock, U.S. Department of Energy
Jack Kay, Drexel University
Richard Kayser, National Institute of Standards and Technology
William H. Kirchhoff, U.S. Department of Energy
Michael L. Knotek, U.S. Department of Energy
Samuel P. Kounaves, Tufts University
Robert L. Lichter, Camille & Henry Dreyfus Foundation
Richard E. Lucier, University of California
W. Gary Mallard, National Institute of Standards and Technology
Robert S. Marianelli, Office of Science and Technology Policy
Paul Maupin, U.S. Department of Energy
David R. McLaughlin, Eastman Kodak Company
L. Eugene McNeese, Oak Ridge National Laboratory
Gregory J. McRae, Massachusetts Institute of Technology
Paul Messina, California Institute of Technology and U.S. Department of Energy
William S. Millman, U.S. Department of Energy
Raul Miranda, National Science Foundation
Janet Nelson, American Chemical Society
Janet G. Osteryoung, National Science Foundation
Aristides Patrinos, U.S. Department of Energy
John B. Pfeiffer, Air Products and Chemicals, Inc.
John A. Pople, Northwestern University
Clint Potter, University of Illinois at Urbana-Champaign
Gintaras V. Reklaitis, Purdue University
Michael E. Rogers, National Institute of General Medicine
Celeste M. Rohlfing, National Science Foundation
L. David Rothman, The Dow Chemical Company
Joel H. Saltz, University of Maryland
Stanley I. Sandler, University of Delaware
Peter Schmidt, Office of Naval Research
Gustavo E. Scuseria, Rice University
John Sessler, Ballistic Missile Defense Organization (retired)
Donald Singleton, National Research Council of Canada
Jeffrey Skolnick, Scripps Research Laboratory
Michael E. Smith, Exxon R&D Laboratory
W. David Smith, E.I. Du Pont de Nemours and Company
Peter R. Taylor, San Diego Super Computer Center, University of California, San Diego

Michael Thompson, Pacific Northwest National Laboratory
John S. Tse, National Research Council of Canada
Marek W. Urban, North Dakota State University
David L. Venezky, Naval Research Laboratory
Andrew B. White, Jr., Los Alamos National Laboratory
Carter White, Naval Research Laboratory
G. Edwin Wilson, University of Akron
William D. Wilson, Lawrence Livermore National Laboratory
William T. Winter, Environmental Science & Forestry, State University of New York, Syracuse
Robert S. Wood, Rohm & Haas Research Laboratories
Jeff Yellets, Allied Signal
Cynthia Zoski, University of Rhode Island

Staff

Douglas J. Raber
David Grannis
Ruth McDiarmid
Sybil Paige

Appendix B

Origin of and Information on the Chemical Sciences Roundtable

In April 1994, the American Chemical Society (ACS) held an Interactive Presidential Colloquium entitled "Shaping the Future: The Chemical Research Environment in the Next Century." The report from this colloquium identified several objectives, including the need to ensure communication on key issues among government, industry, and university representatives.[1] The rapidly changing environment in the United States for science and technology has created a number of stresses on the chemical enterprise. The stresses are particularly important with regard to the chemical industry, which is a major segment of U.S. industry, makes a strong, positive contribution to the U.S. balance of trade, and provides major employment opportunities for a technical work force. A neutral and credible forum for communication among all segments of the enterprise could enhance the future well-being of chemical science and technology.

After the report was issued, a formal request for such a roundtable activity was transmitted to Dr. Bruce M. Alberts, chairman of the National Research Council (NRC), by the Federal Interagency Chemistry Representatives (FICR), an informal organization of representatives from the various federal agencies that support chemical research. As part of the NRC, the Board on Chemical Sciences and Technology (BCST) can provide an intellectual focus on issues and fundamentals of science and technology across the broad fields of chemistry and chemical engineering. In the winter of 1996, Dr. Alberts asked BCST to establish the Chemical Sciences Roundtable to provide a mechanism for initiating and maintaining the dialogue envisioned in the ACS report.

The mission of the Chemical Sciences Roundtable is to provide a science-oriented, apolitical forum to enhance understanding of the critical issues in chemical science and technology affecting the government, industrial, and academic sectors. To support this mission, the Chemical Sciences Roundtable will do the following:

[1] *Shaping the Future: The Chemical Research Environment in the Next Century*, American Chemical Society Report from the Interactive Presidential Colloquium, April 7-9, 1994, Washington, D.C.

- Identify topics of importance to the chemical science and technology community by holding periodic discussions and presentations, and gathering input from the broadest possible set of constituencies involved in chemical science and technology.
- Organize workshops and symposia, and publish reports on topics important to the continuing health and advancement of chemical science and technology.
- Disseminate the information and knowledge gained in the workshops and reports to the chemical science and technology community through discussions with, presentations to, and engagement of other forums and organizations.
- Bring topics deserving further, in-depth study to the attention of the NRC's Board on Chemical Sciences and Technology. The roundtable itself will not attempt to resolve the issues and problems that it identifies —it will make no recommendations, nor provide any specific guidance. Rather, the goal of the roundtable is to ensure a full and meaningful discussion of the identified topics so that the participants in the workshops and the community as a whole can determine the best courses of action.