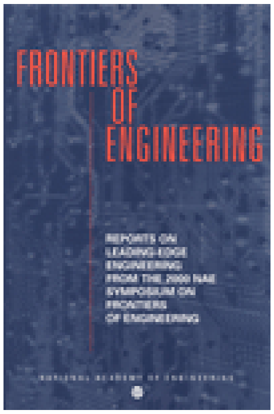


Frontiers of Engineering: Reports on Leading-Edge Engineering From the 2000 NAE Symposium on Frontiers in Engineering

National Academy of Engineering

ISBN: 0-309-50260-8, 144 pages, 6 x 9, (2001)

SIXTH ANNUAL SYMPOSIUM ON FRONTIERS OF ENGINEERING

NATIONAL ACADEMY OF ENGINEERING

NATIONAL ACADEMY PRESS
Washington, D.C. 2001

NATIONAL ACADEMY PRESS - 2101 Constitution Ave., N.W. - Washington, D.C. 20418

NOTICE: This publication has been reviewed according to procedures approved by a National Academy of Engineering report review process. Publication of signed work signifies that it is judged a competent and useful contribution worthy of public consideration, but it does not imply endorsement of conclusions or recommendations by the NAE. The interpretations and conclusions in such publications are those of the authors and do not purport to represent the views of the council, officers, or staff of the National Academy of Engineering.

Funding for the activity that led to this publication was provided by Cummins Engine Company, DaimlerChrysler Corporation Fund, Defense Advanced Research Projects Agency, Department of Defense Research and Engineering, Hewlett-Packard Company, Lockheed Martin Corporation, Lucent Technologies, Microsoft Corporation, National Aeronautics and Space Administration, Parsons Brinckerhoff, Inc., Pratt & Whitney, Sandia National Laboratories, Science Applications International Corporation, Xerox Foundation, Xerox Research and Technology, and the National Academy of Engineering Fund.

International Standard Book Number 0-309-07319-7

Additional copies of this report are available from National Academy Press, 2101 Constitution Avenue, N.W., Lockbox 285, Washington, D.C. 20055; (800) 624-6242 or (202) 334-3313 (in the Washington metropolitan area); Internet, http://www.nap.edu

Printed in the United States of America

THE NATIONAL ACADEMIES

National Academy of Sciences
National Academy of Engineering
Institute of Medicine
National Research Council

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 M. 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 M. Alberts and Dr. William A. Wulf are chairman and vice chairman, respectively, of the National Research Council.

ORGANIZING COMMITTEE

MICHAEL L. CORRADINI (Chair), Associate Dean, Academic Affairs; Professor, Nuclear Engineering and Engineering Physics, University of Wisconsin

BRENDA M. BOHLKE, Vice President, Parsons Brinckerhoff Inc.

CARLA E. BRODLEY, Associate Professor, School of Electrical and Computer Engineering, Purdue University

PETER T. CUMMINGS, Distinguished Professor, Department of Chemical Engineering, University of Tennessee, Knoxville

CHANG-BEOM EOM, Professor, Department of Materials Science and Engineering, University of Wisconsin

PATRICK HANRAHAN, Canon USA Professor, Computer Science and Electrical Engineering Departments, Stanford University

DEIRDRE R. MELDRUM, Associate Professor, Department of Electrical Engineering, University of Washington

RICHARD M. RATLIFF, Senior Vice President and Chief Architect, Strategic Architecture Team, Sabre Inc.

PATRICK M. SHANAHAN, Vice President and General Manager, 757 Programs, Commercial Airplane Group, The Boeing Company

GANESH SKANDAN, Vice President, Research and Development, Nanopowder Enterprises, Inc.

JOHN YIN, Associate Professor and Cargill Faculty Fellow, Department of Chemical Engineering, University of Wisconsin

Staff

JANET R. HUNZIKER, Program Officer

MARY W. L. KUTRUFF, Administrative Assistant

Preface

In 1995 the National Academy of Engineering (NAE) initiated the Frontiers of Engineering Symposium program, which every year brings together 100 of the nation's future engineering leaders to learn about cutting-edge research and technical work in different engineering fields. On September 14-16, 2000, the National Academy of Engineering held its sixth Frontiers of Engineering Symposium at the Academies' Beckman Center in Irvine, California. Symposium speakers were asked to prepare extended summaries of their presentations, and it is those papers that are contained here. The intent of this book, and of the five that precede it in the series, is to describe the content and underpinning philosophy of this unique meeting and to highlight some of the exciting developments in engineering today.

GOALS OF FRONTIERS OF ENGINEERING

The practice of engineering is changing. Not only must engineers be able to thrive in an environment of rapid technological change and globalization, but engineering is becoming more interdisciplinary. The frontiers of engineering are frequently occurring at the intersections of engineering disciplines, which compels researchers and practitioners alike to be aware of developments and challenges in areas other than their own.

At the three-day Frontiers of Engineering symposium, 100 of this country's best and brightest engineers, ages 30 to 45, learn from their peers about what is happening at the leading edge of engineering. This has great value for the participants in a couple of ways. First, it broadens their knowledge of current developments in other fields of engineering, leading to insights that may be applicable to the furthering of their own work. Second, because the engineers come from a variety of institutions in academia, industry, and government and from many different engineering disciplines, it allows them to make contacts with and learn from individuals whom they would not ordinarily meet in their usual round of professional meetings. This networking, it is hoped, will lead to collaborative work, facilitating the transfer of new techniques and approaches across fields.

The number of participants at each meeting is kept at 100 to maximize the opportunity for interaction and exchange among the attendees, who are invited to attend after a competitive nomination and selection process. The choice of topics and speakers for each year's meeting is carried out by an organizing committee composed of engineers in the same 30- to 45-year-old cohort as the participants. Each year different topics are covered, and, with few exceptions, different individuals participate.

The speakers at the Frontiers of Engineering symposium have a unique challenge—to make the excitement of their field accessible to a technically sophisticated but nonspecialist audience. To achieve the objectives of the meeting, speakers are asked to provide a brief overview of their fields and to address such questions as: What are the frontiers in your field? What experiments, prototypes, and design studies are completed and in progress? What new tools and methodologies are being used? What are the current limitations on advances? What are the controversies? What is the theoretical, commercial, societal, and long-term significance of the work? Many elements of these topics are captured in the papers as well.

CONTENT OF THE 2000 SYMPOSIUM

The four broad areas that provided the framework for the 2000 meeting were systems engineering, visual simulation and analysis, engineering challenges and opportunities in the genomic era, and nanoscale science and technology. In the Systems Engineering session, the International Space Station, battlefield management, and software development provided the context for the discussion. Here the theme was "managing complexity," in particular, delivering well-engineered and tested products in an era of rapidly shrinking time to market. The three environments described provided some interesting contrasts and similarities. Visual simulation, the second area covered, is the application of ideas from physics, mathematics, and computer science to the production of rich imagery by computer. Applications of visual simulation include entertainment, training, virtual prototyping, industrial design, art, and scientific and information visualization. The speakers in that session amply demonstrated through talks on physically-based animation, data mining and visualization, and multi-resolution methods for modeling, simulation, and visualization that visual simulation is more than just pretty pictures. The four speakers in the Genomics session described the challenges and opportunities brought about by the sequencing of the human genome. Talks covered an overview of the Human Genome Project and the subject of genomics, the characterization of proteins in a small model genome, a bioengineering approach to understanding developments in molecular biology and genomics, and ethical questions generated by advances in genomics. The last talk set the stage for small group discussions on the merits and perils resulting from the development and application of genomic technologies. The symposium concluded with a session on the topic of nanotechnology, which is the science and engineering of making materials, functional structures, and devices on the order of a nanometer scale (1 nanometer = 10^{-9} m). The talks here provided an introduction to the field with presentations on the synthesis, processing, and application of functional nanostructured inorganic particles; carbon-based nanotubes; and nanoscale semiconductor devices. (See Appendixes for complete program.)

As has been done in previous years, a distinguished engineer was invited to address the Frontiers of Engineering participants at dinner on the first evening of the symposium. At the 2000 meeting, Robert Lucky, corporate vice president for applied research at Telcordia Technologies, Inc., spoke about the future of information technology and raised some provocative questions about information technology's limits, intellectual property, and the impact of networking. Dr. Lucky's remarks are contained in this volume.

As part of an ongoing process to make these meetings even more useful to participants, the attendees were asked to evaluate the Frontiers symposium. This feedback once again confirmed the value of the event. Attendees found that being informed about engineering areas with which they were not as familiar was very useful and had the potential to affect their research and technical work. Others noted that the opportunity to interact with engineers from other sectors and disciplines was broadening and inspiring. Many noted that with engineering becoming more interdisciplinary, this meeting filled a unique and much-needed niche in the profession.

The National Academy of Engineering is deeply grateful to the following organizations for their support of the Sixth Annual Symposium on Frontiers of Engineering: Defense Advanced Research Projects Agency, Department of Defense – DDR&E-Research, National Aeronautics and Space Administration, DaimlerChrysler Corporation Fund, Hewlett-Packard Company, Microsoft Corporation, Science Applications International Corporation, Cummins Engine Company, Lockheed Martin Corporation, Lucent Technologies, Parsons Brinckerhoff, Inc., Pratt & Whitney, Sandia National Laboratories, Xerox Foundation, and Xerox Research and Technology. NAE would also like to thank the members of the Symposium Organizing Committee (see p. *iv*), particularly its chair, Michael Corradini, for their work in planning and organizing the event.

Contents

SYSTEMS ENGINEERING

VISUAL SIMULATION AND ANALYSIS

ENGINEERING CHALLENGES AND OPPORTUNITIES IN THE GENOMIC ERA

NANOSCALE SCIENCE AND TECHNOLOGY

DINNER SPEECH

APPENDIXES

SYSTEMS ENGINEERING

Systems Engineering Challenges of the International Space Station

MARK D. JENKS
Boeing Space and Communications Group
Huntsville, Alabama

The development of the International Space Station (ISS) has presented a variety of unique systems engineering challenges. Some are simply extensions of the classical integration issues encountered with any complex development project, but many are unique to ventures of the massive scale and scope of the ISS. Here I address these unique challenges, focusing on areas where traditional thinking may be either of limited value or, in some cases, counterproductive in understanding the critical issues associated with this class of very large-scale, technically and organizationally complex projects.

PROGRAM OVERVIEW

The ISS program may well be the largest single international engineering project in history. It involves the direct participation of 16 nations, 88 launches (37 Shuttle, 9 Proton, 31 Progress, and 11 Soyuz), and 160 space walks (over twice the total number of space walk hours completed by NASA since Ed White's initial walk in 1964). As of September 2000, the station consists of the Russian-built Zarya and Zvezda modules and the U.S.-built Unity Node. The station begins permanent operations with the arrival of the Expedition One crew in November 2000, and, once completed, will provide an unprecedented capability for cooperative international research—both publicly and privately funded.

UNIQUE SYSTEMS ENGINEERING CHALLENGES

While some of the program's challenges have no direct precedent, they are more than simply interesting case studies. The continued globalization of the world economy, the importance of space-based communication and defense systems, and the basic human drive to extend the boundaries of knowledge and achievement all suggest that large, complex multinational enterprises like the ISS will become increasingly important over the decades to come. The primary challenges identified for discussion here are (1) the extended development cycle, (2) unique test and verification requirements and constraints, and (3) the scale and complexity of the required infrastructure.

EXTENDED DEVELOPMENT CYCLE

For the typical commercial development project, minimizing time to market is often the critical challenge, and development cycles are measured in months (or even weeks). In contrast, the challenge of a publicly funded megaproject like the ISS is that of dealing with a protracted development cycle, lengthened by both the scale and scope of the project as well as the funding realities of the political process. Even compared to typical development cycles for complex, integrated vehicles (such as commercial aircraft and defense aerospace systems), the flow for a project the magnitude of the ISS is unprecedented. Whereas the development flow for a major new commercial jetliner is on the order of 5 years, the ISS will have been under development for more than 15 years by the time Phase Two (three-person permanent crew capability) is completed in 2001—the original preliminary design work having been initiated in 1985. After several design concept iterations (many driven by funding changes), the Space Station Freedom configuration was replaced by the Alpha configuration in 1993, and it was again modified to the final ISS configuration—with joint Russian participation fully defined—in 1994.

Technology and hardware obsolescence are among the primary difficulties faced in long-development-cycle projects like the ISS, leading to unique challenges in the areas of personnel turnover, skills and knowledge retention, as well as infrastructure and supplier base erosion. In many high-tech industries (such as internet service providers), the focus is on maintaining cutting-edge skills among a workforce in which turnover is an accepted element of a rapidly expanding industry. In contrast, the larger challenge for the ISS is maintaining key engineers with the legacy design knowledge and development history for the vehicle's various components and subsystems. Funding disruptions and uncertainties in follow-on and sustaining activities further magnify these challenges. These issues are difficult enough to address within the prime contractor organization, but with hundreds of major and subtier contractors producing low-volume, technically sophisticated components (often with little commercial follow-on potential), the challenge of maintaining the personnel, support equipment, and basic production and maintenance capability for all of the critical program hardware becomes a first-order design consideration.

In addition to obsolescence issues and knowledge retention, the length of the development cycle poses an even more fundamental problem: the inability to generate and sustain a comprehensive experience base. Broadly viewed, the history of American human spaceflight consists of three "phases" spanning approximately 40 years: (1) the Apollo-era projects (consisting of the Mercury, Gemini, Apollo, and Skylab programs), (2) the Space Shuttle, and (3) now the ISS. With the resulting frequency of only one or two major "programs" per generation, there is limited opportunity to build a comprehensive experience base, both in terms of knowledgeable engineers and readily accessible engineering data. This operational experience base is significant because, despite the sophistication of analysis techniques and simulation technologies, there remains a substantial element of iterative, "cut and try" optimization in the basic engineering process. Analytical methodologies require correlation, and operational experience nearly always generates significant learning. Whereas the fundamental engineering approach to system development has evolved with an assumption of a robust operational feedback mechanism (the use of extensive field data generated over an extended period under a wide range and combination of operating conditions), human space technologies have not been afforded this opportunity.

As a result, extensive high-fidelity ground-based testing and simulation capabilities, along with an increased focus on capture and analysis of all available operational data, have become essential. Additionally, enormous effort must be expended to ensure first-time success. The cost of failure is often extremely high, and the cost/risk trade is therefore much different in character from most ground-based applications. It is important to note that this paradigm is likely to shift as the ISS becomes permanently occupied. With operational experience will come a reduction in the real risk associated with many follow-on engineering activities, and the immediate cost/risk trade should more often result in selecting less costly options. Assuring that this shift is properly incorporated into the design and verification philosophy will be critical to enhancing the efficiency of future development activities.

TEST AND VERIFICATION CHALLENGES

Testing and verification of the ISS have involved a number of unique requirements and constraints. The major contributing factors include the on-orbit assembly process, the series of vehicle "stages" defining the incremental steps in the assembly process (each with unique functional performance requirements), and the requirement to operate and support human activity under the extreme environmental conditions of space. Add to this that the system is designed, built, and deployed by a consortium of 16 nations, and its parts delivered to their final assembly location by the most complex space vehicle ever built (the Space Shuttle), and the result is arguably the most difficult test and verification program ever attempted.

The operational environment, a zero-gravity vacuum with temperature swings of hundreds of degrees within hours, poses the most obvious verification challenges. Testing under vacuum conditions is most difficult for large elements and complex mechanisms. Two examples are the thermal vacuum testing of the Common Berthing Mechanism (CBM), used to berth the non-Russian pressurized elements of the station, and the leak testing of the U.S. Laboratory. In the case of the CBM, the 20-foot thermal vacuum chamber at Marshall Space Flight Center was used to qualify the active and passive segments of the berthing mechanism under vacuum and a wide range of thermal conditions. A sophisticated resistive load system and software simulating the dynamics of the shuttle and station robotic arms were used to validate the berthing system throughout its operating range. In the case of the U.S. Laboratory leak testing, the entire element was lifted into the large vacuum chamber at Kennedy Space Center, where a low concentration of helium was introduced into the element to determine total leakage to vacuum. Both of these tests involved use of assets developed during the Apollo project, upgraded with advanced analytical and experimental techniques to further extend the precision and accuracy of test results.

The one-g environment poses additional challenges when simulation of zero gravity is impractical. Perhaps the best example is the large solar array wings. Deployment of the wings in a flightlike manner under one-g conditions is impractical. The structure is not designed to withstand gravity loads, and the scale of the system makes simulation of zero gravity via counterbalancing cost prohibitive. In this case a combination of analysis and subcomponent testing is used to verify system performance.

Human thermal vacuum (HTV) testing and neutral buoyancy testing are most commonly used in the development and validation of tasks requiring direct human interaction. HTV testing is typically used to assess the acceptability of planned extravehicular activities (EVAs) that require suited crew members to operate mechanical equipment under thermal vacuum conditions. Operational procedures and human factor issues can be assessed with either actual flight hardware or high-fidelity engineering units to closely simulate the equipment's operational characteristics. Neutral buoyancy testing is used for similar development and validation goals, when zero gravity, rather than thermal vacuum effects, is the critical simulation parameter. Although the mechanical systems simulations are typically of lower fidelity, neutral buoyancy testing allows realistic assessment of zero-gravity operational procedures and is the preferred means for validating large-scale operations and for conducting crew training. Zero-gravity simulation also can be accomplished in a specially outfitted KC-135, which can provide a simulated zero-gravity environment for a series of 30-second windows by flying successive parabolic flight maneuvers.

Another unique element of the ISS verification program is the frequent use of "protoflight" versus conventional qualification and acceptance testing. The conventional hardware verification approach includes qualification testing (in which the basic hardware design is verified) of a dedicated qualification article, along with less severe acceptance testing of the flight units (to verify proper workmanship and quality control during manufacturing). But because of the cost and complexity of many ISS components, a "protoflight" concept has been adopted, whereby only flight units are built. These are then subjected to an intermediate level of testing to both qualify the design and verify workmanship. This approach offers the opportunity for reduced cost and development time, but it also puts a premium on the use of analysis to assure that adequate testing is completed so as to demonstrate performance margins without potentially damaging the flight hardware.

The various vehicle "stages" consist of incremental combinations of the approximately 40 different elements composing the "assembly complete" configuration. (In contrast, the lunar landing missions involved just 3 or 4 vehicle configurations). Each stage has a unique set of functional performance requirements, and, as a result, the system verification process is an order of magnitude more complex than for a typical vehicle. In most cases, the mating elements are joined on orbit, having never previously been assembled. When possible, some elements have been brought together on the ground for integrated testing. For instance, a multi-element integrated test (MEIT) program brought together several of the U.S. flight elements in the Space Station Processing Facility (SSPF) at Kennedy Space Center. Using flight software and simulations of the other interfacing elements, researchers tested critical functions of the station in various assembly stages. In addition, a comprehensive series of cable and hose fit checks were performed to assure proper on-orbit mating of all fluid and electrical connections between station elements.

In cases in which manufacturing schedules (as in the case of the U.S. Node and the Joint Airlock) or geography (as in the case of the U.S. Node and the Russian Zarya module) prevent physical integration and fit checks, advanced digital measurement techniques are employed to determine the as-built configuration of elements that will be physically assembled for the first time on orbit. An extensive Digital Pre-assembly (DPA) program is used to document the as-built geometry of each element and to electronically "assemble" the station elements prior to launch.

INFRASTRUCTURE SCALE AND COMPLEXITY

Anyone who has visited the Kennedy Space Center and seen the Vehicle Assembly Building (VAB)—originally built to process the Saturn V and currently used to assemble the Space Shuttle and its external tanks and solid rocket boosters—and the crawler-transporter—used to move the final assembled vehicle and mobile launch platform from the VAB to the launch pad—has an appreciation for the massive infrastructure requirements of a major space project. The ISS program has benefited substantially from the existence of the Apollo-era infrastructure. Updated and tailored for use with the Shuttle and now the ISS, this infrastructure is critical to execution of the ISS program. In addition, such facilities as the SSPF, built specifically for processing the ISS elements, represent a major new investment in ground handling, support, and test equipment. Mission control centers in Houston and Moscow, along with mission support facilities at the various development sites, represent an equally extensive investment for on-orbit control and operations.

Probably the most complex (and unique) infrastructural elements of the program are the launch vehicles themselves. With 88 total launches (U.S. and Russian) scheduled, assessing the cost and risk associated with the launch infrastructure and incorporation of these elements into basic design trades is extremely difficult. Future programs, particularly those driven by purely commercial incentives, will require an increased focus on the cost and risk associated with infrastructure and support elements of various design options. Minimizing the additional infrastructure requirements and reducing the cost of developing and maintaining that infrastructure will be key to the economic viability of future human space projects.

While minimizing the required asset base is a key focus across many of today's manufacturing industries, properly valuing the existing infrastructure and minimizing additional investment requirements are more complex for multinational projects, particularly those with a legacy of heavy government involvement. As international joint ventures become a mainstay of global business, properly valuing and integrating technology and infrastructure contributions of partners in large-scale international ventures is an issue of significant consequence—beyond the ISS.

SUMMARY

Current trends in macroeconomics, defense systems, and advanced communications seem to suggest that large, complex, multinational enterprises like the ISS will become increasingly important over the decades to come. Here I have outlined a variety of unique systems engineering challenges associated with the development of the ISS, including the protracted development cycle, test and verification requirements, and infrastructure considerations.

The extended development cycle offers major challenges in the areas of technical obsolescence and skills retention. Additionally, the limited number of new programs minimizes the opportunities to build and maintain a comprehensive engineering experience base. With the start of continuous ISS operations in November 2000, this constraint will be softened, offering the opportunity to make a major shift in the basic development paradigm that has existed since the start of human spaceflight.

Verification of a system assembled and operated on orbit requires unique testing technologies and places a premium on advanced simulation and analysis techniques. The facilities and equipment to support this complex verification task, along with the massive infrastructure associated with launch processing and on-orbit operations, represent a huge investment and a major challenge to developing a commercially viable human spaceflight industry.

This paper has focused on the definition and description of these unique challenges, but the more critical job is to extend the lessons of the ISS to better prepare the engineering community for the challenges of similar projects in the years to come.

Battlefield Management

Mark W. Maier
The Aerospace Corporation
Chantilly, Virginia

Battlefield management is the modern term for the thousands-of-years-old practice of controlling men and weapons on the battlefield—what we used to call "command and control." "Battlefield management" is partially a euphemism (an antiseptic take on a violent act) and partially an admission of reality (modern battles are balancing acts of objectives much more complex than "surrender or die"). It is an especially important topic for the United States, because the interaction of technology and national security policy has led to an exceptionally complex real-time control problem on the battlefield. The problem is not at all "classical," in the sense of corresponding to a well-structured optimal control or communications problem; instead, it contains all of the pragmatic and important complexities of most interesting, real problems. It is fundamentally a systems problem—spanning the domains of computer science, operations research, aerospace engineering, and communications engineering.

At the most basic level, there are two battle management architectures (in the sense of organizing structures): (1) hierarchically decentralized and (2) centralized. The hierarchically decentralized structure is the traditional command and control pattern. Commanders give operational orders to their subordinates, each responsible for an overall mission in a given area. As orders pass down, each unit is responsible for finding and engaging enemy targets in its area of responsibility. This system works well for land combat and integrated air/land operations, but much less well for air combat, especially strategic combat over whole theaters. In contrast, large-scale air campaigns are now run in a centralized way, through the air tasking order (ATO). The air campaign commander lays out the targeting plan, and much of the operational planning remains at air campaign headquarters. This is particularly necessary since targets are attacked with complex arrays of weapons (aircraft, cruise missiles, and ballistic missiles) that are launched by different services from widely separated locations, and everything must be synchronized with complex combat support.

One way to break this overall systems challenge into smaller, technically meaningful areas is to look at it in terms of driving realities and representative challenges. Here I discuss the five operational realities and five representative technical challenges.[1]

FIVE OPERATIONAL REALITIES

The nature of military conflict has changed considerably in the last 20 years and is likely to change more in the next 20. The United States gained much of the military leadership it enjoys today through the superiority of its systems engineering. One illustration of this is the extent to which major systems achievements of the 1960s either have not been replicated or have been replicated only in a limited way by other nations, even though the underlying technology has become relatively well known (e.g., the Minuteman ICBM system, submarine-launched ballistic missiles, and global command and control). This is in contrast to what was assumed at the time, when it seemed that whatever had been done by us would be done by others (see, for example, Kahn, 1967). The five operational realities of modern warfare for the United States are

1. strategic air warfare with precision conventional weapons as a primary element of U.S. war policy;
2. the likelihood of high-intensity warfare becoming standoff warfare, with decision time lines becoming shorter than weapon/platform flight times;
3. the political significance of "hide-and-seek" warfare;
4. the primacy of peace enforcement and other prolonged, low-intensity conflicts; and
5. the central place of collaborative systems (systems no one explicitly designs or owns) in military operations.

The 1991 Gulf War (and its continuing military conflict) and the 1999 Kosovo/Serbia conflict share an unexpected characteristic: both were dominantly strategic air wars. A strategic air war is one in which one side attempts to produce a political effect (acquiescence to demands) by attacking the other side's societal infrastructure instead of the immediately adjacent military forces in the field. The Gulf War air campaign was a mixture of strategic and tactical: strategic strikes on political, energy, and industrial targets, mixed with tactical attacks

[1] The opinions expressed herein are solely mine and do not represent the position of The Aerospace Corporation, the United States Air Force, or other agencies of the U.S. government.

on the armies deployed in Kuwait and southern Iraq. The Kosovo/Serbia campaign was even more a strategic campaign; most observers credit attacks on the economic infrastructure of Serbia and the resulting threat to the Serbian leaders' political power as the decisive element. In the Gulf War the air campaign was concluded by a classic air/land campaign. In the Kosovo/Serbia conflict no ground forces were even prepared to conduct such a campaign.

The strategic campaigns conducted by the United States are "effects based." That is, they are used to produce specific and complex effects, such as political acquiescence. These campaigns differ from the older theories of strategic warfare in that the goal of the campaign is not to destroy the enemy nation's capacity to go to war but only to cause a politically determined effect. The challenge, as we will see, is that conducting such a campaign poses complex cognitive problems to intelligence gathering and, thus, greatly complicates the engineering problem of designing intelligence systems. The reality is that strategic air campaigns are back to stay, but they now must be conducted in a carefully measured fashion to produce effects at a degree removed from their direct military effects.

Battle management is commonly associated with high-intensity tactical warfare and missile defense. The critical changes are in the time lines and ranges of engagement. Time lines are not only shorter but are likely to pass through a critical inversion, when the targeting and decision cycle must be less than weapon flight times. The normal method of employing heavy weapons is to find the target and then launch the weapon. But when the weapon flies several hundred kilometers from launch, perhaps launched from an aircraft operating from the continental United States, the find-and-launch strategy can work only against fixed or nearly immobile targets. This is still feasible in forms of strategic warfare, as in the point above, but it becomes impossible in a high-intensity armored campaign or in hide-and-seek warfare.

The third new reality is what can be called "hide-and-seek" warfare. Until the second day of the Gulf War, the basic strategy of a military facing a large air assault had been to directly defend against that assault—that is, the air defense system sought to destroy a sufficient fraction of attacking aircraft so that the air assault could not be sustained long enough to do serious damage to the defending country. In both Iraq and Kosovo, this would have meant expending as many surface-to-air missiles as possible while defending critical targets and keeping the air force in the air, fighting, until shot down. In both cases, at least after the first day, the defending military largely curtailed its attempt to actually stop the air assault and concentrated instead on taking the occasional potshot at moments favorable for downing an attacking aircraft.

The Iraqi Scud missile campaign was similar. The missiles were not fired for military effect (e.g., salvoed at the air bases supporting the attacking fighters) but were, instead, shot in small numbers, spread over weeks, at targets (e.g., Israeli and Saudi cities) intended to produce political effects. There are some similarities between this style of air-ground hide-and-seek and classic guerilla warfare. The challenge to battlefield management is clear. In a classic air/land battle targets may be mobile, but they are out and engaged. The reality of hide-and-seek warfare is that targets are mobile and fleeting, every loss or mis-strike is of large political significance, and both sides are seeking to use their weapons for political effects as much as military.

The fourth new reality is peace enforcement and low-intensity warfare. It is entirely possible that the Gulf War will be the last high-intensity air/land battle for decades. Whether or not high-intensity warfare returns, it seems very likely that long-duration, low-intensity conflicts like Kosovo, Bosnia, Sierra Leone, and Iraq will continue and will be the major focus of engaged U.S. forces. These conflicts have a low but ever-present violence level, are conducted in urban and interurban terrain, have complex rules of engagement, resemble police operations as much as military operations, and are conducted by international coalitions.

The last new reality is that the United States rarely fights alone, and the mix of allies in any given conflict is unpredictable. This means that the management system running the battle comes into existence only when the alliance is formed. No single organization planned it in advance, no unified program office acquired it, and no single commander runs it. It is a collaborative system, in the sense that it is composed of autonomous elements that collaborate voluntarily with retained autonomous management. Using a phrase coined for intelligent transportation systems, they are “systems that no one owns.” We need to account for this reality explicitly in future systems.

FIVE TECHNICAL CHALLENGES

If the preceding are the realities, how can they be cast in technical terms? Five representative technical challenges are

1. making a two-orders-of-magnitude reduction in the planning and execution time line of conventional air operations;
2. replicating small unit operational concepts, including an extension to air operations, to thousand-kilometer theaters of war;
3. building combat target identification systems adequate for hide-and-seek warfare;
4. designing intelligence, surveillance, and reconnaissance systems that can meaningfully support effects-based strategic campaigns; and
5. creating automated management systems without centralized control over either acquisition or operation.

The centerpiece of U.S. air operations has been the ATO, which has been developed on a 72-hour time cycle. The cycle includes the selection of targets and the vetting of them through intelligence processes, the collection of reconnaissance imagery and its dissemination to operational units, detailed flight planning (including ingress synchronization, defense suppression support, and mission support), the briefing of pilots or other weapon operators, platform/weapon flight, and battle damage assessment. The problem is that 72 hours is much too long on a mobile battlefield. In fact, it is easy to create scenarios in which the decision-to-execution time line needs to be 1 hour. This implies a two-orders-of-magnitude reduction in the ATO process, at least for some targets and scenarios.

One reduction approach would be to thoroughly reengineer the existing processes, which would mean automating everything possible and cutting communications delays to near zero. This in itself would pose a very complex systems and software engineering challenge—building a distributed computing system that spanned the globe, had an assurance level acceptable for life-critical operations, was dynamically assembled from multinational components, and was robust against deliberate attack.

Unfortunately, even if it were possible, it is unlikely that this reengineering would effectively achieve the desired cycle-time reduction, for some parts of the current cycle, like flight time, simply cannot be cut very far. Instead, we need to rethink how the process is organized. One alternative command and control architecture is that of small unit operations, both ground and air. For example, close air support operations have never been run on multiday time cycles. They must react to the pace of infantry ground combat, which is rarely longer than hours and is often only minutes. Instead of the elaborate preplanning structure of the ATO, close air support operations use forward air controllers to direct weapon-carrying aircraft, which are launched into holding patterns with very wide target boxes.

Reengineering the ATO process is a good example of the technological problem of systems integration. We want to compose complex information systems from large components (computers, operating systems, existing protocol stacks, object libraries, and so forth) that will possess systemwide properties of security, robustness, time boundedness, and so on. But our engineering methods are not up to the task. We define and can analyze security and real-time behaviors only for relatively simple systems. Our methods do not, as far as we know, scale multimillion lines of code compositions of black box products. Even when we abandon the desire for properties by proof and go for a pragmatic risk management approach, we are not in much better shape. The heuristics for building such systems are limited, and the analytical backing for understanding risk is much less.

If we look at it another way, this is an exceptionally complicated distributed optimization problem. We really desire some assignment of weapons to targets (and reconnaissance resources and combat support resources) that is "optimal." But the assignment problem is distributed in time (varying windows for weapons flyout and target vulnerability) and is stochastic (targets appear and disappear, and things do or do not hit what they were aimed at). The centralized ATO process tries to examine the whole problem in a time window and come up with a block solution. The local control process makes rolling assignments and reassignments divided on geographic control regions. It is probably hopeless to strive for an optimal solution, but we should examine what heuristic and analytic guidance can be given to the designers beyond simulation and trial.

The challenge of hide-and-seek warfare concerns both time lines and identification. The players in these battles may expose themselves anywhere from hours to only a few minutes. Even when operating they are likely to be camouflaged, and the field of battle will often be strewn with decoys and noncombatants that resemble legitimate targets. In "old-fashioned" warfare one often dealt with uncertainty by hitting anything that might be a target and by using area weapons when not exactly sure where to shoot. Hopefully, this will remain unacceptable in future wars.

Targeting intelligence and assessing battle damage concentrate on the things that can be observed: buildings, tanks, or missiles destroyed. But the goal of the conflict typically is not to destroy facilities, tanks, or missiles. Rather, the goal is to induce a political change. So, how do we close the cognitive gap between physical observables and political action? It is here that we really arrive at the divide between technology and politics. Force is still a political tool (indeed, it is more of a political tool now than it perhaps ever was). The goal in building technical systems that support military operations is fundamentally political. Recognizing that we really need to close a cognitive gap is simply recognizing the real purposes to which the technology is being put. This is the challenge of putting technology to human ends.

Lastly, the acquisition and operational problem of systems without owners is also a technological problem. We have successful examples of such systems in everyday operation (e.g. the Internet, open source software), but how can the technical and social mechanisms that make them successful be adapted to multinational military systems? Is it even sensible to try and transfer the lessons over? This problem—the problem of architecting collaborative systems—is not unique to software or military battle management. It occurs in many other applications and must be faced aggressively (Maier, 1998).

CONCLUSIONS

Future battlefield management poses both familiar and unfamiliar challenges in systems engineering. On the familiar side, our development process must produce systems that are acceptably robust, secure, usable, and effective. Less familiar are the complexity and functional richness of the components we must now integrate. Architecturally, we face the problem of coordinated technical, doctrinal, and organizational change. Our technology has changed the ways we fight, and our doctrine is beginning to codify new ways of fighting. This, in turn, demands new organizations to fully exploit. The change to technology, doctrine, and organization should go hand in hand. But their coordination is not a matter of chance—it is a matter of design. At this higher level we must architect organizations and systems at the same time.

REFERENCES

Kahn, H. 1967. The Year 2000: A Framework for Speculation on the Next Thirty-Three Years. New York: Macmillan.

Maier, M. W. 1998. Architecting principles for systems-of-systems. Systems Engineering 1(4):267–284.

ADDITIONAL RECOMMENDED READINGS

Air Force Science Advisory Board. Report on the Joint Battlespace Infosphere. [Online]. Available: www.sab.hq.af.mil.

Alberts, D., J. Gartska, and F. Stein. 1999. Network Centric Warfare. DOD C4ISR Cooperative Research Program. [Online]. Available: www.dodccrp.org.

Krygiel, A. 1999. Behind the Wizard's Curtain: An Integration Environment for a System of Systems. DOD C4ISR Cooperative Research Program. [Online]. Available: www.dodccrp.org.

Maier, M. W., and E. Rechtin. 2000. The Art of Systems Architecting. Second edition. Boca Raton, Fla.: CRC Press.

Rechtin, E. 1999. Systems Architecting of Organizations: Why Eagles Can't Swim. Boca Raton, Fla.: CRC Press.

Software Development at Microsoft

Marvin M. Theimer
Microsoft Research
Redmond, Washington

BIG SOFTWARE SYSTEMS ARE HARD TO BUILD

There are countless examples of software projects that have consumed vast numbers of resources and then been scrapped (Saltzer, 2000). In fact, one study indicates that about 30 percent of all projects get scrapped, while 50 percent are delivered with significant budget overruns, shipping delays, or a significant fraction of their functionality left out (Standish Group, 1995).

Microsoft's business is designing, building, and shipping large software systems and applications. An extreme example is the Windows 2000 operating system, which contains roughly 29 million lines of code and which required the efforts of some 4,000 people to bring to fruition (Freeman, 1999). Most of Microsoft's other products also contain millions of lines of code and have development teams that number in the hundreds.

There are two key factors that make developing software systems so difficult. The first is the complexity of the potential interactions among all components of a system. Complex interactions make it difficult to test or verify that a system meets all the requirements of its design specifications. The second is the high rate of change to which current software systems are subjected. High rates of system evolution imply constant redesigns and reimplementations of many system components.

Complexity of Interactions

The number of possible system configurations an operating system or application may have to run on is huge. This creates a problem for testing that is exacerbated because many of the underlying hardware and third-party software products a system must interact with do not implement required specifications in a fully correct manner. In order to work correctly, the system must be able to work around many of the resulting problems, since a vendor often might be unwilling or unable to change the relevant aspects of their product. And if a vendor is of any size or importance, the alternative of not working with their offerings might be unacceptable.

A second kind of interaction complexity stems from the fact that software engineering technologies are currently unable to encode all the various kinds of dependencies between interacting system components in a systematic, machine-checkable manner. Technologies such as type-safe programming languages have enabled the elimination of certain kinds of programming errors; however, many kinds of dependencies are still only documented as internal coding comments that can, at best, be checked by ad hoc means. Examples include locking conventions for concurrent access to shared data and fault-handling conventions for dealing with exceptional cases, such as out-of-memory errors.

Rapid System Evolution

Exacerbating the problem of building complex systems is rapid system evolution—a fundamental aspect of current high-tech markets, which implies that the requirements for a successful product might change frequently and sometimes dramatically. For example, at Microsoft about 30 percent of the contents of the specification for a particular version of a product typically changes during its development cycle (Cusumano and Selby, 1995).

In addition to market evolution, hardware capabilities consistently double every year or two. A consequence is that, every 3 to 7 years, a system may require fundamental redesign in order to take advantage of the roughly 10x improvement in hardware resources that have become available.

SOFTWARE DEVELOPMENT AT MICROSOFT

Managing Development Projects

Cusumano and Selby (1995) give a good description of the key principles that enable development projects at Microsoft to remain manageable. Paraphrasing, these principles are as follows:

- Have a short vision statement that also defines what a product is not supposed to offer.
- Guide feature selection and prioritization with models and data about user activities.
- Have multiple incremental milestones, buffering between milestones, and integrated development and bug fixing.

- Define a modular, horizontal design architecture with a correspondingly decoupled project structure.
- Control the project's scope via fixed project resources.

One of the keys to successful product management is a means for controlling what features a product should include and in what order they should be implemented. Very early on, projects define a short vision statement, the purpose of which is to give a succinct, coherent description of what the product is and—equally importantly—what the product is not. This statement prevents products from evolving into things that are entirely different from what they were meant to be.

However, the vision statement still allows for a far larger set of possible features than there are developer resources available to implement. The choice of which features to include and how important they are is determined by creating models of user activities. A user activity is defined to be a more or less self-contained activity, such as writing a letter or doing a business financial model. By examining which sets of features are needed for any given user activity, a project can ensure that the set of features actually implemented will allow users to successfully complete some set of their normal activities while avoiding the provision of feature sets that only partially support various activities.

A common problem with development projects is that people do not like to report bad news until the last possible minute. Similarly, if bug testing and fixing get put off until the later stages of a project, nasty surprises may pop up at a time too late to fix. To avoid this, Microsoft projects have multiple incremental milestones and integrated development and bug fixing. That is, products are built in multiple increments, with each incremental version having to reach a level of quality that, in theory, is "ready to ship."

A key aspect of defining a realistic project schedule is the inclusion of time buffers—typically representing about one-third of the total time scheduled—between incremental milestones. These buffers are solely for dealing with unforeseen circumstances, such as the late delivery of a needed system component by another project team.

In order to manage the complexity of interacting system components, as well as of interacting developers, the architecture of a product is structured to be as modular and "horizontally decoupled" as possible. Modules are forced to interact through narrow, well-defined interfaces so that developers can design and implement each module separately and in an independent manner. Each module is assigned to a single developer or to a small team of developers so that there is a well-defined understanding of who is responsible for each piece of the project.

Limiting the size of the team working on a module is of critical importance. Since the overhead of coordination and communication grows with the square of the team size, teams typically consist of only a few people. For example, the entire Windows 2000 file system, which consists of about 200,000 lines of code, was mostly done by a 4-person team. One way to extend the leverage of a small team is to staff it with very good people. The best developers are considered to be as much as 10 times better than an average developer. Consequently, the most difficult parts of a system are typically given to small teams consisting primarily of very senior developers.

Finally, in order to ensure that they do not go on endlessly or grow ever larger, projects are given fixed resources. That is, a project is typically forced to cut functionality rather than acquire additional resources or substantially delay its shipping schedule. This is one of the most important differences between how Microsoft develops software and how various famous failed software projects were run (Saltzer, 2000). A key reason for why this approach is acceptable is Microsoft's strategy of continual incremental upgrades for all of its products: cutting functionality from a project is a much less onerous task when the option exists of providing it in the next product release cycle.

Software Development Strategies

The following principles, described in Cusumano and Selby (1995) and paraphrased here, describe the key software development strategies Microsoft employs:

- Employ many small teams of developers in parallel but require them to synchronize each other's changes on a frequent basis.
- Always have a working version.
- Require everyone to use a common set of tools and have everyone colocated physically.
- Test continuously.

All of these principles represent ways to manage the complexity of interactions. For example, although projects try to structure their architecture as an assembly of more or less independent modules, the interactions between them still must be constantly checked and tested. Hence, a new master version of the product is built and tested daily to check for unexpected problems. Similarly, changes made in each module are propagated to everyone else as soon as possible so that unforeseen interaction problems can be flushed out early.

A key requirement for this style of development is that there always be a working version of the product. Although this might not be possible for the very earliest incremental versions of a new product, it is maintained in all other cases at Microsoft.

Another form of system complexity comes from component interaction requirements that are encoded as coding comments. It is vital to have quick access to the developers of any given component and to avoid communication problems stemming from the use of incompatible tools by different groups and projects. The advantage of colocation of developers using a common set of tools cannot be overemphasized. Consider, for example, when some component is used in a way that was never envisioned by its developers; the only truly effective way to understand any puzzling interactions it might exhibit is to go talk to the developers and perhaps even sit down with them and explore a live example of the problem.

Probably the most important strategy Microsoft uses for software development is continuous testing in as many circumstances as possible. Microsoft employs roughly as many dedicated test personnel as software developers. Testers are paired with developers and work together closely throughout the lifetime of a project.

A key facility that enables large-scale testing is automated test scripts and software tools to aid simplified deployment thereof. Automated test scripts enable arbitrary users to run a variety of regression and stress tests without having to know much about the underlying code. Test scripts are run by "everyone." Limited regression and stress tests are run as "quick" tests by a developer before he or she checks any code into a project's master version. Full tests are run against each new daily build of the master version by as many people as possible, on as many different machine configurations as possible. Even upper management participates.

Both developers and testers use a wide variety of tools to help them test their software. Debugging versions of a product typically contain tens of thousands of checked assertion statements. Various program analysis tools have been developed to detect such things as the use of uninitialized variables. Debugging versions also typically contain code that checks for memory allocation errors and corrupted data structures. Yet another technique used is fault injection. Code is added to a system to artificially cause faults to occur in various subsystems and to produce incorrect input parameters to and output results from called functions.

When a product is nearing completion, it is tested via actual use outside the project but internal to Microsoft. Finally, a beta version is sent out to a large audience of external volunteers. Given the extraordinary number of test cases that can occur for large software systems such as Windows 2000, beta tests involving hundreds of thousands of users are necessary to explore even a fraction of all the possible system configurations that can occur in practice.

LIFE IN THE INTERNET WORLD

Microsoft manages to routinely deliver large, complex software systems and applications. However, with the advent of its .NET Internet initiative, its software development process will be facing some fundamental new challenges. Chief among these will be extremely short development cycles and the requirement that the systems it will be providing be able to run continuously, irrespective of any faults that might occur, any hardware changes that are needed, and any software upgrades that are done.

Short development cycles are the norm in the Internet world: companies adapt their business models to changing market requirements on a frequent basis, and things such as fixes for security holes have to be deployed as soon as possible after someone reports them. This makes extensive testing almost impossible and requires the creation of a new generation of tools that will assist in getting software "right the first time."

Continuous operation will require that developers focus on things like reliability, maintainability, and dynamic scalability, in addition to all their usual concerns. Unfortunately, these are so-called crosscutting issues that tend to affect the design of almost every component in a system.

Web services are also becoming more and more distributed in their implementation, with pieces of an application often running concurrently on several different machines. This will require the creation of an entire new generation of debugging, monitoring, and testing tools that can coordinate and analyze the behavior of activities spanning multiple machines.

REFERENCES

Cusumano, M., and R. Selby. 1995. Microsoft Secrets: How the World's Most Powerful Software Company Creates Technology, Shapes Markets, and Manages People. New York: Free Press.

Freeman, E. 1999. Building gargantuan software: 4,000 programmers do Windows 2000. Scientific American Presents 10(4):28–31.

Saltzer, J. H. 2000. Coping with complexity. Invited talk at the 17th ACM Symposium on Operating Systems Principles, Kiawah, S. C., December 12–15, 1999. A brief summary and pointer to the presentation slides is presented in Operating Systems Review 34(2):7–8.

Standish Group. 1995. CHAOS (Application Project and Failure). Standish Group Study. West Yarmouth, Mass.: Standish Group.

Visual Simulation and Analysis

Physically Based Animation

David Baraff
Pixar Animation Studios
Richmond, California

Animation is a painstaking business. Traditional two-dimensional hand-animated movies require 24 frames of animation per second (or 30, if you animate directly for video), which translates to about 1,500 frames per minute. It was recognized early on that a master animator's time was wasted drawing all those frames; instead, senior animators began drawing characters in key poses at significant points of the action, and more junior animators ("tweeners") became responsible for drawing the characters in the in-between frames.

Roughly the same process happens in computer animation today. An animator produces what are known as "key frames"—that is, positions for a character at a small number of frames—and the computer interpolates smoothly between them (typically with some sort of spline curve). Additionally, the animator is not responsible for directly drawing the geometric shape of his or her character; this portion of the task, known as "rendering," is also done by the computer. Instead, the animator is given a relatively small number of controls with which to manipulate the character. A simplistic model of a human might allow the animator to control one or two degrees of freedom per joint in the human (knees, elbow, hips, neck, and so on) and might result in only a couple dozen control knobs. However, a character that needs to convey real emotional content typically will have hundreds of control knobs, allowing the animator to delicately manipulate facial features, breathing, fingers, and the like. As a result, top-quality computer animation is still an expensive proposition.

Physically based animation had its origins in the early 1980s, with the goal of easing an animator's burden. The initial premise was that animators work very hard trying to make realistic animation, and animation is realistic when it conforms to what we see in the real world; hence, realistic animation must have a lot to do with the physics of our world. If we could build animation systems that worked directly from the principles of physics (Figure 1), then animators clearly could have an easier time animating, and, in many cases, we could dispense with the animators entirely!

INANIMATE ANIMATION?

As it turns out, the initial premise of physically based animation overlooks a key aspect of what animators need to be good at: acting. Consider an action as (seemingly) simple as human walking. Can we use an animation system that understands physics to realistically animate a human walking? The answer, in fact, is yes and no: we can simulate a human walking in some way but not necessarily in the style or manner that we might want for a particular application or scene in a movie. (And walking is, indeed, amazingly complex, both as a movement and as a way of conveying an emotional state—a character's walk can easily convey that he or she is happy, tired, scared, hurt, confused, and so forth.)

Early on, it became obvious that merely running a physical simulation of a system and using the result as animation resulted in "inanimate animation." That is, if you want to produce animation of models or objects that need to seem "alive," then the control aspect of your model is as important (or perhaps even more important) that the underlying physics of the model. Coupling physically based animation with requisite control systems is an ongoing and difficult research problem, but it takes us away from our intended topic.

SECONDARY ANIMATION

A much more fruitful niche for physically based animation has turned out to be secondary motion. There are times when the requirements of an animation will overwhelm any animator, and there is no choice but complete automation of the process. An excellent example of this is clothing in computer animation.

While a complex character might have hundreds of degrees of freedom (DOFs) for an animator to manipulate, moderately realistic (or even stylized) clothing has tens of thousands of DOFs, because of the geometric complexity of even the most modest garment (Figure 2). In addition, there are strong constraints (relationships among the controls of the model) on how clothing behaves under motion in order for it to look like clothing. Using physics to simulate the motion of clothing becomes the obvious (and only) realistic choice. Similarly, automatic physics-based motion of fire, liquids, dust/rain/snow, avalanches, and so on, are all prime targets for physically based animation.

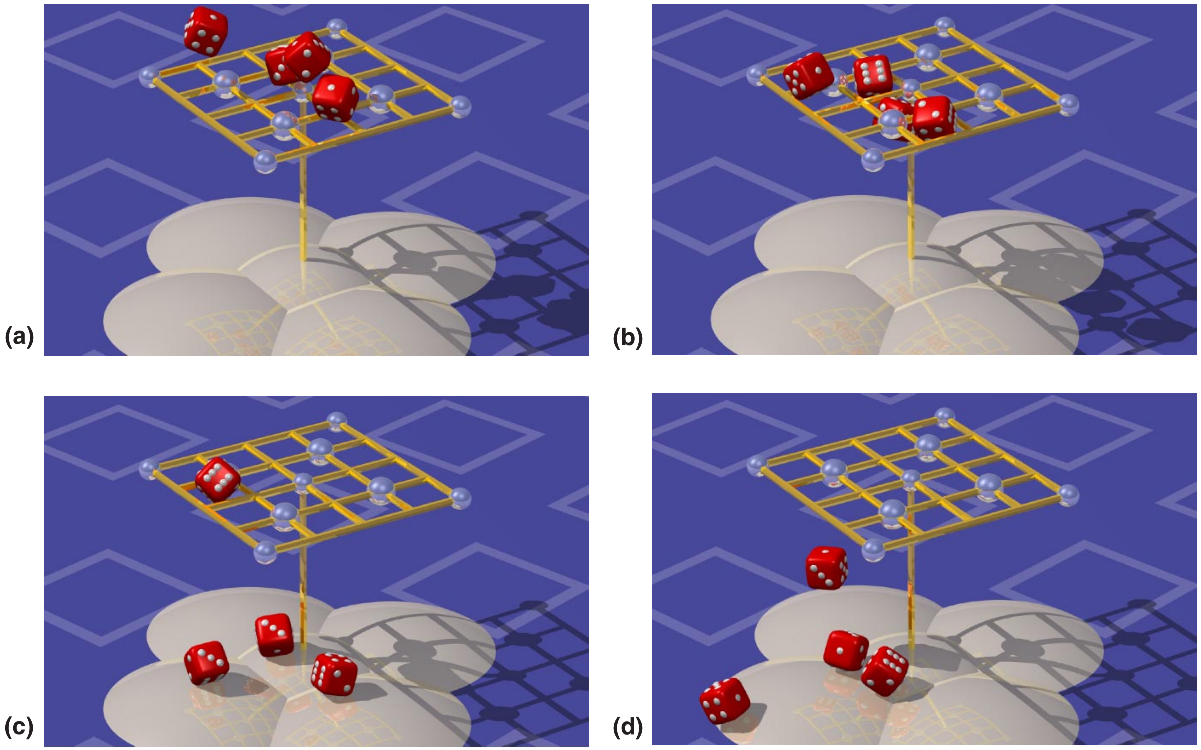


FIGURE 1 Simulation as animation. A pure rigid-body simulation of dice falling through a lattice produces realistic animation without the talents of a skilled animator. SOURCE: ©1990 David Baraff.

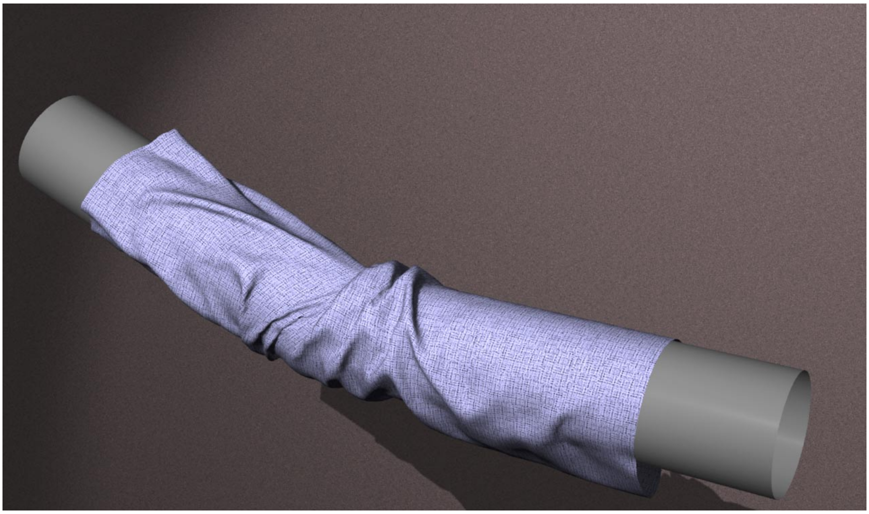


FIGURE 2 Cloth simulation. A single sleeve, simulated as a mesh with more than 14,000 completely independent particles. SOURCE: ©1997 Physical Effects, Inc. Reprinted with permission.

UNDERPINNINGS

It would not be inaccurate to say that physically based animation is, to a large extent, simply physical simulation, provided, however, that one keep in mind that the purpose of the simulation often has a decidedly different flavor from most traditional engineering activities. At a high enough level, physically based animation can be simply described as follows: starting from a description of a system's initial state, and giving a differential equation describing the behavior of the system, we wish to follow the evolution of the system over time. Equivalently, we are solving an initial-valued differential equation. The differential equation, of course, takes its form from the particulars of the world we are trying to animate.

This high-level description obviously hides many details. For example, the differential equation may involve either explicitly or implicitly various constraints on the system over time. The state of the system may be discontinuous (usually not in space, but velocity discontinuities are often fundamental to treatments of rigid bodies in contact). If the number of variables describing the system state is high, we may have to deal with very large systems of coupled equations. In some cases even the inherent computational complexity of the simulation itself may be questionable; for example, some treatments of rigid-body simulation give rise to individual steps for which the solution is NPcomplete! Finally, when it comes to animation, strict realism may not be (and often is not) the desired goal; animation, in the end, exists because of the desire to view the world not as it really is but as the artist chooses it to be. Conformance of a simulation to the animator's world often turns out to be a much harder challenge than conformance to the "real" world.

Data Mining and Visualization

Ronny Kohavi
Blue Martini Software
San Mateo, California

ABSTRACT

Data mining is the process of identifying new patterns and insights in data. As the volume of data collected and stored in databases grows, there is a growing need to provide data summarization (e.g., through visualization), identify important patterns and trends, and act upon the findings. Insight derived from data mining can provide tremendous economic value, often crucial to businesses looking for competitive advantages. A short review of data mining and important theoretical results is provided, followed by recent advances and challenges.

INTRODUCTION

> *Yahoo!'s traffic increased to 680 million page views per day on average. . . . Yahoo!'s communication platform delivered 4.4 billion messages . . . in June [2000].*
>
> —Yahoo! press release, July 11, 2000

The amount of data stored on electronic media is growing exponentially fast. Today's data warehouses dwarf the biggest databases built a decade ago (Kimball and Merz, 2000), and making sense of such data is becoming harder and more challenging. Online retailing in the Internet age, for example, is very different from retailing a decade ago, because the three most important factors of the past (location, location, and location) are irrelevant for online stores.

One of the greatest challenges we face today is making sense of all these data. Data mining—or knowledge discovery—is the process of identifying new patterns and insights in data, whether for understanding the human genome in order to develop new drugs, for discovering new patterns in recent census data in order to warn about hidden trends, or for understanding your customers better at an electronic webstore in order to provide a personalized one-to-one experience. The examples here are from the e-commerce world, but data mining has been used extensively in multiple domains, including many scientific applications. The paper is also restricted to structured mining; significant literature exists on text mining and information retrieval.

The paper is organized as follows. The next section introduces data mining tasks and models, followed by a quick tour of some theoretical results. Next, a review of the recent advances is presented, followed by challenges and a summary.

DATA MINING TASKS AND MODELS

> *The most exciting phrase to hear in science, the one that heralds new discoveries, is not "Eureka!" (I found it!) but "That's funny. . . ."*
>
> —Isaac Asimov

Data mining, sometimes referred to as "knowledge discovery" (Fayyad et al., 1996), is at the intersection of multiple research areas, including machine learning (Dietterich and Shavlik, 1990; Kearns and Vazirani, 1994; Mitchell, 1997; Quinlan, 1993), statistics (Breiman et al., 1984; Friedman et al., in press; Fukunaga, 1990), pattern recognition (Bishop, 1996; Duda et al., 2000; Ripley and Hjort, 1995), databases (Shafer et al., 1996; Srikant and Agrawal, 1995), and visualization (Cleveland, 1993; Tufte, 1983). Good marketing and business-oriented data mining books are also available (Berry and Linoff, 2000; Berson et al., 1999; Dhar, 1997). With the maturity of databases and constant improvements in computational speed, data mining algorithms that were too expensive to execute are now within reach.

Data mining serves two goals:

1. Insight: Identify patterns and trends that are comprehensible so that action can be taken based on the insight. For example, characterize the heavy spenders on a website or people who buy product X. By understanding the underlying patterns, one can personalize and improve the website. The insight may also lead to decisions that affect other channels, such as brick-and-mortar stores' placement of products, marketing efforts, and cross-sells.

2. Prediction: Build a model that predicts (or scores) based on input data. For example, a model can be built to predict the propensity of customers to buy product X based on their demographic data and browsing patterns on a website. Customers with high scores can be used in a direct marketing campaign. If the prediction is for a discrete variable with a few values (e.g., buy product X or not), the task is called "classification"; if the prediction is for a continuous variable (e.g., customer spending in the next year), the task is called "regression."

The majority of research in data mining has concentrated on building the best models for prediction. Part of the reason, no doubt, is that a prediction task is well defined and can be objectively measured on an independent test set. A dataset that is labeled with the correct predictions is split into a training set and a test set. A learning algorithm is given the training set and produces a model that can map new, unseen data into the prediction. The model then can be evaluated for its accuracy in making predictions on the unseen test set. Descriptive data mining, which yields human insight, is harder to evaluate yet necessary in many domains, because the users may not trust predictions coming out of a black box or because, legally, one must explain the predictions. For example, even if a Perceptron algorithm (Minsky and Papert, 1987) outperforms a loan officer in predicting who will default on a loan, the person requesting a loan cannot be rejected simply because he is on the wrong side of a 37-dimensional hyperplane; legally, the loan officer must explain the reason for the rejection.

The choice of a predictive model can have a profound influence on the resulting accuracy and on the ability of humans to gain insight from it. Some models are naturally easier to understand than others. For example, a model consisting of if-then rules is easy to understand, unless the number of rules is too large. Decision trees are also relatively easy to understand. Linear models get a little harder, especially if discrete inputs are used. Nearest-neighbor algorithms in high dimensions are almost impossible for users to understand, and nonlinear models in high dimensions, such as neural networks, are the most opaque.

One way to aid users in understanding the models is to visualize them. MineSet (Brunk et al., 1997), for example, is a data mining tool that integrates data mining and visualization very tightly. Models built can be viewed and interacted with. (Several movies are available at http://www.sgi.com/software/mineset/demos.html.) Figure 1 shows a visualization of the Naive-Bayes classifier. Given a target value—in this case those in the U.S. working population who earn more than $50,000—the visualization shows a small set of "important" attributes (measured using mutual information or cross-entropy). For each attribute a bar chart shows how much "evidence" each value (or range of values) of that attribute provides for the target label. For example, higher education levels (right bars in the education row) imply higher salaries because the bars are higher. Similarly, salary increases with age up to a point and then decreases, and salary increases with the number of hours worked per week. The combination of a back-end algorithm that bins the data and computes the importance of hundreds of attributes and a visualization that shows the important attributes makes this a very useful tool that helps identify patterns. Users can interact with the model by clicking on attribute values and seeing the predictions the model makes.

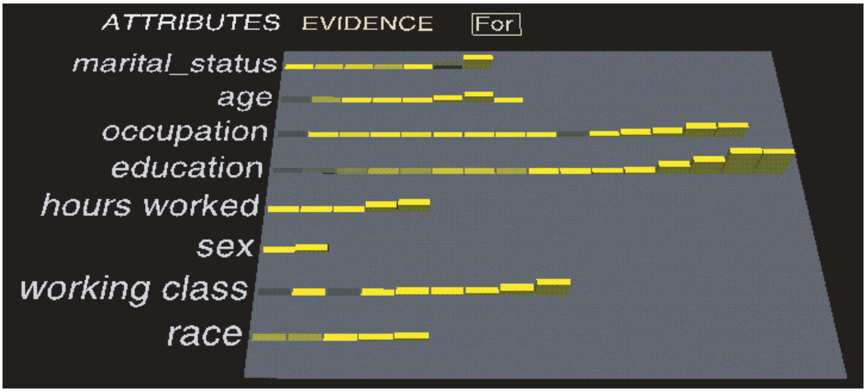


FIGURE 1 A visualization of the Naive-Bayes classifier. SOURCE: Reprinted with permission from AAAI Press (Brunk et al., 1997).

DATA MINING THEORY

> *Reality is the murder of a beautiful theory by a gang of ugly facts.*
>
> —Robert L. Glass (1996)

The following is a short review of some theoretical results in data mining.

- *No free lunch*. A fundamental observation is that learning is impossible without assumptions. If all concepts are equally likely, then not only is learning impossible but no algorithm can dominate another in generalization accuracy (Schaffer, 1994; Wolpert, 1995). The result is similar to the proof that data compression is not always possible (yet everyone enjoys the savings provided by data compression algorithms). In practice, learning is very useful because the world does not present us with uniform worst-case scenarios.
- *Consistency*. While parametric models (e.g., linear regression) are known to be of limited power, nonparametric models can be shown to learn "any reasonable" target concept, given enough data. For example, nearest-neighbor algorithms with a growing neighborhood have been shown to have asymptotically optimal properties under mild assumptions (Fix and Hodges, 1951). Similar results exist for the consistency of decision tree algorithms (Gordon and Olshen, 1984). While asymptotic consistency theorems are comforting because they guarantee that with enough data the learning algorithms will converge to the target concept one is trying to learn, our world is not so ideal. We are always given finite amounts of data from which to learn, and rarely do we reach *asymptopia*.

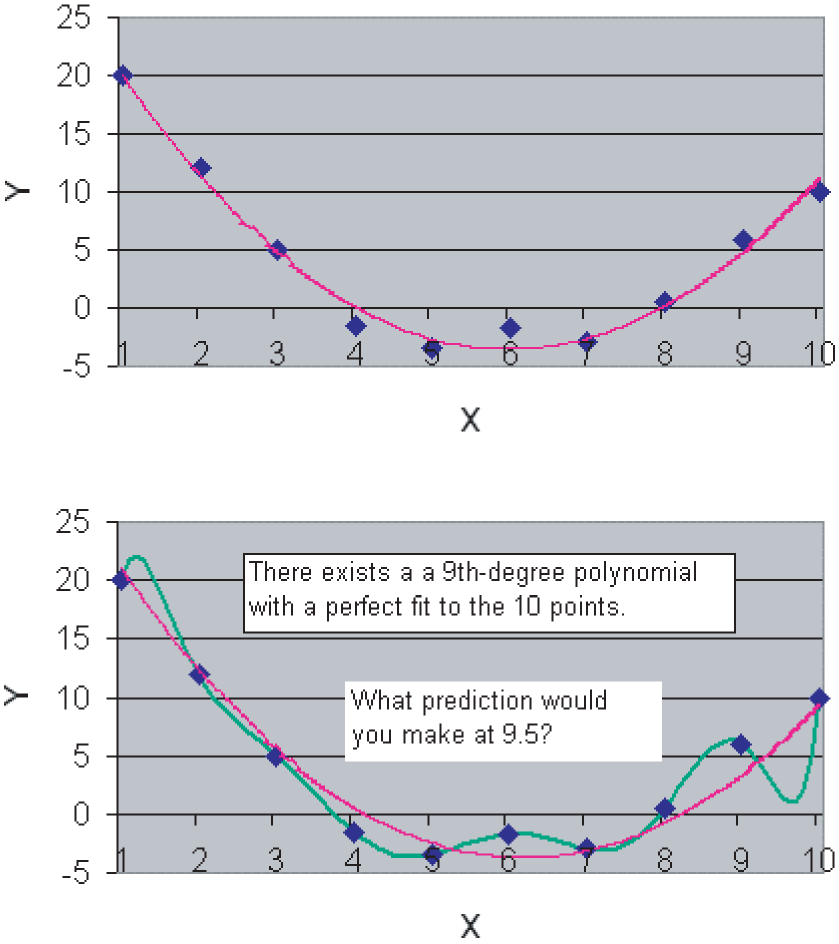
An excellent example of the problem of nearest neighbors not being so "near" is as follows (Friedman et al., in press). Assume a 20-dimensional unit ball (radius = 1) centered at the origin with 100,000 points uniformly distributed. The median distance from the origin to the closest point is 0.55, more than halfway to the boundary. Most points, therefore, are closer to the boundary of the sample space than to another point! In a few dimensions, standard visualization methods work well; in higher dimensions our intuition is commonly wrong, and data mining can help.

- *PAC learning*. Probably Approximately Correct (PAC) learning (Kearns and Vazirani, 1994; Valiant, 1984) is a concept introduced to provide guarantees about learning. Briefly, assuming that the target can be described in a given hypothesis space (e.g., disjunctions of conjunctions of length k), a PAC learning algorithm can learn the approximate target with high probability. The two parameters typically given as input to a PAC learning algorithm are ε and δ. The algorithm must satisfy the condition that at least $(1-\delta)$ fraction of the time, the error between the actual target concept and the predictions made is bounded by ε. PAC learning theory defines bounds on the number of examples needed to provide such guarantees.

One of the more interesting results in PAC learning theory is that a weak learning algorithm, which can classify more accurately than random guessing (e.g., $\varepsilon < 0.5$), can always be boosted into a strong learning algorithm, which can produce classifiers of arbitrary accuracy (Schapire, 1990). (More training data will be needed, of course.) This theoretical result has led to interesting practical developments, mentioned below.

- *Bias-variance decomposition*. The expected error of any learning algorithm for a given target concept and training set size can be decomposed into two terms: the bias and the variance (Geman et al., 1992). The importance of the decomposition is that it is valid for finite training set sizes—not asymptotically—and that the terms can be measured experimentally. The bias measures how closely the learning algorithm's average guess (over all possible training sets of the given training set size) matches the target. The variance measures how much the learning algorithm's guess varies for different training sets of the given size. Researchers have taken many unsuccessful and painful routes trying to improve a learning algorithm by enlarging the space of models, which can reduce the bias but may also increase the variance. For example, Figure 2 shows how 10 data points, assumed to be slightly noisy, can be reasonably fit with a quadratic polynomial and perfectly fit with a ninth-degree polynomial that overfits the data. A learning algorithm trying to fit high-degree polynomials will generate very different polynomials for different training sets and, hence, have high variance. A learning algorithm that always fits a linear model will be more stable but will be biased for quadratic and higher-order models.

Making the analogy to decision tree models, finding the smallest decision tree that perfectly fits the data (an NP-hard problem) takes a long time and often results in worse generalizations than using a simple greedy algorithm that approximately fits the data. The reason is that the smallest perfect trees generat-



There exists a a 9th-degree polynomial with a perfect fit to the 10 points.

What prediction would you make at 9.5?

FIGURE 2 The upper figure shows a quadratic fit to data, but the fit is not perfect. The lower figure shows a ninth-degree polynomial fit that perfectly passes through all the data points. If the data are expected to contain some noise, the model on the top will probably make a better prediction at x = 9.5 than the squiggly model on the bottom, which (over) fits the data perfectly.

ed for similar data sets of the same size vary significantly in structure and predictions, and, hence, the expected error has a large variance term. For several algorithms it is known how to move along this bias-variance tradeoff through regularization techniques.

RECENT ADVANCES

> *The advancement of the arts, from year to year, taxes our credulity and seems to presage the arrival of that period when human improvement must end.*
> —Henry Elsworth, U.S. Patent Office (1844)

Below is a brief summary of recent advances in the field of data mining. A description of several advances specific to machine learning can be found in *AI Magazine* (Dietterich, 1997).

- *Multiple model learning*. Two learning techniques developed in the last few years have had a significant impact: Bagging and Boosting. Both methods learn multiple models and vote them in order to make a prediction, and both have been shown to be very successful in improving prediction accuracy on real data (Bauer and Kohavi, 1999; Quinlan, 1996). Bagging (Breiman, 1996) generates bootstrap samples by repeatedly sampling the training set with replacement. A model is built for each sample, and they are then uniformly voted. Boosting algorithms—and specifically the AdaBoost algorithm (Freund and Schapire, 1997)—generate a set of classifiers in sequence. Each classifier is given a training set in which examples are reweighted to highlight those previously misclassified.
- *Associations*. A common problem in retailing is to find combinations of products that, when bought together, imply the purchase of another product. For example, an association might be that the purchase of hotdogs and a Coke® implies the purchase of chips with high probability. Several algorithms were developed to find such associations for market basket analysis (Srikant and Agrawal, 1995). Given a minimum support (percentage of the data that has to satisfy the rule) and a minimum confidence (the probability that the right-hand side is satisfied given the left-hand side), the algorithms find all associations. Note that unlike prediction tasks, this is a descriptive task where the result is well defined and the algorithms must be sound and complete. The main observation in these algorithms is that in order for a combination of size L to have minimum support, each of its subsets of size (L – 1) must have minimum support.
- *Scalability (both speed and data set size)*. Several advances have been made in scaling algorithms to larger data sets and parallelizing them (Freitas and Lavington, 1997; Provost and Kolluri, 1999; Shafer et al., 1996).

CHALLENGES

Laggards follow the path of greatest familiarity. Challengers, on the other hand, follow the path of greatest opportunity, wherever it leads.
—Gary Hamel and C. K. Prahalad (1994)

Selected challenging problems are described below.

- *Make data mining models comprehensible to business users.* Business users need to understand the results of data mining. Few data mining models are easy to understand, and techniques need to be developed to explain or visualize existing ones (e.g., Becker et al., 1997), or new models that are simple to understand with matching algorithms need to be derived. This is particularly hard for regression models. A related problem is that association algorithms usually derive too many rules (e.g., 100,000), and we need to find ways to highlight the "interesting" rules or families of associations.
- *Make data transformations and model building accessible to business users.* An important issue not mentioned above is the need to translate users' questions into a data mining problem in relational format. This often requires writing SQL, Perl scripts, or small programs. Even defining *what* the desired transformations and features should be is a knowledge-intensive task requiring significant understanding of the tasks, the algorithms, and their capabilities. Can we design a transformation language more accessible to business users? Can we automatically transform the data?
- *Scale algorithms to large volumes of data.* It has been estimated that the amount of text in the Library of Congress can be stored in about 17 terabytes of disk space (Berry and Linoff, 2000). The package-level detail database used to track shipments at UPS is also 17 terabytes. Most data mining algorithms can handle a few gigabytes of data at best, so there are three to four orders of magnitude to grow before we can attack the largest databases that exist today. In addition, most algorithms learn in batch, but many applications require real-time learning.
- *Close the loop: identify causality, suggest actions, and measure their effect.* Discoveries may reveal correlations that are not causal. For example, human reading ability correlates with shoe size, but wearing larger shoes will not improve one's reading ability. (The correlation is explained by the fact that children have smaller shoe sizes and cannot read as well.) Controlled experiments and measurements of their effects can help pinpoint the causal relationships. One advantage of the online world is that experiments are easy to conduct: changing layout, emphasizing certain items, and offering cross-sells can all be done easily and their effects can be measured. For electronic commerce, the World Wide Web is a great laboratory for experiments, but our learning techniques need to be improved to offer interventions and take them into account.

- *Cope with privacy issues*. Data mining holds the promise of reducing the amount of junk mail we receive by providing us with more targeted messages. However, data collection can also lead to abuses of the data, raising many social and economic issues. This is doubly true in the online world, where every page and every selection we make can be recorded.

SUMMARY

> *You press the button, and we'll do the rest.*
> —Kodak advertisement

Taking pictures and developing them (or loading them into a computer) has become a trivial task: there is no need to focus, adjust the shutter speed and aperture, or know anything about chemistry to take great pictures. Data mining and related technologies have had significant advances, but we have yet to build the equivalent of the point-and-click cameras. This short review of the basic goals of data mining, some theory, and recent advances should provide those interested with enough information to see the value of data mining and use it to find nuggets; after all, almost everyone has access to the main ingredient needed: data.

ACKNOWLEDGMENTS

I thank Carla Brodley, Tom Dietterich, Pedro Domingos, Rajesh Parekh, Ross Quinlan, and Zijian Zheng for their comments and suggestions.

REFERENCES

Bauer, E., and R. Kohavi. 1999. An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. Machine Learning 36:105–139.

Becker, B., R. Kohavi, and D. Sommerfield. 1997. Visualizing the simple Bayesian classifier. Presented at KDD Workshop on Issues in the Integration of Data Mining and Data Visualization, Newport Beach, Calif., August 17, 1997.

Berry, M. J. A., and G. Linoff. 2000. Mastering Data Mining: The Art and Science of Customer Relationship Management. New York: John Wiley & Sons.

Berson, A., K. Thearling, and S. J. Smith. 1999. Building Data Mining Applications for CRM. New York: McGraw-Hill.

Bishop, C. M. 1996. Neural Networks for Pattern Recognition. New York: Oxford University Press.

Breiman, L. 1996. Bagging predictors. Machine Learning 24:123–140.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone. 1984. Classification and Regression Trees. Belmont, Calif.: Wadsworth International Group.

Brunk, C., J. Kelly, and R. Kohavi. 1997. MineSet: An integrated system for data mining. Pp. 135–138 in Proceedings of the Third International Conference on Knowledge Discovery and Data Mining, D. Heckerman, H. Mannila, D. Pregibon, and R. Uthurusamy, eds. Cambridge, Mass.: AAAI Press. Website: http://mineset.sgi.com.

Cleveland, W. S. 1993. Visualizing Data. Summit, N.J.: Hobart Press.

Dhar, V. 1997. Seven Methods for Transforming Corporate Data into Business Intelligence. New York: Prentice–Hall.

Dietterich, T. G. 1997. Machine-learning research: Four current directions. AI Magazine 18(4):97–136.

Dietterich, T. G., and J. W. Shavlik, eds. 1990. Readings in Machine Learning. San Francisco: Morgan Kaufmann.

Duda, R. O., P. E. Hart, and D. G. Stork. 2000. Pattern Classification. New York: John Wiley & Sons.

Fayyad, U. M., G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, eds. 1996. From data mining to knowledge discovery: An overview. Pp. 1–34 in Advances in Knowledge Discovery and Data Mining. Cambridge, Mass.: MIT Press and the AAAI Press.

Fix, E., and J. Hodges. 1951. Discriminatory Analysis—Nonparametric Discrimination: Consistency Properties. Technical Report 21-49-004, Report No. 04. Randolph Field, Tex.: USAF School of Aviation Medicine.

Freitas, A. A., and S. H. Lavington. 1997. Mining Very Large Databases With Parallel Processing. New York: Kluwer Academic Publishers.

Freund, Y., and R. E. Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences 55(1):119–139.

Friedman, J., T. Hastie, and R. Tibshirani. In Press. The Elements of Statistical Learning: Prediction, Inference and Data Mining.

Fukunaga, K. 1990. Introduction to Statistical Pattern Recognition. San Diego: Academic Press.

Geman, S., E. Bienenstock, and R. Doursat. 1992. Neural networks and the bias/variance dilemma. Neural Computation 4:1–48.

Glass, R. 1996. The relationship between theory and practice in software engineering. Communications of the ACM 39(11):11–13.

Gordon, L., and R. A. Olshen. 1984. Almost sure consistent nonparametric regression from recursive partitioning schemes. Journal of Multivariate Analysis 15:147–163.

Hamel, G., and C. K. Prahalad. 1994. Competing for the Future. Boston: Harvard Business School Press.

Kearns, M. J., and U. V. Vazirani. 1994. An Introduction to Computational Learning Theory. Cambridge, Mass.: MIT Press.

Kimball, R., and R. Merz. 2000. The Data Webhouse Toolkit: Building the Web-Enabled Data Warehouse. New York: John Wiley & Sons.

Minsky, M. L., and S. Papert. 1987. Perceptrons—Expanded Edition: An Introduction to Computational Geometry. Cambridge, Mass.: MIT Press.

Mitchell, T. M. 1997. Machine Learning. New York: McGraw-Hill.

Provost, F., and V. Kolluri. 1999. A survey of methods for scaling up inductive algorithms. Data Mining and Knowledge Discovery 3(2):131–169.

Quinlan, J. R. 1993. C4.5: Programs for Machine Learning. San Francisco: Morgan Kaufmann.

Quinlan, J. R. 1996. Bagging, boosting, and C4.5. Pp. 725–730 in Proceedings of the 13th National Conference on Artificial Intelligence. Cambridge, Mass.: MIT Press and the AAAI Press.

Ripley, B. D., and N. L. Hjort. 1995. Pattern Recognition and Neural Networks. New York: Cambridge University Press.

Schaffer, C. 1994. A conservation law for generalization performance. Pp. 259–265 in Machine Learning: Proceedings of the Eleventh International Conference. San Francisco: Morgan Kaufmann.

Schapire, R. E. 1990. The strength of weak learnability. Machine Learning 5(2):197–227.

Shafer, J., R. Agrawal, and M. Mehta. 1996. SPRINT: A scalable parallel classifier for data mining. Pp. 544-555 in Proceedings of the 22nd International Conference on Very Large Databases, T. M. Vijayaraman, A. P. Buchmann, C. Mohan, and N. L. Sarda, eds. San Francisco: Morgan Kaufmann.

Srikant, R., and R. Agrawal. 1995. Mining generalized association rules. Pp. 407–419 in Proceedings of the 21st International Conference on Very Large Databases, U. Dayal, P. M. D. Gray, and S. Nishio, eds. San Francisco: Morgan Kaufmann.

Tufte, E. R. 1983. The Visual Display of Quantitative Information. Cheshire, Conn.: Graphics Press.

Valiant, L. G. 1984. A theory of the learnable. Communications of the ACM 27:1134–1142.

Wolpert, D. H., ed. 1995. The relationship between PAC, the statistical physics framework, the Bayesian framework, and the VC framework. The Mathematics of Generalization: The Proceedings of the SFI/CNLS Workshop on Formal Approaches to Supervised Learning. Reading, Mass.: Addison-Wesley.

Digital Geometry Processing

Peter Schröder
Department of Computer Science
California Institute of Technology
Pasadena, California

Wim Sweldens
Lucent Technologies
Murray Hill, New Jersey

There have been three waves of multimedia so far: sound, images, and video. At present we are witnessing the arrival of the fourth wave of multimedia: three-dimensional geometry. Each of these waves was initiated by increases in acquisition capabilities, compute power, storage capacity, and transmission bandwidth for the respective type of data. As Moore's law moved along, we first saw digital sound in the 1970s, digital images in the 1980s, and digital video in the 1990s. An average PC will soon be able to handle digital geometry.

Each new digitization wave brings with it the need for new processing tools. Typical elements of a signal processing toolbox are denoising, compression, transmission, enhancement, detection, analysis, editing, and so forth. While analog circuitry can handle only the most basic processing tasks, as soon as the data are digitized, a whole new realm of algorithms becomes feasible. This has led to an explosion in digital signal processing research, aimed at the development of suitable mathematical representations, their manipulation, and associated computational paradigms. One example is the ubiquitous use of digital sound, from portable CD players to musical instruments and digital cellular phones. More recently, increasing network bandwidth and compute power have contributed to a revolution in the distribution of music over networks and on personal computers. Similar observations can be made about images and video.

In a sense, the cheap and plentiful availability of a data type has led to a spur in the development of methods to wring usefulness from these data, which, in turn, has stimulated progress in the underlying technology to support the respective type of data.

In a very similar way we are now witnessing the arrival of a new data type: three-dimensional geometry. To be sure, geometric modeling has been around for a long time, but geometry was created "by hand," in a tedious custom process. Strides in the acquisition of three-dimensional geometry through low- and high-end three-dimensional scanners are now making digital proxies of geometry available for processing in a much broader way. Example sources of such geometry range from the sculptures of Michelangelo[1] to manufactured products[2] and the earth itself.[3]

Once again, we need to build a toolbox of fundamental algorithms and mathematics to process this new class of digital geometry data. This time the task is significantly more challenging than before.

FOURIER OR NOT FOURIER

Multimedia data of the first three waves can be modeled easily as being defined on a section of Euclidean geometry. Sound is defined as a function of time—that is, on a one-dimensional line. Similarly, images are naturally defined as functions over a section of a two-dimensional plane. Finally, video is most naturally modeled as a function over a section of three-dimensional space—two dimensions in space and one in time. Another useful abstraction is that of sampling. A given datatype is digitized by sampling it at regular intervals (either space or time) and converting the measured values into binary numbers. The act of sampling and the regular spacing between samples make Fourier analysis and the Fourier transform the quintessential tools to understand and manipulate the content of such signals. If the underlying mathematical space were not Euclidean, the regular spacing of samples could not be achieved. For example, consider a sphere. There are no equispaced sampling patterns on the sphere beyond those given by the five Platonic solids. A digital image produced by one of today’s cameras, however, is a regular matrix of picture elements, each representing a sample of the image irradience on the two-dimensional image plane.

This regular sampling allows the computation of the Fast Fourier Transform (FFT), one of the most popular algorithms in digital signal processing. Once in the Fourier domain, it is easy to remove noise, for example, or to enhance the image.[4] Recently, new multiresolution and time-frequency-based methods such as wavelets have proven to be a valuable alternative to the Fourier transform in many applications. These methods introduce the notion of scale into the analysis. For example, some phenomena being measured may occur at a broader scale while others appear at a finer scale. Adding this element allows us to zoom in and out of particular features in the data, yet even these methods still rely on the Fourier transform for their design and analysis.

[1] Digital Michelangelo Project, Stanford University; models with more than 10^9 samples.

[2] Reverse engineering and noninvasive inspection.

[3] See the recent shuttle radar topography mission, for example.

[4] In practice, algorithms often do not perform an actual Fourier transform to achieve these effects. Nevertheless, the Fourier transform is essential for understanding how to design and analyze such algorithms.

Geometry, however, relies in an essential way on a non-Euclidean setting: curves, surfaces, and, more generally, manifolds. Unfortunately, curved geometry does not readily admit a Fourier transform, let alone a fast transform algorithm. Additional complications arise in software and hardware implementations. What used to be simple arrays or streams of regularly arranged data is now a complex topological data structure requiring much more sophisticated algorithms to handle. In general, there is no way around these difficulties when the underlying geometry is not flat. For example, consider data defined on a sphere, such as measurements that are a function of direction. One could map the sphere to a section of the plane and then use regular sampling and Fourier transforms. It is well known, however, that no such mapping can avoid singularities resulting in uncontrollable artifacts. Instead, one needs to take the essential nature of the sphere into account. These arguments apply even more so when the underlying surface is more complicated.

Luckily, some recent techniques inspired by wavelet constructions and commonly referred to as "multiresolution algorithms" offer a basis for the development of a digital geometry processing toolbox. These techniques, developed at the intersection of mathematics (wavelet analysis) and computer science (graphics), are based on subdivision (Zorin and Schröder, 2000), lifting (Sweldens, 1997), and second-generation wavelets (Daubechies et al., 1999) and carry over to the curved geometry setting.

THE POWER OF SEMIREGULAR MESHES

We have seen that it is impossible to keep the regularity—that is, the regular Cartesian arrangement of samples—of the Euclidean setting when going to surfaces. However, we can do almost as well using so-called semiregular meshes. These are constructed through a process of recursive quadrisection, an idea that originated in the area of subdivision. In a coarse mesh consisting of a generally small number of triangles, each triangle is recursively split into four subtriangles. In standard subdivision, where this approach is used to produce smooth surfaces, the new point positions are computed based on local averaging. A few of the steps in such a sequence are shown in Figure 1. Instead, we use point positions that are samples of the desired geometry, in effect adding displacements—or wavelet details—to a subdivision surface. This combination of subdivision and details, or wavelets, can be compared to the Euclidean geometry signal processing setting with its low- and high-pass filters. Although the results from that setting are not immediately applicable to the surface setting, they do provide guidance.

Because of the recursive quadrisection process by which these sampling patterns are built, we can build fast hierarchical transforms. These generalize the fast wavelet transform to the surface setting. While it is much harder to prove smoothness and approximation properties in this more general setting, first re-

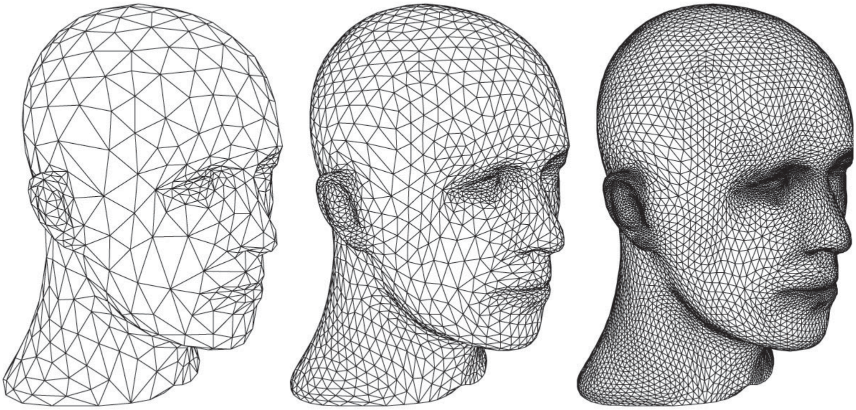


FIGURE 1 Example of subdivision for a surface, showing three successive levels of refinement. On the left is an initial triangle mesh approximating the surface. Each triangle is split into four according to a particular subdivision rule (middle). On the right the mesh is subdivided in this fashion once again.

sults exist and are extremely encouraging. For example, these constructions can be used for very efficient progressive transmission algorithms. Imagine a file format for geometry that has the property that the first bits in the file provide a rough outline of the shape, and as more bits arrive, more and more details of the shape are revealed.

CONCLUSIONS

The fourth wave of multimedia—that of geometry—creates new mathematical and algorithmic challenges that cannot be answered with straightforward extensions of signal processing techniques from the Euclidean setting. However, ideas from multiresolution, subdivision, and second-generation wavelets provide the foundation of a new digital geometry processing apparatus based on semiregular meshes.

REFERENCES

Daubechies, I., I. Guskov, P. Schröder, and W. Sweldens. 1999. Wavelets on irregular point sets. Philosophical Transactions of the Royal Society of London A 357:2397–2413.

Sweldens, W. 1997. The lifting scheme: A construction of second generation wavelets. SIAM Journal on Mathematical Analysis 29(2):511–546.

Zorin, D., and P. Schröder, eds. 2000. Subdivision for modeling and animation. Course notes from the 27th International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH), New Orleans, July 23–28, 2000.

Engineering Challenges and Opportunities in the Genomic Era

The Human Genome Project: Elucidating Our Genetic Blueprint

Eric D. Green
National Human Genome Research Institute
National Institutes of Health
Bethesda, Maryland

The human genome consists of approximately 3 billion base pairs (bp) of DNA contained within 24 chromosomes that range in size from approximately 50 million to 260 million bp. Encoded within this DNA are an estimated 25,000 to 75,000 genes and the necessary elements that control the regulation of their expression. A listing of the sequence of the human genome's bases in single-letter symbols (G, A, T, C) would fill about 13 sets of the *Encyclopaedia Britannica*, approximately 750 megabytes of computer disk space, or roughly 1 CD-ROM.

THE HUMAN GENOME PROJECT

The Human Genome Project is a coordinated, international effort to map and sequence the human genome and, in parallel, that of several well-studied model organisms. In the United States, the National Institutes of Health (www.nhgri.nih.gov) and the Department of Energy (jgi.doe.gov) orchestrate the relevant programs, with the project officially beginning on October 1, 1990. Since its inception, the Human Genome Project has been associated with carefully crafted, milestone-oriented goals, the most recent set being established in 1998 (Collins et al., 1998).

The first phase of the project involved constructing relatively low-resolution maps of the human genome and refining the approaches for large-scale DNA sequencing. The second phase has focused on establishing the complete sequence of the human and other vertebrate genomes, as well as beginning to decipher the encoded information in a systematic fashion.

MAPPING THE HUMAN GENOME

Genomic maps are organizational representations that reflect specific features of the corresponding DNA. There are three major classes of genomic maps: cytogenetic, genetic, and physical (Figure 1). In each case, the coordinates on which the maps are based reflect the experimental method(s) used to establish the relative positions of landmarks.

- *Cytogenetic maps.* Cytogenetic maps represent the appearance of chromosomes when properly stained and examined microscopically. Particularly important is the appearance of differentially staining regions (called bands) that render each chromosome uniquely identifiable.

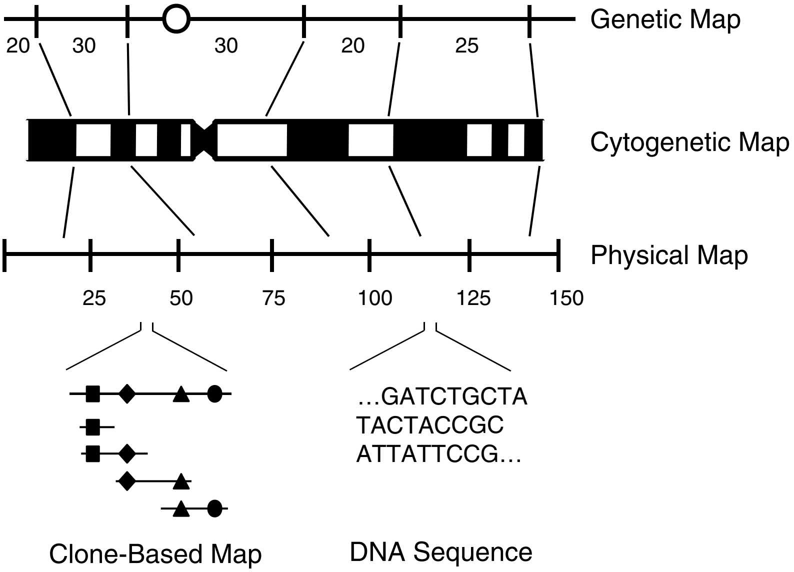


FIGURE 1 Schematic representations of the genetic, cytogenetic, and physical maps of a human chromosome. For the genetic map, the positions of several hypothetical genetic markers are indicated, along with the relative genetic distances between them. The circle indicates the position of the centromere. For the cytogenetic map, the classic Giemsa-banding pattern of a chromosome is shown. For the physical map, the approximate physical locations of the above genetic markers are indicated, along with the relative distances between them in millions of bp. Two types of physical maps—an overlapping clone-based map and the highest-resolution physical map, the actual DNA sequence—are depicted along the bottom. SOURCE: Reprinted with permission from The McGraw-Hill Companies (Green, 2001).

- *Genetic maps.* Also known as linkage maps or meiotic maps, genetic maps depict the relative locations of genetic markers across a stretch of DNA. Genetic markers reflect sequences that vary among different individuals (i.e., they are polymorphic). Such sequence variation allows genetic markers to be followed during passage from one generation to the next.
- *Physical Maps.* Physical maps depict the relative locations of physical landmarks across a stretch of DNA. The conventional approach for physical mapping involves fragmenting and cloning the DNA, and then reassembling the pieces in an organized fashion. A collection of ordered, overlapping clones is called a contig, since the clones together contain a contiguous segment of DNA.

The development of high-quality genetic and physical maps of the human genome represented key early activities of the Human Genome Project. These goals were largely reached by the mid-1990s (Figure 2). Attention then turned to sequencing the human genome.

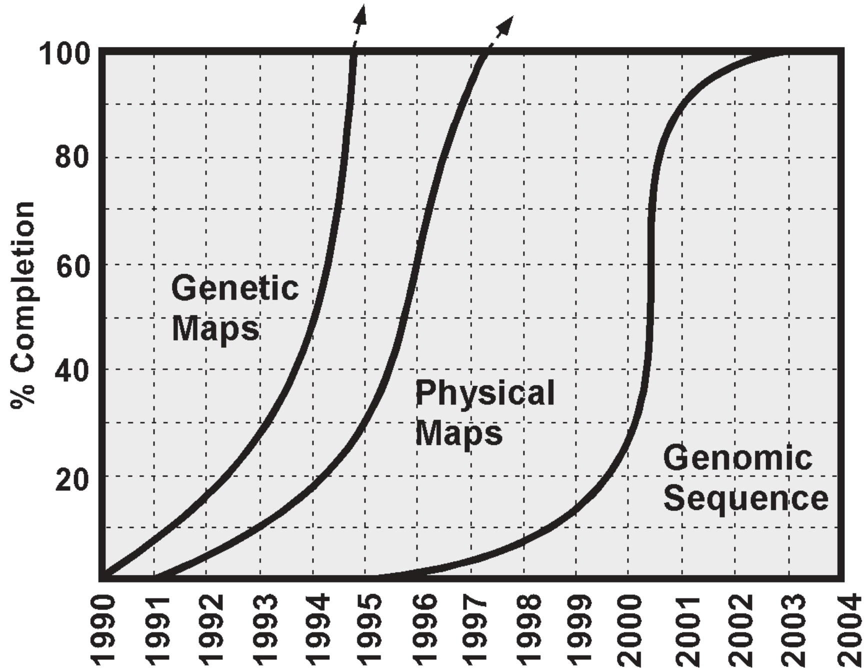


FIGURE 2 Timetable for human genome analysis in the Human Genome Project. The approximate schedule for completing the human genetic map, physical map, and complete genomic sequence is depicted. SOURCE: Reprinted with permission from The McGraw-Hill Companies (Green, 2001).

SEQUENCING THE HUMAN GENOME

The highest-resolution physical map is the actual DNA sequence, which depicts the precise order of bases (adenine [A], guanine [G], cytosine [C], and thymine [T]). A handful of techniques have been developed for sequencing DNA. The approach described by Sanger and coworkers in 1977 (Sanger et al., 1977)—dideoxy chain-termination sequencing—is the most widely used sequencing method. This technique involves the *in vitro* synthesis of DNA molecules in the presence of artificial (dideoxy) bases, which prevent chain extension when incorporated into a growing DNA strand. The resulting population of DNA molecules, which terminate at different base positions, is then analyzed by gel electrophoresis and the relative migration of the various DNA fragments is used to deduce the sequence of the starting DNA. The most contemporary methods for large-scale DNA sequencing involve the incorporation of fluorescent tags into the DNA, followed by their real-time detection using laser-based instrumentation (Hunkapiller et al., 1991; Smith et al., 1986).

In the past decade, there have been numerous advances in fluorescence-based DNA sequencing. In parallel, researchers have gained tremendous experience through the Human Genome Project's efforts to sequence several model organisms, including the yeast *S. cerevisiae* (genome-www.stanford.edu/Saccharomyces), the bacterium *E. coli* (www.tigr.org/tdb/mdb/ mdbcomplete.html), the nematode *C. elegans* (www.sanger.ac.uk and genome.wustl.edu/gsc/gschmpg.html), and the fruit fly *D. melanogaster* (www.fruitfly.org). As a result, the project has focused most recently on sequencing the human genome, with the aim of generating a complete human genome sequence by 2003 (Figure 2). This endeavor has involved two major steps (Waterston and Sulston, 1998). First, suitable overlapping clones are being selected for sequencing—specifically, well-ordered sets of bacterial artificial chromosome (BAC) clones, each containing approximately 100,000 to 200,000 bp of DNA. Second, individual clones are being subjected to a process known as shotgun sequencing (Wilson and Mardis, 1997).

Shotgun sequencing (Figure 3) begins with the generation of subclones that each contain a small (e.g., ~2000 bp), random piece of the starting clone (e.g., BAC). Sequence reads are then obtained from one or both ends of a large number of subclones. Sufficient sequence data are generated such that each base of the starting DNA is read, on average, 6 to 10 times. Computational tools are then used to analyze the resulting sequences so as to identify those that overlap to form sequence contigs (each associated with an assembled consensus sequence). The initial shotgun sequencing data typically result in the assembly of a small number of sequence contigs. The next phase involves generating sequence data in a highly directed fashion so as to fill in the remaining gaps, merge the sequence contigs together, and refine any low-quality regions. This "finishing" process involves a number of specialized computational and experimental

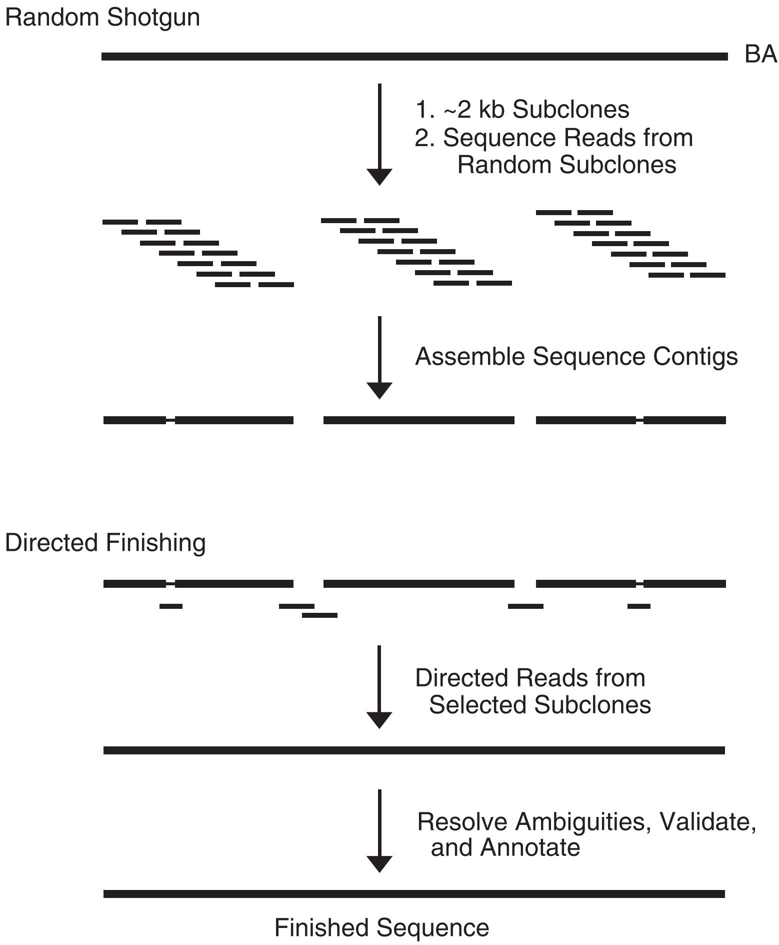

Random Shotgun
BAC
1. ~2 kb Subclones
2. Sequence Reads from Random Subclones
Assemble Sequence Contigs
Directed Finishing
Directed Reads from Selected Subclones
Resolve Ambiguities, Validate, and Annotate
Finished Sequence

FIGURE 3 Shotgun DNA sequencing. The dominant genomic sequencing strategy being used in the Human Genome Project is shotgun sequencing, which consists of two major phases. In the first random shotgun phase a genomic clone (e.g., BAC) is subcloned into approximately 2 kb fragments. Sequence reads are then obtained from one or both ends of a large number of randomly selected subclones. Sufficient sequence data are generated such that each base of the starting clone is read, on average, about 6 to 10 times. These redundant sequence data are then analyzed with various computational tools, allowing the assembly of sequence contigs. Typically, only a handful of gaps between sequence contigs and other problems (e.g., low-quality regions) remain at this stage. In the second directed finishing phase, additional data are generated to complete the sequence, most often by obtaining directed reads from strategically selected subclones. This usually allows the merger of the remaining sequence contigs and the refinement of any low-quality regions to yield a final contiguous, high-accuracy (i.e., finished) sequence. SOURCE: Reprinted with permission from The McGraw-Hill Companies (Green, 2001).

tools, requires highly trained technical personnel, and often involves generating data from particularly difficult sequences (Wilson and Mardis, 1997). The last step of shotgun sequencing involves careful review of the entire assembled sequence, which includes both checking for any ambiguities or problem areas and analyzing the sequence for features of interest.

Several additional points about sequencing the human genome deserve mention. First, this activity is being carefully coordinated among five major and about twelve smaller sequencing centers around the world so that the sequence can be completed as efficiently as possible. Second, all new sequence data being generated by the project are made freely available over the Internet every night (e.g., via the public sequence database GenBank [www.ncbi.nlm.nih.gov]). Note that this includes both the final, finished sequence of individual clones and the evolving, preliminary sequence data of clones whose analyses are still in progress. Finally, the public Human Genome Project's effort to sequence the human genome is occurring in parallel with the much-publicized private effort being performed at the company Celera Genomics (see www.celera.com). The latter group is using a different (albeit complementary) sequencing strategy, whereby the entire human genome is shotgun sequenced en masse (called a "whole-genome shotgun"; Venter et al., 1998).

In the summer of 2000, a major milestone was reached in the Human Genome Project, with completion of a "working draft" sequence for approximately 90 percent of the human genome. In essence, this reflects the acquisition of preliminary sequence data for virtually all of the readily clonable human DNA. This sequence will be refined (i.e., finished to high accuracy) over the next 2 to 3 years. Indeed, the complete, finished sequence for two human chromosomes has already been achieved.

SUMMARY

The Human Genome Project is one of the most important projects—if not *the* most important project—ever undertaken in biomedical research. It is fundamentally an endeavor to develop tools for the study of biology and medicine. These tools reflect both an information resource in the form of genetic maps, physical maps, and the underlying sequence of the human genome and that of several model organisms, as well as an ever-increasing number of experimental technologies that are becoming standard techniques in the armamentarium of biomedical researchers. In this regard, an exciting new "genomic revolution" has started and will forever change the way research is performed. The new and powerful foundation of genetic information is empowering investigators to tackle complex problems relating to disease, development, and evolution that were previously unapproachable. Ultimately, the central legacy of the Human Genome Project will be the provision to future generations of scientists and clinicians of an unprecedented resource—the human "genetic blueprint"—that will allow them to understand fundamental features of human biology (Lander, 1996) and to elucidate the genetic bases of disease (Green, 2001).

REFERENCES

Collins, F. S., A. Patrinos, E. Jordan, A. Chakravarti, R. Gesteland, L. Walters, and Members of the DOE and NIH Planning Groups. 1998. New goals for the U.S. Human Genome Project: 1998–2003. Science 282:682–689.

Green, E. 2001. The Human Genome Project and its impact on the study of human disease. Pp. 259–298 in The Metabolic and Molecular Bases of Inherited Disease, C. R. Scriver, A. L. Beaudet, W. S. Sly, and D. Valle, eds. New York: McGraw-Hill.

Hunkapiller, T., R. J. Kaiser, B. F. Koop, and L. Hood. 1991. Large-scale and automated DNA sequence determination. Science 254:59–67.

Lander, E. S. 1996. The new genomics: Global views of biology. Science 274:536–539.

Sanger, F., S. Nicklen, and A. R. Coulson. 1977. DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences of the USA 74:5463–5467.

Smith, L. M., J. Z. Sanders, R. J. Kaiser, P. Hughes, C. Dodd, C. R. Connell, C. Heiner, S. B. H. Kent, and L. E. Hood. 1986. Fluorescence detection in automated DNA sequence analysis. Nature 321:674–679.

Venter, J. C., M. D. Adams, G. G. Sutton, A. R. Kerlavage, H. O. Smith, and M. Hunkapiller. 1998. Shotgun sequencing of the human genome. Science 280:1540–1542.

Waterston, R., and J. E. Sulston. 1998. The Human Genome Project: Reaching the finish line. Science 282:53–54.

Wilson, R. K., and E. R. Mardis. 1997. Shotgun sequencing. Pp. 397–454 in Genome Analysis: A Laboratory Manual. Vol. 1: Analyzing DNA, B. Birren, E. D. Green, S. Klapholz, R. M. Myers, and J. Roskams, eds. Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press.

Current Genomic Research:
The Proteins Have It

LYNNE J. REGAN
Department of Molecular Biophysics and
Biochemistry and Department of Chemistry
Yale University
New Haven, Connecticut

Proteins play crucial and essential roles in all aspects of cellular function in all organisms. They are composed of linear strings of discrete subunits called amino acids. There are 20 naturally occurring amino acids, which can be grouped into different classes based on the chemical nature of their side chains: acidic, basic, polar, aliphatic, aromatic. All the information necessary and sufficient to specify the final three-dimensional structure that is adopted by a protein is encoded by its amino acid sequence. The amino acid sequence of each protein is itself directly encoded in the chromosomal DNA, which is passed from generation to generation.

A single nucleotide base change in the chromosomal DNA that results in a single amino acid change in a single protein can have devastating, if not fatal, consequences for an individual organism and his or her offspring. Sickle cell anemia, in which the protein hemoglobin forms long aggregates that dramatically reduce its oxygen-carrying capacity and give rise to the characteristic sickling distortion of red blood cells, is a consequence of a mutation of a single amino acid from the negatively charged aspartic acid to the aliphatic valine. There are numerous other examples of diseases that can be linked directly to protein mutation.

Cloned human proteins produced in bacteria are of considerable therapeutic importance. Human growth hormone is administered to children for the treatment of dwarfism (and has been abused by athletes to promote muscle formation!); granulocyte colony-stimulating factor is administered to patients undergoing chemotherapy to promote white blood cell formation; and, of course, thousands of diabetics self-administer insulin daily. These proteins are targets of redesign to optimize their biological, physical, and pharmacokinetic properties for therapeutic purposes. Many human proteins are also potential drug targets.

To cite just a few examples, inhibition of liver phosphorylase would provide a route to the maintenance of adult onset diabetes, inhibition of tumor-promoting proteins provides a mechanism for novel cancer therapy, and inhibition of protein aggregation is a potential strategy by which to combat Alzheimer's disease.

Proteins that are specific to an invading organism are targets for specific inhibition to counteract infection; examples range from the fairly recently developed anti-AIDS drugs, which target HIV's unique protease, to penicillin, which targets the enzymes involved in the bacteria-specific process of cell-wall synthesis.

Protein structure can be classified in a hierarchical fashion (Figure 1). The first level is the amino acid sequence or "primary structure," and the next is the "secondary structure," which is the regular repetition of backbone dihedral angles in a linear stretch of amino acids that gives rise to a common structural unit. The two dominant types of secondary structure in proteins are -helices and -sheets. Elements of secondary structure are connected by loops or turns, which allow them to fold back on themselves and to associate to form a globular "tertiary structure." In some proteins the folded tertiary structure of a monomer is the active form, whereas in others monomers associate with other similar or dissimilar subunits to give specific higher-order complexes or "quaternary structure."

The different tertiary structures or "folds" that combinations of a-helices and b-sheets give rise to are not a continuum but, rather, can be grouped into distinct structural classes. The tertiary folds are specified by the geometrical constraints that are imposed by optimally packing together the elements of secondary structure. A convenient classification scheme is the CATH scheme of Orengo and Thornton (Orengo et al., 1997), in which a protein structure is classified according to its class, architecture, topology, and homology. A sample CATH classification of a protein is shown in Figure 2. If we perform such a classification on all proteins for which structures are known at high resolution by X-ray crystallography or nuclear magnetic resonance, the distribution of structural classes can be illustrated on a CATHerine wheel (Figure 3). We see that approximately one-quarter of all folds are α-helical, one-quarter are all β-sheet, and one-half are a mixture of α-helix and β-sheet.

But how can this information be combined with the results of the genome scale sequencing? The first thing to do when a novel gene is sequenced is to perform a sequence-alignment search against all known sequences from all organisms. If a match is found with a protein of known structure—one of the fold classes discussed above—a model for the structure of the novel protein can be built on the basis of its sequence homology with proteins of known fold. The greater the number of homologous sequences available, the better the alignment and the better the homology model that can be made.

How far can we get with this approach? Is a homology model, rather than an actually determined three-dimensional structure, "good enough"? The answer to this question largely depends on the specific goal of a project. For some purposes, knowing the fold and having a homology model of the protein of

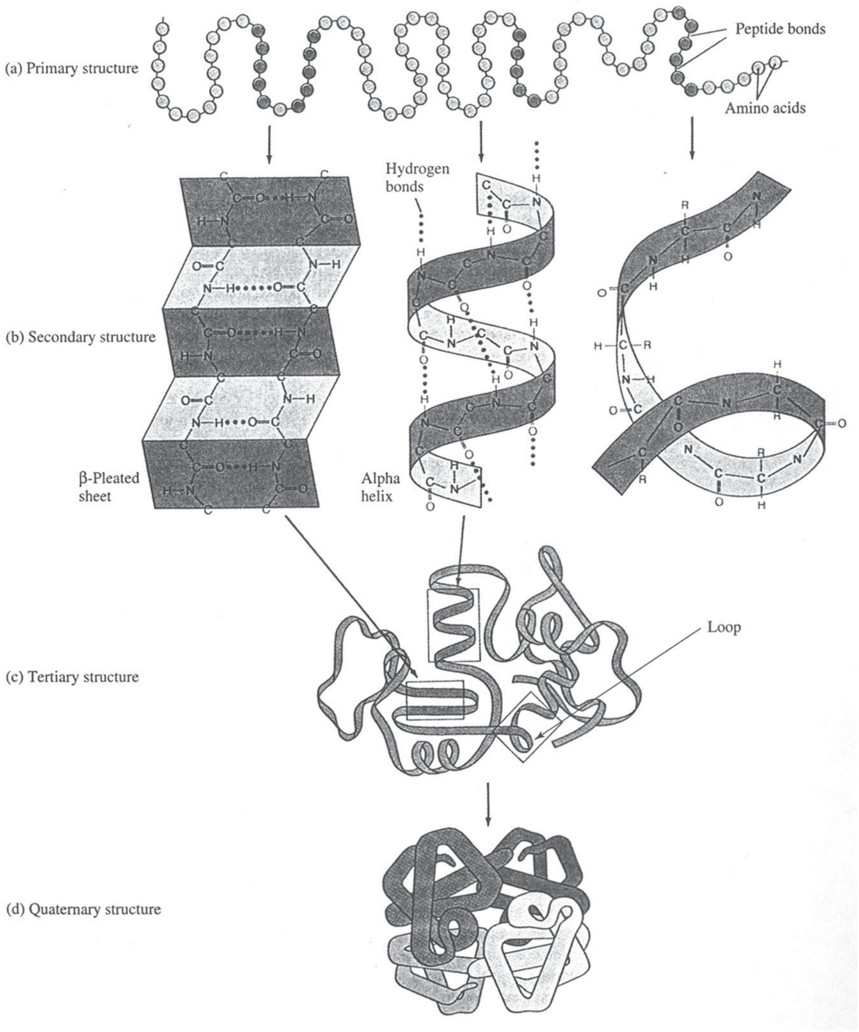


FIGURE 1 The hierarchical nature of protein structure. SOURCE: Reprinted with permission of Wadsworth, an imprint of the Wadsworth Group, a division of Thomson Learning (Boyer, 1999).

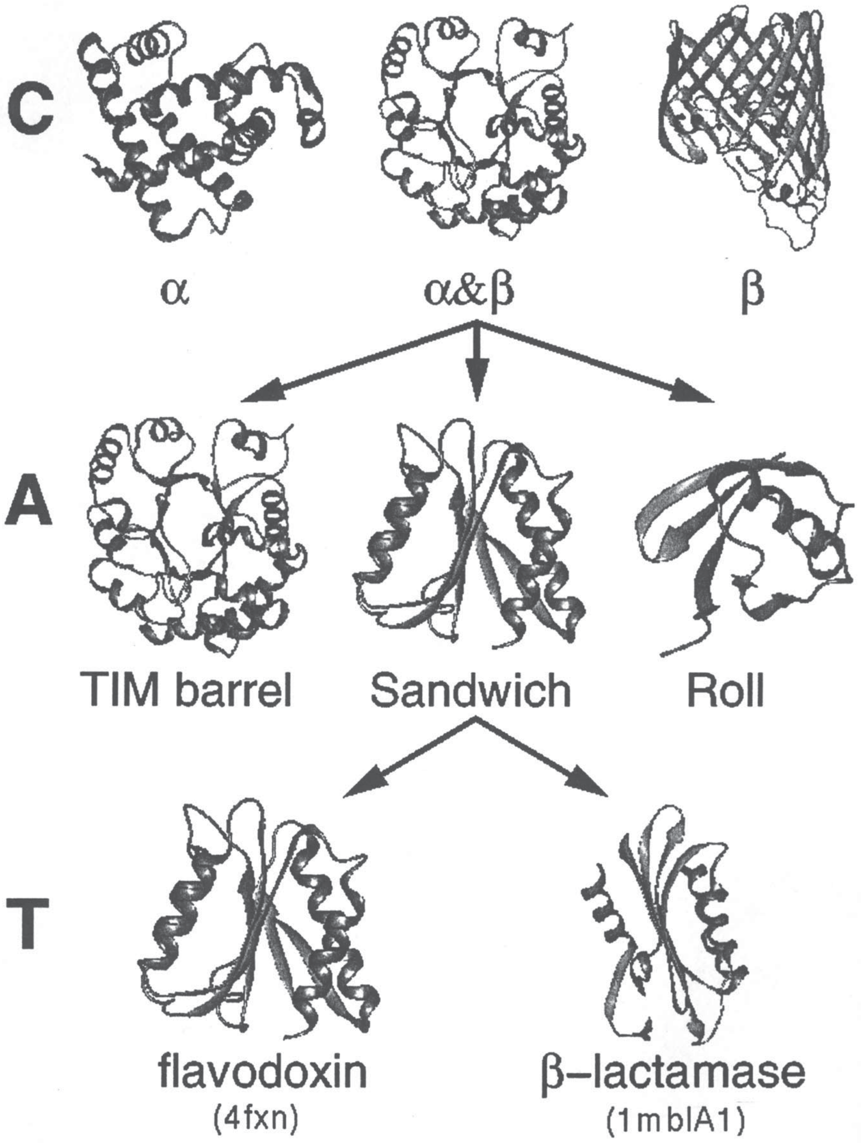


FIGURE 2 The class, architecture, topology, homology (CATH) classification of a protein. SOURCE: Department of Biochemistry and Molecular Biology, University College, London. This figure can be viewed in color at <http://www.biochem.ucl.ac.uk/bsm/cath_new/cath_info.html>.

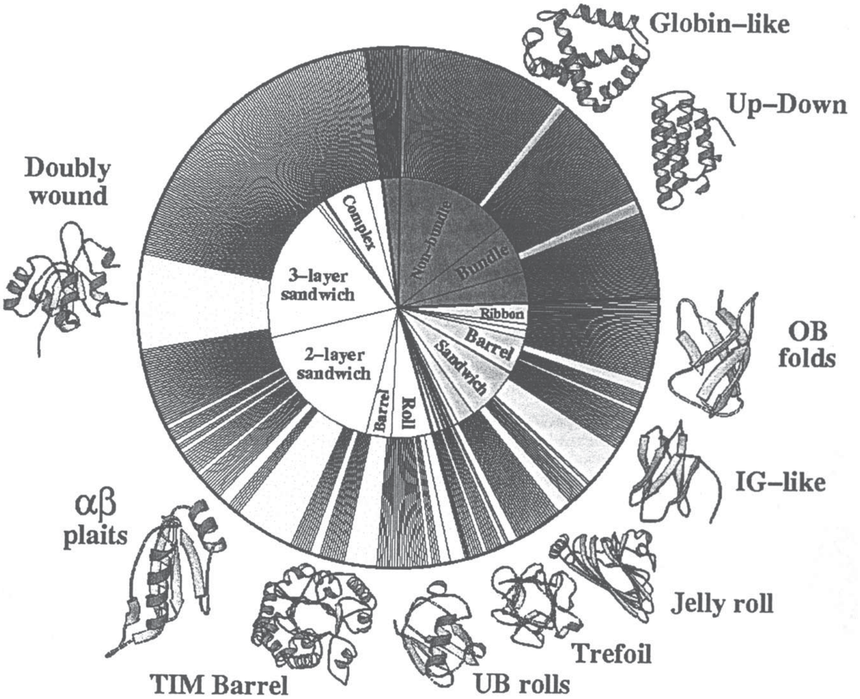


FIGURE 3 CATHerine wheel illustration of the distribution of fold classes in the database of known protein structures. SOURCE: Department of Biochemistry and Molecular Biology, University College, London. This figure can be viewed in color at <http://www.biochem.ucl.ac.uk/bsm/cath/lex/CATHerine_orig.html>.

interest will be sufficient. Will a homology model provide a good enough structural model for rational drug design? The answer to this question is subject to debate. One relevant piece of information is that the goodness of the model depends upon the percent sequence identity of the unknown protein and the target fold. It was first pointed out by Chothia and Lesk (1986) that there is a direct and predictable relationship between the percent sequence identity of two proteins and the backbone root mean square deviation of the two structures.

Having discussed how we can make use of sequence information to predict structure, based on our understanding of sequence-structure relationships and homology modeling to proteins of known fold, we must now consider what fraction of all proteins in an organism either have known structure or can be modeled reasonably satisfactorily by homology. Figure 4 shows the complete distribution of folds and sequences in the genome of *Mycoplasma genitalium* (MG), which is, to date, the smallest known self-replicating organism. Less than

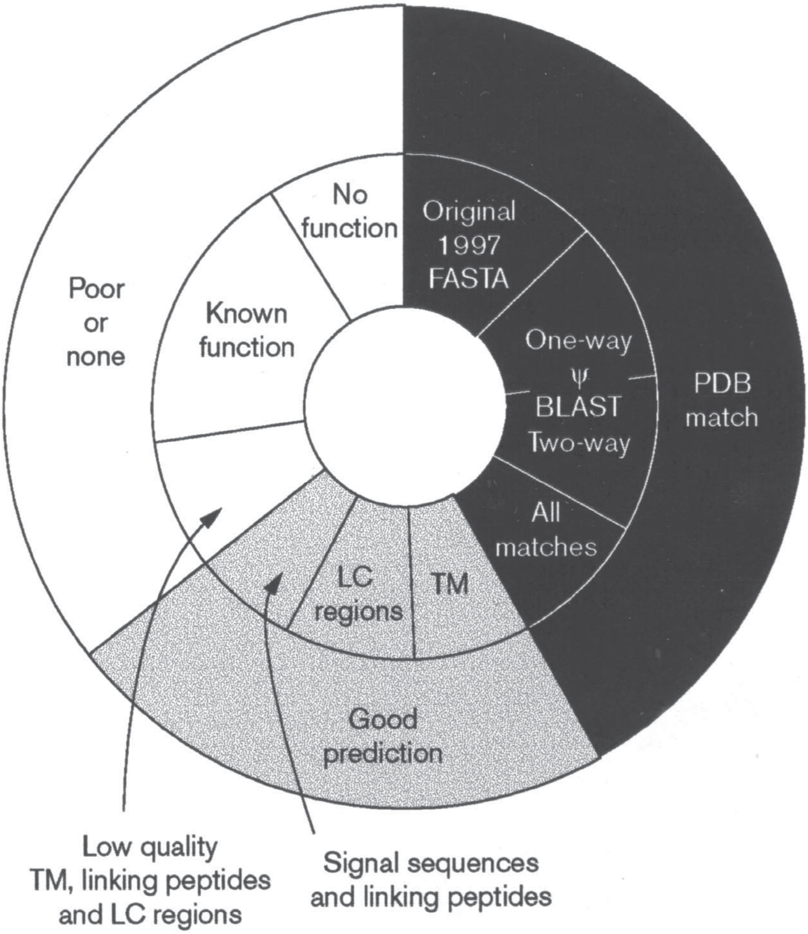


FIGURE 4 Illustration of the distribution of protein sequences in the genome of *Mycoplasma genitalium*. SOURCE: Reprinted with permission from Elsevier Science (Teichmann et al., 1999).

half of the protein structures have been determined experimentally, about a third have structures that can be predicted "well" by homology modeling, and about half are completely unknown. This distribution, which is similar in other organisms, provides an illustration of the scale of missing "fold space." Filling in these missing structures via a number of different strategies is a goal of current genomics research. This information is important not only because it provides fundamental insights into how different protein folds can be used for the same or different functions but also because it is of practical importance in drug design.

There are two complementary approaches to the characterization of unknown proteins in structural genomics projects. First is the "low hanging fruit" approach, in which proteins are cloned, expressed, purified and structures solved, but, initially, at each step only the proteins that behave well and are easy to work with are carried forward to the next step. Second is the rational preselection of representatives of predicted new fold classes, or proteins that may have unusual physical properties, and a concentrated focus on characterizing these proteins, regardless of how easy or difficult it is to work with each protein. Each approach has a number of advantages and disadvantages, and at this stage in structural genomics research, there is clearly a need for both.

As an example, let's consider our work on MG (Balasubramanian et al., 2000). MG has the smallest genome known for any self-replicating organism, encoding approximately 483 proteins (the exact number of predicted proteins varies slightly, depending on the identification method used). Of these, 202 are structurally uncharacterized, 70 are both functionally and structurally uncharacterized, and 25 are completely structurally and functionally uncharacterized over their entire length. Of these 25, 15 are unique to MG, and 10 have homologues in related organisms. We have expressed, purified, and characterized 12 of these 25 proteins. Seven behave like "normal" proteins, display substantial secondary structure, and likely represent novel folds. These are candidates for high-resolution structure determination. One protein is unstructured and may require a partner molecule (either another protein subunit or a nucleic acid or other cellular component) in order to fold, and two display unusual thermodynamic properties: they are highly helical and extremely resistant to thermal denaturation. These latter two proteins are highly conserved from MG to man.

What stages in this work provide engineering challenges and opportunities? Currently, the most common means of production of the large amounts of protein required for biophysical and structural characterization is expression in live bacteria. A better high-throughput method would speed the process. Purification of large quantities of protein to a reasonable level of purity can be accomplished quite readily by attaching a universal "tag," which allows all proteins to be purified by that same method, regardless of their individual chemical nature. Development of an additional universally applicable purification step, which would easily allow proteins to be purified to the high levels required for crystallization, is important. At the moment, a more significant problem than these is that at least half of all proteins, when expressed in bacteria, do not partition into the soluble phase but, rather, aggregate and partition as unfolded protein into the insoluble fraction. This means that they must be refolded before characterization can be begun, or bacterial growth conditions must be manipulated to "coax" as much protein as possible into the soluble phase. In our work with the uncharacterized proteins of MG, obtaining reasonable quantities of pure folded protein certainly was the rate-limiting step.

In summary, this brief overview is intended to provide a sampling of the ways in which studying protein structure and function in the "genomic era" furnishes new challenges and opportunities and is likely to give rise to a host of unexpected and exciting discoveries.

ACKNOWLEDGMENTS

I thank all members of the North East Structural Genomics Consortium, but especially my collaborator at Yale, Mark Gerstein, for his insights and advice on illustrations for my presentation. I acknowledge Christine Orenga and Janet Thornton for the CATH protein classification system and for the illustrations in this article that are taken from their work.

REFERENCES

Balasubramanian, S., T. Schneider, M. Gerstein, and L. Regan. 2000. Proteomics of *Mycoplasma genitaliumi*: Identification and characterization of unannotated and atypical proteins in a small model genome. Nucleic Acids Research 28:3075–3082 and references therein.

Boyer, R. 1999. Concepts in Biochemistry. Pacific Grove, Calif.: Brooks/Cole Publishing Co.

Chothia C., and A. M. Leske. 1986. The relation between the divergence of sequence and structure in proteins. The EMBO Journal 5(4):823–826.

Orengo, C. A., A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, and J. M. Thornton. 1997. CATH: A hierarchic classification of protein domain structures. Structure 5(8):1093–1108.

Teichmann, S. A., C. Chothia, and M. Gerstein. 1999. Advances in structural genomics. Current Opinion in Structural Biology 9:390–399.

Bioengineering for the Science and Technology of Biological Systems

Douglas A. Lauffenburger
Division of Bioengineering & Environmental Health,
Department of Chemical Engineering,
and Biotechnology Process Engineering Center
Massachusetts Institute of Technology
Cambridge, Massachusetts

THE "-OMICS" ERA: SIMULTANEOUSLY "DATA RICH" AND "DATA POOR"

In the past two decades we have seen two dramatic revolutions in biology: first that of molecular biology followed by that of genomics. Certainly, these revolutions are closely related, for genomics would not even be possible without the tools created by molecular biology. Each, however, generates its own type of data and understanding. Molecular biology, on the one hand, ultimately offers reductionist knowledge concerning molecular mechanisms governing functions at all higher hierarchical levels: cell, tissue, organ, and organism. Genomics, on the other hand, at least promises global knowledge concerning relationships of genetically encoded information and operation of the physiological system for which it serves as the core program. These types of knowledge must converge, of course, because genetic information must be transcribed and translated into physicochemical molecular mechanisms in order for us to be able to actually carry out programmed operations.

Combining expectations for continually accelerating generation of both types of knowledge, scientists and engineers anticipate working in a "data-rich" era in the coming decades. Relatively speaking, there is no question that this should be the case, compared to the scattered mist of hard data previously available for biomolecular interactions. At the same time, the extent to which even an imminent avalanche of data of this sort can be presumed to cover the enormously complex and intricate network of interactions at every level of physiological hierarchy must be considered to be tiny for the foreseeable future. Moreover, even as the extent of information grows, it will remain merely information until it can be organized into meaningful conceptual understanding. These quite serious limitations are relevant no matter how the term *genomics* is broadened to comprise not solely the field aimed at discovering gene sequences but also those aimed at elucidating gene expression (the "transcriptome") and protein structure and function (the "proteome").

It is widely appreciated that gene sequence and expression data will be inadequate for understanding the operation of biological systems, for multiple reasons. First, even complete DNA sequence information obviously cannot indicate what genes are actually expressed under any given conditions. Messenger RNA levels for even the full transcriptome, likewise, do not adequately represent the levels of the proteins they spawn. Researchers have demonstrated mathematically, for instance, that the dynamic behavior of a gene regulation network may not be properly predicted solely from gene expression data (Hatzimanikatis et al., 1999a). Moreover, protein levels in themselves cannot satisfactorily characterize molecular mechanisms, since locations (i.e., in various intracellular compartments or multimolecular assemblies) and states (such as phosphorylation on various amino acid residues or cleavage from proforms) substantially alter what the proteins are doing. Thus, knowledge of the reagents (i.e., expression profiles) alone does not offer substantial insight into modes of interactions and mechanisms of operation, which determine ultimate functional behavior. Second, this litany of molecular players—DNA, RNA, proteins—neglects the capacity of other classes of biological molecules (e.g., oligosaccharides, lipids) to regulate biological functions in an epigenetic manner. Finally, living organisms are not autonomous, closed systems. Rather, they operate within the context of environment, which strongly influences every step along the hierarchy outlined above—as well as the genomic content itself.

These arguments should caution us all regarding both the pace and scope of what can be understood about biological systems by whatever combination of "-omics" might be proferred. Basically, my conclusion is that we will remain "data limited" for quite some time yet, despite at the same time surely becoming data rich.

Finally, it should be stressed that the issue raised here is relevant both to the future of biological science, in terms of understanding how biological systems work, and to the future of biotechnology, in terms of enabling creation of products and processes that benefit societal needs.

A BIOENGINEERING APPROACH TO PROGRESS IN THIS ERA

This situation is ideal for bringing to bear a classical engineering mindset on this new biological science and biotechnology—that of an "integrated systems," "design principles and parameters" orientation to the analysis and synthesis of biological processes based on a fundamentally molecular foundation. Both basic scientific understanding and innovative technological development should be advanced by this kind of approach.

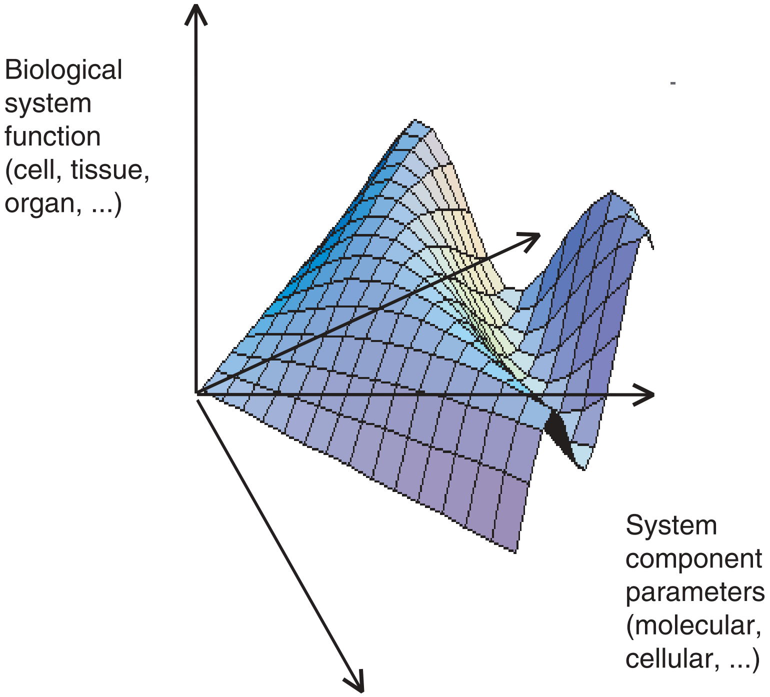

Biological system function (cell, tissue, organ, ...)
System component parameters (molecular, cellular, ...)

FIGURE 1 Illustration of an engineering perspective on biological systems as a dynamic function dependent on component properties.

What engineers generally bring to the research table is a predilection to analyze a complicated system in terms of principles useful for manipulating that system, with the goal of making it operate in some intended fashion (see Figure 1). This perspective emphasizes elucidating and working with an understanding about the system that is admittedly incomplete, but sufficient for “input/output” dependencies to be constructed with enough reliability at the desired level of description that it will do what is needed under most circumstances. Moreover, and just as important, engineers are taught to view a complicated system in terms of its essentially hierarchical nature: a relationship that is mechanistic at one level is phenomenological at another, with parameters empirically measured at one level being useful for prediction of system behavior at the next higher level and being ultimately derivable and predictable from properties of the constituents at the next lower level.

In essence, in this approach the data-limited nature of whatever system is being tackled is accepted, and an adequate balance is found between what is fundamentally known with certainty and what must be assumed or empirically related. The proof of adequacy, of course, lies within the objective of the exercise. For any objective short of universal predictive capability from first principles, this ought to suffice.

A number of examples of this approach can be offered. Some that provide insightful illustration are at perhaps the most ambitiously comprehensive end of the spectrum: the PhysioLab models being developed by Entelos, Inc. (www.entelos.com) for a variety of pathophysiological conditions, with the objective of determining useful drug targets and interpreting clinical trial data. For instance, the Asthma PhysioLab model attempts to account for how the resistance to air flow in the lungs is governed by cell-level functions (e.g., airway smooth muscle cell contractility, airway epithelial cell permeability, white blood cell accumulation in tissues, and secretion of inflammatory mediators), which are, in turn, regulated by cytokine/receptor interactions and consequent intracellular signaling pathways (see Figure 2a).

Computational simulations of this model yield particular realizations for how any variable in the system (levels of molecular species, degrees of cell behavioral activities, extents of tissue functions) depends on system component properties, measurable or assignable. In principle, the computational simulations could more generally produce a “state space” diagram, in which key system parameters (potentially characterizing asthmatic patients in comparison to healthy patients) govern whether airway resistance in response to an environmental challenge stays at a normal physiological state or jumps to a dangerous pathological state (see Figure 2b). In turn, effects of a putative drug regimen can be examined to determine whether it could bring the system back into the safe state. This approach is what will eventually bring the promise of pharmacogenomics to fruition. For any individual patient, information concerning foundational genetic variation (e.g., single-nucleotide polymorphisms) could be related to gene expression and protein property issues that are involved in this overall dynamic system. Now, it is presumptuous to expect that a fully complete model for all relevant physicochemical mechanisms at each level in the space-time hierarchy (e.g., gene expression regulation, signal transduction, cell functions [proliferation, death, differentiation, migration, secretion, contractility], tissue mechanics and transport) might even at the end of the 21st century be available with quantitatively determined parameters. However, it is cowardly to wait to pursue this kind of approach until complete information is on hand.

This tension between being data rich and data limited is, in fact, the history of engineering science and technology. Motor vehicles were manufactured and driven for decades without complete data on the combustion reactions taking place within the engine, yet they moved people and items very effectively, to the overall great benefit of society. Of course, realizing technologies derived from restricted basic scientific understanding has drawbacks: witness environmental issues surrounding the products of those engines. Advances in combustion chemistry and catalytic reactors have helped to reduce the drawbacks in that particular system, certainly, although in medical technologies cost/benefit analyses are sure-

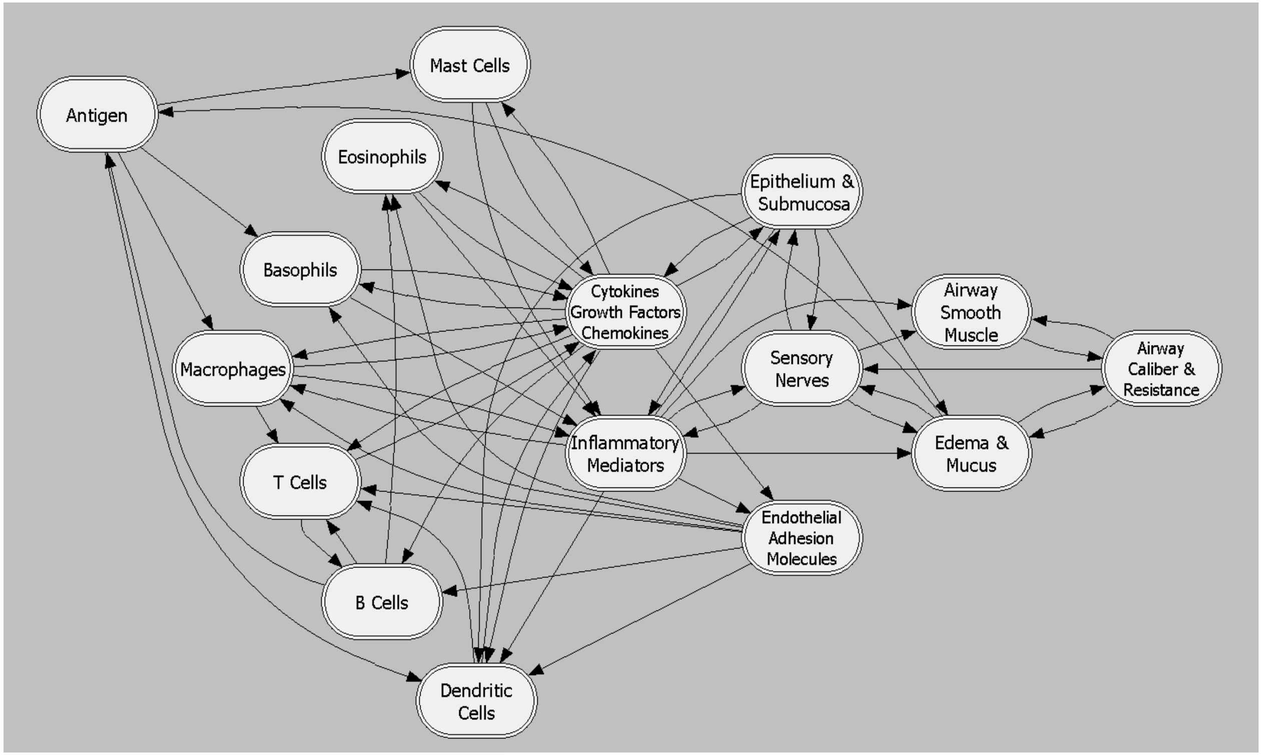


FIGURE 2a Schematic framework for Entelos Asthma PhysioLab, showing cell, tissue, and organ components involved in the computation model. Underlying each component is a set of hierarchical models with increasing detail. SOURCE: Reprinted with permission from Entelos, Inc.

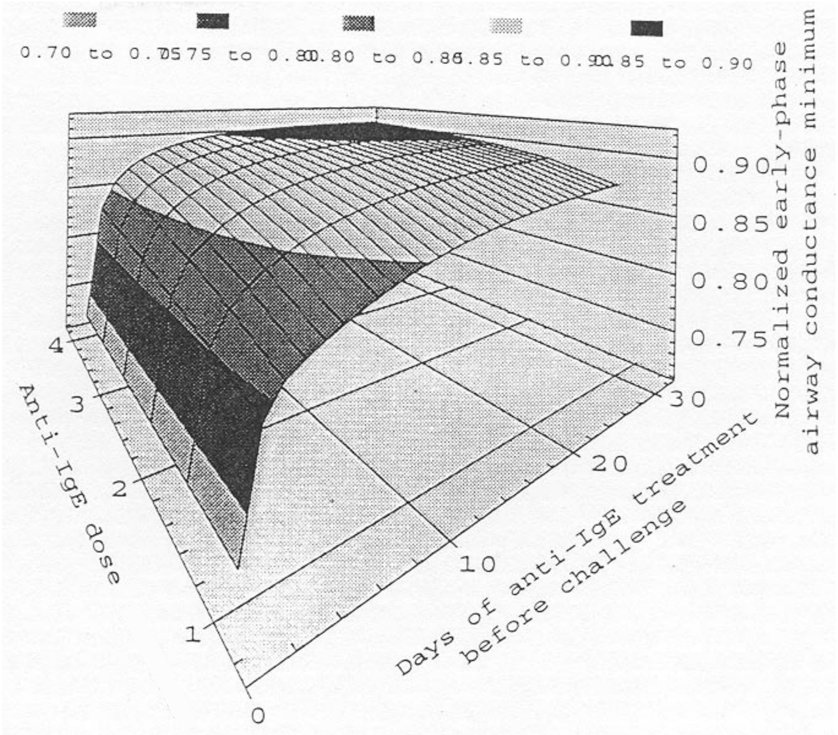


FIGURE 2b Example dynamic systems relationship from Entelos Asthma PhysioLab computer simulations, showing airway conductance minimum as a function of anti-IgE dose level and period. SOURCE: Reprinted with permission from Entelos, Inc.

ly even more difficult to resolve. The point is that the data-limited nature of any system must be considered cautionary, but not necessarily paralyzing.

To be useful, these kinds of systems modeling and analysis approaches need not be aimed at a high-level physiological system. Complexities are just as daunting, yet the benefit of this approach for understanding and manipulation is as exciting, even at the level of an individual cell. An excellent example of this is the systems models for cell cycle control that have been developed over the past decade (e.g., Chen et al., 2000; Hatzimanikatis et al., 1999b). A “wiring diagram” that contextualizes the cell cycle molecular regulatory network has been provided (Kohn, 2000), and to look at it initially is to regret contemplating a systems modeling description (see Figure 3a). However, rather than rushing headlong into writing differential equations for all components and interactions, Tyson’s and Bailey’s groups at Virginia Polytechnic Institute and State Univer-

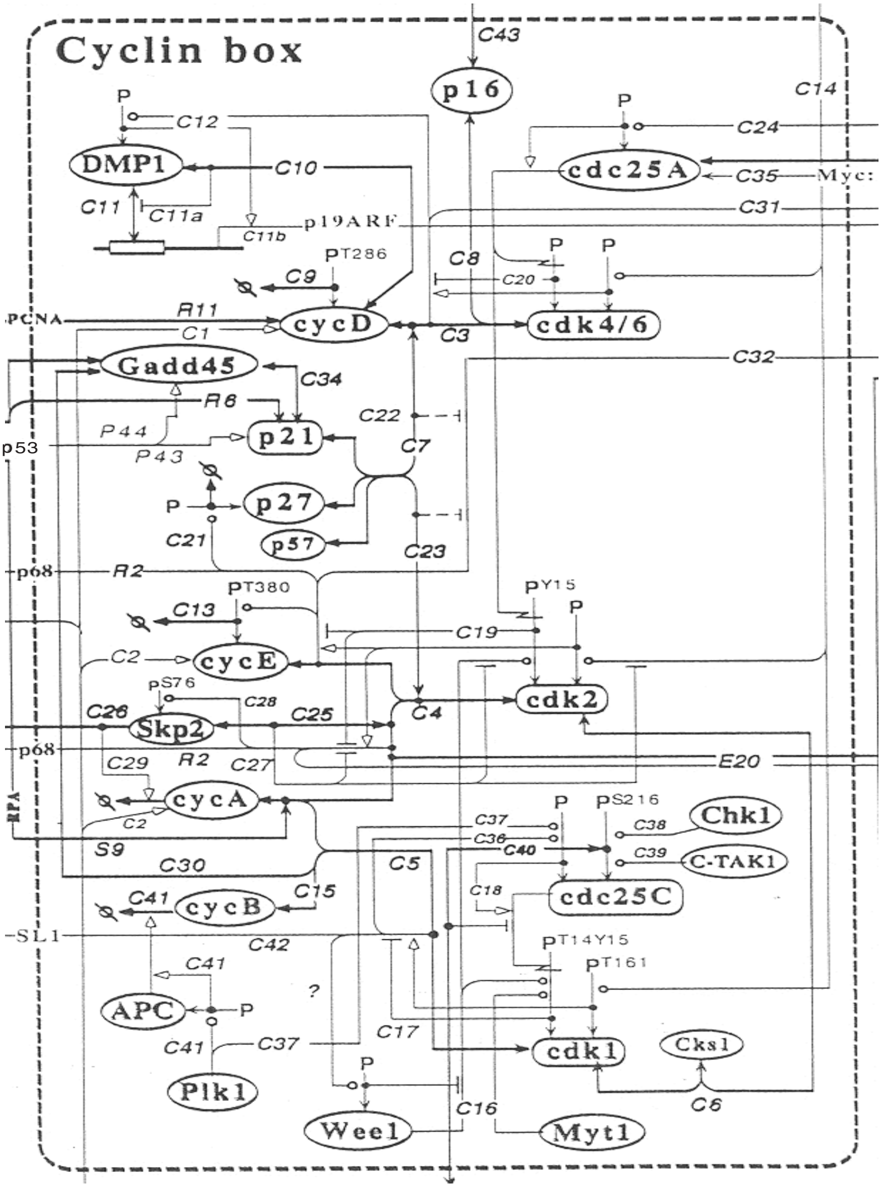


FIGURE 3a Biomolecular "circuit diagram" for the cell cycle. SOURCE: Reprinted with permission from the American Society for Cell Biology (Kohn, 2000).

sity and ETH Zürich, respectively, have identified key underlying "modules" that are at the core of the dynamic behavior of this system and then incorporated additional network elements around this core to explain and predict more and more experimental data concerning wider aspects of its regulation (see Figure 3b). Again, one can aim for construction of dynamic systems behavior versus parameter relationships, such as a bifurcation diagram for understanding and predicting conditions under which cells will progress through a DNA synthesis checkpoint or not (see Figure 3c). It should be emphasized that an important element of biological systems analysis along these lines is likely to be a consideration of stochastic issues, since most biological data represent either individuals or individuals distributed across a population—whether at the level of cell, organ, or organism (e.g., Arkin et al., 1998).

This success provides a concrete example of the concept that a modular approach to modeling and understanding highly hierarchical biological systems should be productive (Hartwell et al., 2000; Lauffenburger, 2000). That is, it may be anticipated that crucial core functions of biological systems (e.g., metabolism, force generation) are performed by an identifiably restricted mechanism, but the vast proportion of the overall system's components are found to serve as control and safety networks in order to ensure robustness and reliability of core function

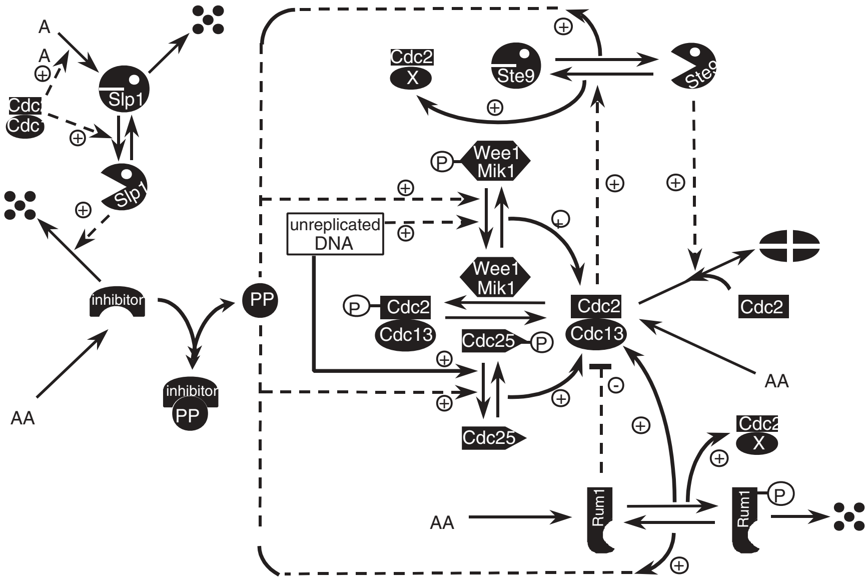


FIGURE 3b Schematic illustration of mathematical model for the cell cycle. SOURCE: Reprinted with permission from the American Society for Cell Biology (Chen et al., 2000).

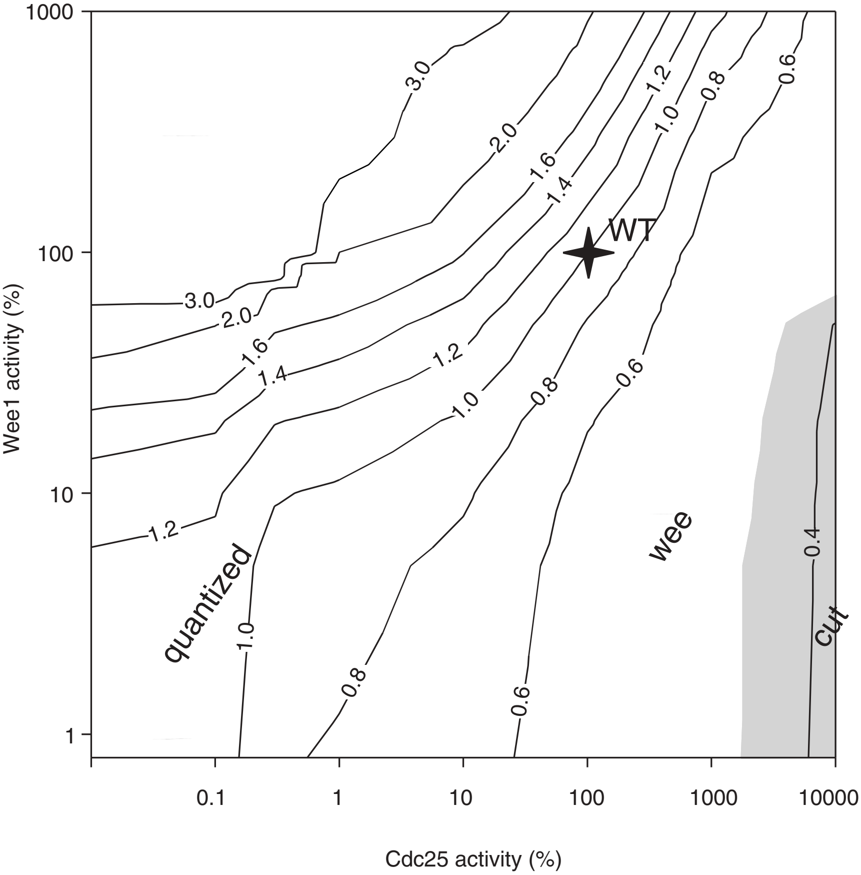


FIGURE 3c Example dynamic systems relationship. Model computations for yeast cell proliferation behavior as a function of activities of two key gene products. SOURCE: Reprinted with permission from the American Society for Cell Biology (Chen et al., 2000).

under highly variable environmental conditions (Carlson and Doyle, 1999). It is conceivable that much of the uncertainty regarding quantitative parameter values might, in fact, be rendered less problematic by this concept of functional and control modules, because their dynamic operation may turn out to be surprisingly insensitive to specific parameter values (von Dassow et al., 2000). This idea will be very interesting to examine with emerging powerful genetic (Rao and Verkman, 2000) and chemical (Schreiber, 2000) molecular intervention methodologies.

Neither of these impressive examples, however, embodies the full scope of the potential engineering systems approach that will need to be pursued, since both focus on directly observable (at least in principle) physicochemical interactions. Since a large proportion of the physicochemical processes involved in molecular and cellular networks—and their corresponding parameters—will remain undetermined for years to come, systems modeling approaches less dependent on a detailed physicochemical framework should be valuable to pursue. The dynamic module approach noted above is one possibility. An early instance of this can be found in the signaling network regulating bacterial chemotaxis. For this phenomenon of biased cellular motor rotation, leading to directional cell movement, the dynamic network behavior of 14 protein components (see Figure 4a), which could require on the order of a few dozen parameters for quantitative physicochemical description, responds to a step change in input to yield a perfectly adapting output result (see Figure 4b). This input/output relationship then permits the full network to be replaceable by an integral feedback control loop that can be characterized by two phenomenological parameters (Yi et al., 2000; see Figure 4c). While at first glance this might seem to be a step backward, the physicochemical detail has not disappeared; rather, it is ultimately responsible for the dynamic behavior, including the quantitative parameter values, so the feedback control loop could be replaced by the detailed subsystem whenever the complete set of information becomes available.

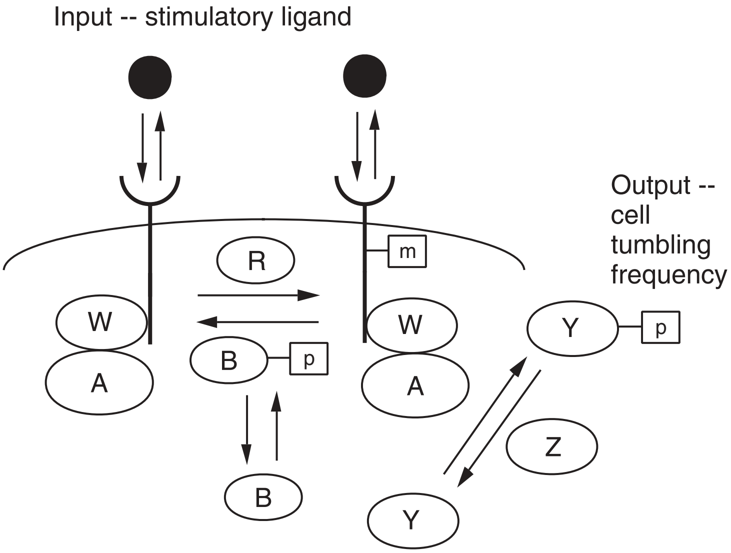

Input -- stimulatory ligand
Output -- cell tumbling frequency
R
m
W
A
B
p
W
A
B
Y
p
Z
Y

FIGURE 4a Signal transduction network for bacterial chemotaxis, showing how cell tumbling frequency (which governs chemotactic locomotion) is regulated by a stimulatory ligand. SOURCE: Reprinted with permission from Lauffenburger. Copyright (2000) National Academy of Sciences, U.S.A.

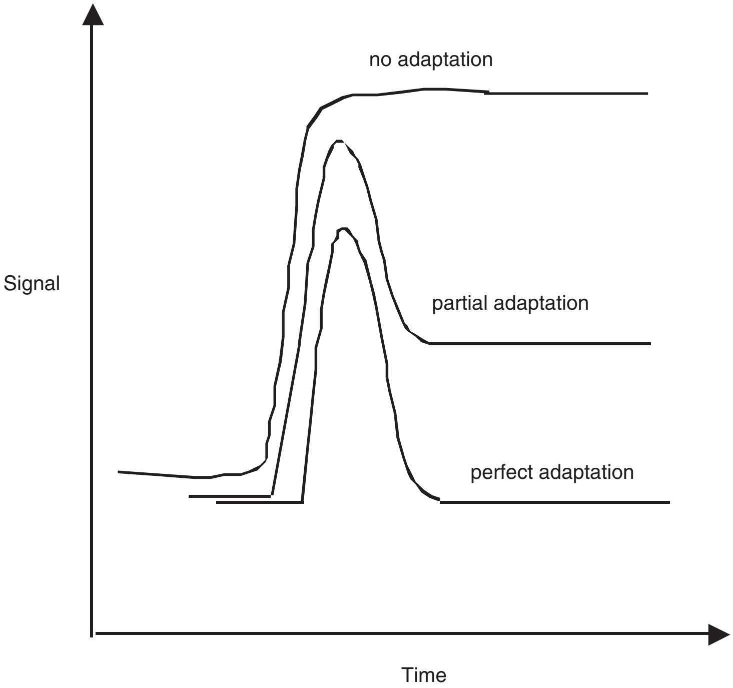


FIGURE 4b Dynamic signal behavior possibilities in response to step input; the network shown in Figure 4a yields "perfect adaptation" behavior. SOURCE: Reprinted with permission from Lauffenburger. Copyright (2000) National Academy of Sciences, U.S.A.

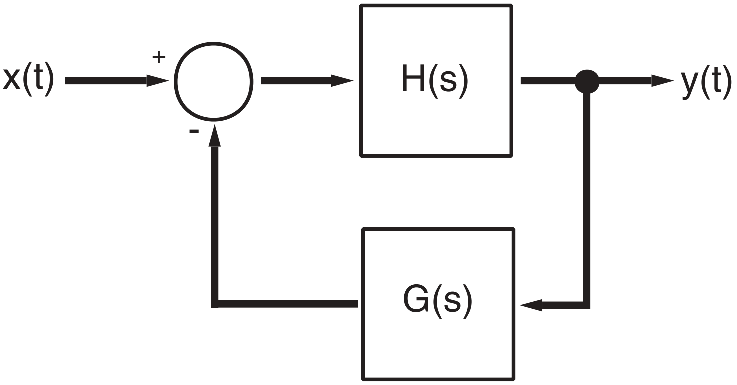


FIGURE 4c Schematic of feedback control module, which can generate the dynamic signal behavior shown in Figure 4b.

Another possibility is employment of more phenomenological relation-focused models, developed from analysis of large but not necessarily mechanism-oriented data sets. A classical reference point for this concept may be found in biomedical engineering treatments of organ-level physiological dynamics, such as an electrocardiogram, which use signal processing techniques to develop relationships useful for analysis and technology creation, despite possessing little direct connection to underlying fundamental mechanisms. An analogous methodology may be valuable for modeling at least some aspects of protein-protein and protein-gene regulatory networks as signal processing elements (Asthagiri and Lauffenburger, 2000; McAdams and Arkin, 1998). Finally, yet a different approach is that of a cybernetic perspective, in which physicochemical mechanistic details are largely replaced by "objective-based" algorithms that characterize programs by which cells might be presumed to manage molecular resources (Varner, 2000; Varner and Ramkrishna, 1999).

We have ourselves begun to embark on an effort to combine some of these concepts, attempting to analyze the cell behavioral response of apoptosis (programmed cell death) versus survival in response to death-promoting and survival-promoting factors by following a combination of gene expression and protein level/state/location dynamics. Out of hundreds of putative death-activating and survival-protective genes—and dozens of protein-based physicochemical kinetic and transport processes—we have selected a subset for microarray expression and proteomic experimental measurement following challenge by a matrix of input stimuli. The question posed is whether we can determine a signal processing algorithm utilized by the cells to make their decision based on information flows through key regulatory networks (e.g., Arkin and Ross, 1994).

Finally, it is important to emphasize that major advances are concomitantly needed in experimental methodologies for quantification of molecular and cellular processes. This need ranges from improvements in surface chemistry and fluorescence imaging for nucleic acid and peptide microarrays to isolation and characterization of large sets of proteins from small cellular samples, to creation of tissue-engineered in vitro organ surrogates for generating more nearly physiological cellular contexts (Griffith et al., 1997), to instrumentation enabling measurement of subtle but important functional properties of and within living organisms such as transgenic mice. In short, the same type of high-throughput acceleration of data gathering that has arisen at the gene expression level must be propagated to higher levels of the biological systems hierarchy.

ACKNOWLEDGMENTS

I would like to express my appreciation to a number of colleagues whose comments and perspectives in recent months have helped shape the thoughts outlined here, including Anand Asthagiri, Jay Bailey, Rick Horwitz, Kelvin Lee, and Steve Wiley.

REFERENCES

Arkin, A., and J. Ross. 1994. Computational functions in biochemical reaction networks. Biophysical Journal 67(2):560–578.

Arkin, A., J. Ross, and H. H. McAdams. 1998. Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected *Escherichia coli* cells. Genetics 149:1633–1648.

Asthagiri, A. R., and D. A. Lauffenburger. 2000. Bioengineering models of cell signaling. Annual Review of Biomedical Engineering 2:31–53.

Carlson, J. M., and J. Doyle. 1999. Highly optimized tolerance: A mechanism for power laws in designed systems. Physical Review E 60(2A):1412–1427.

Chen, K. C., A. Csikasz-Nagy, B. Gyorffy, J. Val, B. Novak, and J. J. Tyson. 2000. Kinetic analysis of a molecular model of the budding yeast cell cycle. Molecular Biology of the Cell 11:369–391.

Griffith, L. G., B. Wu, M. J. Cima, M. J. Powers, B. Chaignaud, and J. P. Vacanti. 1997. *In vitro* organogenesis of liver tissue. Annals of the New York Academy of Sciences 831:382–397.

Hartwell, L. H., J. J. Hopfield, S. Leibler, and A. W. Murray. 2000. From molecular to modular cell biology. Nature 402(6761 Suppl):C47–C52.

Hatzimanikatis, V., L. H. Choe, and K. H. Lee. 1999a. Proteomics: Theoretical and experimental considerations. Biotechnology Progress 15(3):312–318.

Hatzimanikatis, V., K. H. Lee, and J. E. Bailey. 1999b. A mathematical description of regulation of the G1-S transition of the mammalian cell cycle. Biotechnology and Bioengineering 65(6):631–637.

Kohn, K. W. 2000. Molecular interaction map of the mammalian cell cycle control and DNA repair systems. Molecular Biology of the Cell 10:2703–2734.

Lauffenburger, D. A. 2000. Cell signaling pathways as control modules: Complexity for simplicity? Proceedings of the National Academy of Sciences of the USA 97:5031–5033.

McAdams, H. H., and A. Arkin. 1998. Simulation of prokaryotic genetic circuits. Annual Review of Biophysics and Biomolecular Structure 27:199–224.

Rao, S., and A. S. Verkman. 2000. Analysis of organ physiology in transgenic mice. American Journal of Physiology Cell Physiology 279(1):C1–C18.

Schreiber, S. L. 2000. Target-oriented and diversity-oriented organic synthesis in drug discovery. Science 287:1964–1969.

Varner, J. D. 2000. Large-scale prediction of phenotype: Concept. Biotechnology and Bioengineering 69:664–678.

Varner, J. D., and D. Ramkrishna. 1999. Mathematical models of metabolic pathways. Current Opinion in Biotechnology 10(2):146–150.

von Dassow, G., E. Meir, E. M. Munro, and G. M. Odell. 2000. The segment polarity network is a robust development module. Nature 406:188–192.

Yi, T. M., Y. Huang, M. I. Simon, and J. Doyle. 2000. Robust perfect adaptation in bacterial chemotaxis through integral feedback control. Proceedings of the National Academy of Sciences of the USA 97:4649–4653.

Genomics and Ethics

Pilar N. Ossorio
University of Wisconsin Law School
Madison, Wisconsin

The U.S. Human Genome Project (HGP) is the first federally funded science program in which a portion of the budget is set aside for studying the ethical, legal, and social implications (ELSI) of the science. Ten years of ELSI research have produced numerous empirical and theoretical studies. Here I briefly address only two ELSI topics: 1) medical information privacy as applied to human genetics, and 2) ethical problems arising from the use of genetic interventions to change who will live in the future.

GENETICS RESEARCH AND INFORMATION PRIVACY PROTECTION

Numerous parties—scientists, biotechnology companies, pharmaceutical companies, and patient advocacy groups, for instance—have begun to value DNA not only for its biochemical properties but also for the many layers of information it contains. This information can be used to identify health-related traits and to predict the probabilities that individuals will develop certain diseases. In the future, genetic information may provide one means of predicting behavioral tendencies, personality traits, or abilities that are related to the acquisition of resources and social positions (i.e., wealth, desirable jobs, or political office). Because people's DNA contains much sensitive information about them, and because DNA itself can be used as a means of identifying persons, new research raises many problems concerning who should have access to genetic information, under what circumstances, and for what purposes.

Genetic information privacy is a subset of broader questions about the privacy of medical and other personal information. People value privacy for instrumental reasons and, perhaps, for some noninstrumental ones. Instrumental reasons for valuing genetic information privacy are to avoid unfair denial of goods and services, unfair denial of opportunities, or stigmatization. Genetic information might be used as a basis for limiting an individual's access to insurance, denying or limiting employment opportunities, or limiting one's educational opportunities. By creating privacy protections, we can prevent insurers, employers, educational institutions, and others from using or misusing genetic information to harm us.

We may also value information privacy for noninstrumental reasons. Part of developing a personal identity consists of determining who should know what about us. Scholars have argued that information privacy is essential in allowing people to develop as autonomous individuals who have some control over what "face" they present in particular circumstances. Part of showing respect for a person consists of not exposing their private aspects or information. This argument applies to genetic information, because knowledge of your genetic predispositions may affect how people treat you or conceive of you—it may affect their attribution of an identity to you.

There are numerous ways in which genetic information generated during research can become known to "third parties." For instance, if the research involves genetic testing or sequencing and the reporting back of test results, these results may become part of a subject's medical records and/or school records and, therefore, may become accessible to hundreds of people; if the research involves DNA sequencing and the sequence information is linkable to a subject's identity and not stored securely, it may become accessible; and if the research involves banking of bodily materials and those materials can be both linked to the subject and used for DNA sequencing or testing by other researchers or at a later date, genetic information may become "public."

Sequence information raises particularly difficult privacy concerns, because when we obtain sequence data, we do not necessarily understand or recognize all of the information the data contain. The sequence of a gene and its surrounding regions may contain mutations that predispose to a disease, but we may not yet recognize this. Another possibility is that the sequence may contain elements that regulate how and when a gene is turned on and off, but we again might not yet recognize this. Sequence data can be "mined" repeatedly for new information and knowledge: data that at first seem innocuous and unprejudicial might turn out to be far more problematic as our knowledge expands.

Both ethical and legal theorists acknowledge that individual privacy interests or rights must always be weighed against other social goods. In the case of biomedical research, the primary good at stake is great medical advances. However, scientists also recognize that these advances will not take place, or will be delayed, if people refrain from participating in research because they fear losing their insurance or jeopardizing their jobs. Medical quality control and perhaps some other social goods, such as increasing equality of opportunity, could also be served by some disclosures of genetic information.

Society balances these privacy interests against other interests through several different sets of rules and ethical norms; these include federal and state statutes, common law such as privacy torts, institutional rules, and institutional and professional ethics norms. Federal human subjects research regulations currently are considered inadequate with respect to protecting genetic privacy, but they are under review by the National Bioethics Advisory Commission, and recommendations for their revision are forthcoming. Many states have genetic information privacy laws; however, because these laws differ from each other in fundamental ways (such as in their definitions of “genetic information”), they may undermine researcher’s ability to undertake the large-scale, multistate projects that will be necessary for the next phase of genetics discoveries. The U.S. Congress has been attempting for several years now to pass comprehensive legislation concerning medical information privacy (which would cover genetic information generated during research), but the complexity of the issues, in addition to a lack of political consensus, has prevented the passage of legislation.

Most scholars believe that we need antidiscrimination rules in addition to privacy rules. Antidiscrimination rules limit what people or institutions can do with genetic information if they obtain it. We need antidiscrimination rules because we cannot and should not have absolute information privacy; the absolutist approach would undermine other important social goals. Currently, some state laws on genetics include antidiscrimination provisions. In addition, some existing federal laws, such as the Americans with Disabilities Act of 1990 (ADA), may provide a degree of protection against discrimination based on genetic information.

NEW ETHICAL CHALLENGES: FUTURE GENERATIONS’ PROBLEMS

Thus far, the legal and ethical questions I have discussed are difficult, but they do not substantially challenge our models of ethical analysis; we have the intellectual tools to address problems associated with genetic and other medical information privacy. Advances in genetics, however, have also highlighted some areas in which our current ethical and legal theories are inadequate. One such area is the problem of determining what parents, physicians, and the society as a whole are permitted or obligated to do on behalf of future generations.

The notion that people have obligations to those who will live in the future, particularly their descendants, is deeply ingrained in many societies. The degree to which present choices and actions reflect concern for impacts on future generations is one measure of an individual or society’s moral character. With respect to genetics research and the application of genetic technologies, the issue of obligations to future generations arises because genetic technologies can be used in conjunction with reproductive technologies to influence the traits of people who will be born.

One early use of genetics has been to attempt the birth of children who are free of diseases with a known genetic cause (e.g., Tay Sachs disease, cystic fibrosis, Huntington's disease). In the future, we may employ germ-line genetic interventions (germ-line gene therapy) to change the genomes of our offspring. This could be done for therapeutic reasons—to prevent the person who will be born from having a disease—or to enhance or change some trait, such as height, weight, skin color, or psychological predisposition. By changing the germ-line, one could change the genome of an entire lineage. Note, however, that it is not necessarily the case that an entire lineage would be affected. A person with a germ-line modification would have the same reproductive freedom as anybody else, and could choose whether to become a genetic parent, and whether to use assisted reproduction to prevent his or her genetic modification from being transmitted to future generations.

What ethical principles or approaches should guide us as individuals or as a society in determining which genetic interventions are permissible/impermissible or obligatory? One important part of that equation is the child who will be born, and that child's lineage. We want to ask what is best for the future child: What should we do on behalf of the person who will be born? But this leads us to a particular paradox: many of the genetic interventions we might do on behalf of a future person would actually change *who* will be born. If we employ a technology that uses genetic testing to select which embryos should and should not be implanted or aborted, then we are clearly choosing that some people will be born in place of others.

Even if we alter the genome of an existing embryo, we might change who is born. This might be the case because more than one numerical person can arise from the same fertilized egg (conceptus). Identical twins arise from the same conceptus, and they are not the same person, demonstrating that one genome can lead to more than one person. If this is the case with twins, it should also be the case with a conceptus that does not become twins. That conceptus might become one of several different people, depending on events that occur during its development. One event that could occur is that one or some of its genes might be changed. If these changes substantially affect embryonic and perinatal development, they might result in the birth of a person different from the one who would have been born in the absence of genetic intervention.

Why does it matter that our genetic interventions might change who is born? Because our standard ethical theories make it difficult to argue that you can do something on somebody's behalf if your actions mean that they will not be born and instead a different person will come into existence. You cannot say that the life of the person who would have been born with Tay Sachs disease or cystic fibrosis is made better if what you have done is substituted a different person who will not have the disease. And the new person has not been made better off, because he or she would not have been born with a disease in the absence of intervention; he or she would not have been born at all!

This problem, called the "nonidentity problem," throws into doubt the application of traditional ethical notions of harm and benefit, at least with respect to many decisions that one would like to make on behalf of future generations. This is because our standard notions of harm and benefit are "person affecting"; they rely on the possibility that some particular person's interests will be advanced or thwarted by our actions. For instance, our standard notion of preventing disease means preventing people from developing or acquiring the disease, which is different from preventing the disease by preventing the birth of people with the disease.

To address the nonidentity problem, philosophers have proposed that future generations' choices in which we change who is born should be characterized by an "impersonal" ethic. An impersonal principle of beneficence would look something like the following: if in two possible futures the same number of people would be born, it would be worse if those who lived were worse off than others who could have lived. This is a principle that judges the rightness of a choice by weighing two possible future worlds against each other: If we make the better choice, there will be people who are better off, but the reason that choice was better is not because any particular people have been made better off!

If we are to use impersonal principles to guide and evaluate choices about genetic interventions that affect future generations, then there are still many issues to be addressed. Are impersonal harms just as bad as person-affecting harms? What methodology should we use to evaluate the two hypothetical future worlds? Should impersonal principles only apply when the nonidentity problem arises?

Nanoscale Science and Technology

From Clusters to Automobiles: Processing and Applications of Granular Nanomaterials

Horst W. Hahn
Institute of Materials Science
Darmstadt University of Technology
Darmstadt, Germany

Matter with nanometer-size structures has been around for billions of years in the universe. It is currently speculated that the assembly of nanometer-size particles into larger building blocks that eventually exhibit gravitational forces is the basic mechanism of planet formation in the proximity of newly formed stars. On Earth, scientists have been attracted to clusters consisting of a few to a few hundred atoms because of the interesting physical and chemical properties that are dominated by quantum-size effects. More than 15 years ago, a new field of materials science evolved, initiated by the original work of Gleiter (1989), and is now an integral part of the broader field of nanotechnology. This new field is based on the fact that many properties of materials with technological relevance depend on the size of the microstructural features, such as grain size, or are influenced by the large fraction of interfaces that exhibit a disordered structure compared to the usual crystalline order. The detailed description of the internal structure of a material is commonly termed “microstructure.”

Here I present the potential of manmade materials with microstructural features in the nanometer range, along with a few examples of technological relevance. In general, nanomaterials can exhibit nanometer-size features in zero, one, two, and three dimensions. Consequently, these materials are named “clusters/nanoparticles,” “multilayers,” “thin films,” and “nanocrystalline materials.” Their common feature is that the fraction of atoms residing in the interfacial regions, such as surfaces and grain/phase boundaries, is extremely large and can exceed 50 percent of the total number of atoms for dimensions in the order of 5 nm. In order to tailor the physical and chemical properties of nanomaterials, the microstructure has to be controlled on the level of a few nanometers or even on the atomic level, during the entire synthesis and processing. Presented here are gas phase techniques that allow the synthesis of all nanostructured materials irrespective of their dimensionality. Examples of size-dependent properties and their technological applications are given. The materials are not limited to clusters, nor are the applications restricted to automobiles, but include all classes of nanostructured materials mentioned above and apply as well to such application areas as UV protection, catalysis, hard disks, magnetoelectronic devices, and many more, demonstrating the enormous potential of the design of materials on the nanometer scale.

Gas phase processes provide a convenient and versatile way to prepare materials, with excellent control of the microstructure starting from an atomic level. Techniques such as sputtering, Molecular Beam Epitaxy (MBE), and Chemical Vapor Deposition (CVD) are well-established processes, even in industrial production of thin-film devices. The underlying principle is the transfer of atoms/molecules into the gas phase and their deposition onto an appropriate substrate, where the growth is controlled on the atomic level by adjusting temperature, growth rates, and so forth. By employing sophisticated technological modifications of the basic principles, thin films, multilayered structures, and granular materials—that is, nanoparticles embedded in a matrix—can be readily prepared. Simply increasing the gas pressure in the synthesis chamber from the range of the above processes (10^{-11} to 10^{-3} mbar) to the range from 1 mbar to ambient pressure, and/or increasing the precursor partial pressures, changes the growth mechanisms drastically, and a whole new class of nanomaterials can be obtained. Because of the elevated pressure, clusters/nanoparticles are nucleated in the gas phase by the increase of collisions between the precursor atoms or molecules. Consequently, the processes are termed (Inert) Gas Condensation (IGC) and Chemical Vapor Synthesis (CVS), referring to the difference of the growth mechanisms.

Recently, it was discovered that the electrical resistivity of a multilayered structure consisting of alternating layers of a ferromagnetic metal, such as Co, and a nonferromagnetic metal interlayer, such as Cu, is drastically reduced in a magnetic field. The effect was termed Giant Magnetoresistance (GMR), because of the large change of the resistance of up to 100 percent, compared to the much smaller Anisotropic Magnetoresistance (AMR) effect. Besides the interest in gaining a basic understanding of the physical concepts of GMR, the main driver for the immense efforts in the field of magnetoresistance effects has been the potential for applications in electronic data storage and for a large variety of precision sensors for revolutions and angle detection.

A structural prerequisite for the GMR effect is the exact control of the thickness of the interlayers in the range of 1 to 2 nm, that is, 5 to 10 monolayers. Depending on the thickness of the interlayer, either ferromagnetic or antiferromagnetic coupling between the two neighboring magnetic layers is observed. In the case of antiferromagnetic coupling, a magnetic field can change the direction of the magnetic moments of some layers, resulting in the observed decrease of

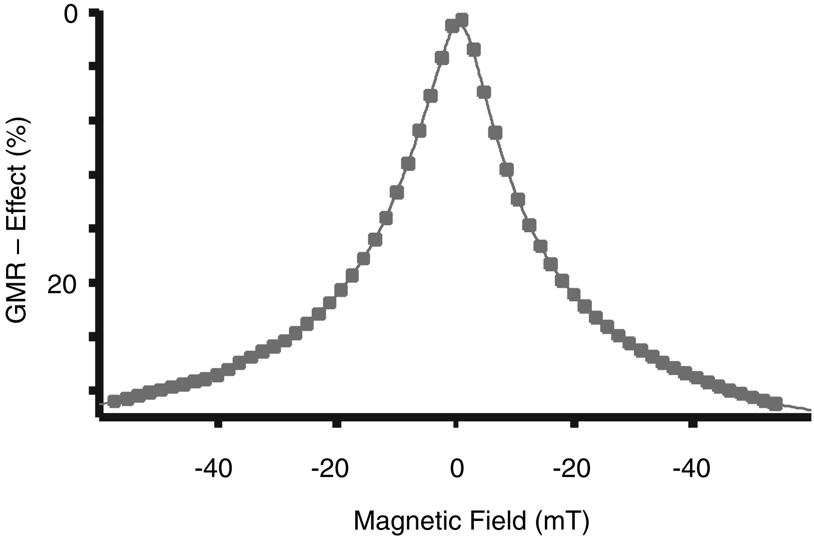


FIGURE 1 Change of the normalized electrical resistance (GMR effect) as a function of the magnetic field for a granular thin film of silver/cobalt prepared by molecular beam epitaxy. SOURCE: Alof, 2000.

the resistance. The change of the resistance as a function of the applied magnetic field is shown in Figure 1.

After the initial discovery of the GMR effect, other structures exhibiting insulating interlayers instead of the nonferromagnetic metals and spin valves with only a few nonperiodic layers have shown comparable resistance effects. The first case results in the Tunneling Magnetoresistance (TMR) effect, which has a tremendous potential for applications in nonvolatile data storage devices. Additionally, granular structures consisting of nanometer-size particles embedded in a metallic matrix have been examined because of the much simpler production and their increased stability at elevated temperatures. In this case, two metals, such as cobalt and silver, with no or little solubility, are codeposited onto heated substrates using sputtering or MBE techniques. The major disadvantage of granular structures is the large magnetic field required to obtain an effect. By optimization of the size, shape, and distribution of the nanoparticles in the matrix, the material can be used for sensor applications—in particular, for high temperature capabilities (Alof, 2000). In the system silver-iron, it was shown by comparison of the fraction of Fe atoms in the interfaces to the Ag matrix determined by Mössbauer spectroscopy with the observed GMR effect that the interfaces play a crucial role for the magnetoresistance effect (Alof et al., 2000). These measurements demonstrate the importance of the interfacial regions that

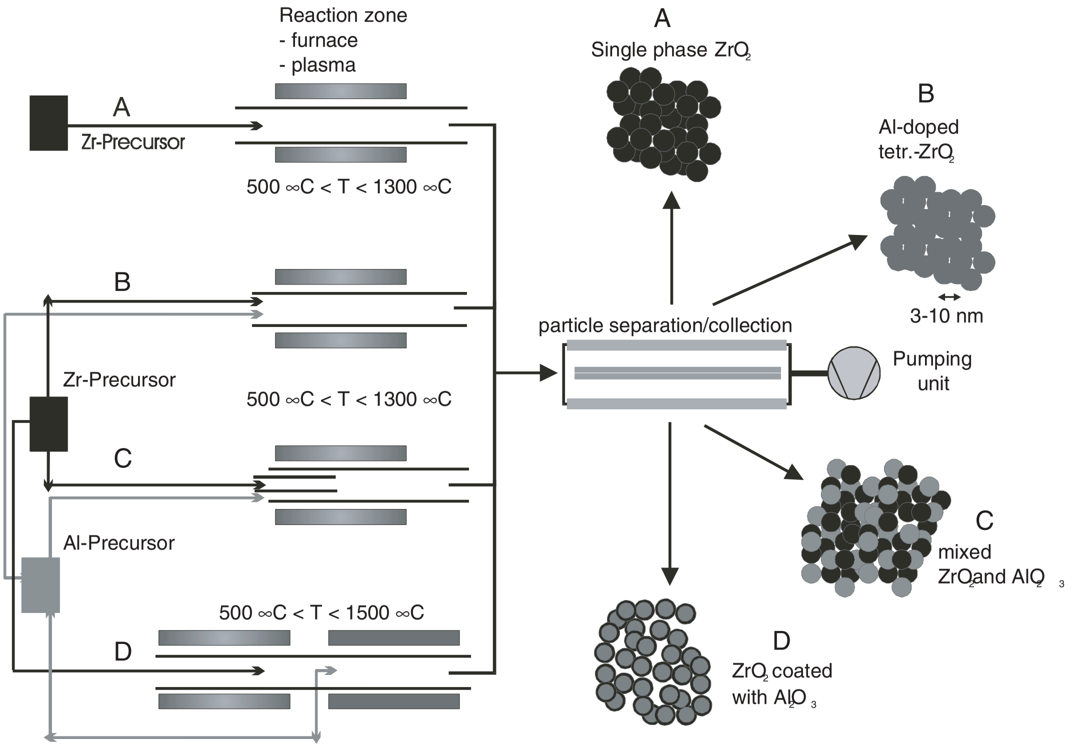


FIGURE 2 Schematic of the different opportunities of Chemical Vapor Synthesis for the preparation of nanocrystalline powders of ZrO_2 and in the system ZrO_2/Al_2O_3, with different elemental distribution. SOURCE: Hahn, private communication.

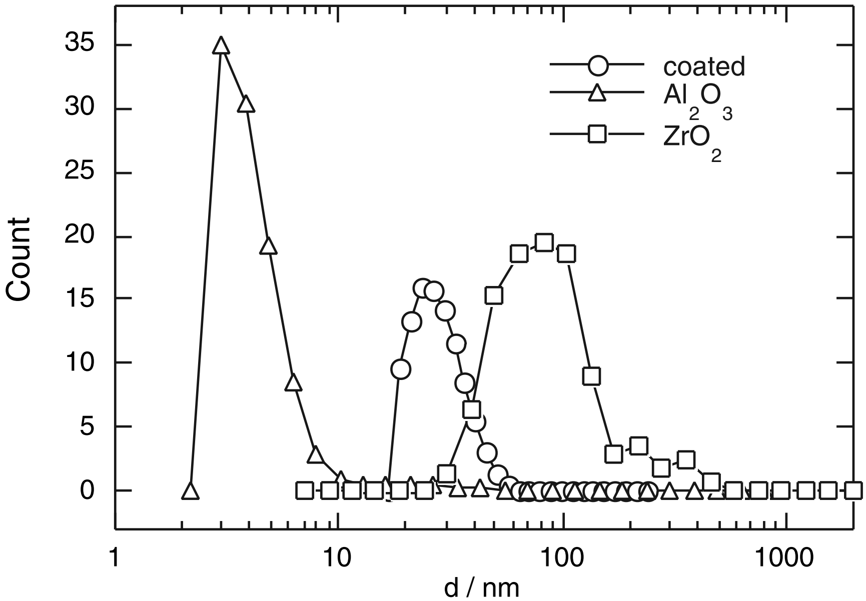


FIGURE 3 Size distribution of dispersions of the pure oxides ZrO_2 and Al_2O_3 in water. Note the improvement of the size distribution of ZrO_2 coated with a thin layer of Al_2O_3. The alumina nanoparticles can be dispersed to the size of the dry as prepared powder. SOURCE: Möller, 2000.

dominate in nanostructured materials and the control of structures and composition on a nanometer scale.

Nanoparticles synthesized by gas phase processes (IGC and CVS) can be used as prepared, after surface functionalization or after further processing, such as compaction and sintering, depending on the properties of interest. Examples are catalytic properties, gas sensing properties, enhanced sintering (Srdic et al., 2000), and superplasticity (Betz et al., 1997). While in zero-dimensional structures the drastic increase of the specific surface area as well as the details of the atomistic surface structure are important, three-dimensional structures are of interest because of the large fraction of internal interfaces that enhance grain-size-dependent properties. In Figure 2 the opportunities of the CVS method are schematically shown for the case of oxide nanoparticles consisting of several cations. The different elemental distribution (doped nanoparticles, binary mixtures on the nanometer scale of two different oxide nanoparticles, and coated/surface functionalized nanoparticles), which can be readily prepared using CVS, results in different properties at the same nominal composition. The effect of surface modification can be seen in Figure 3 for the dispersion behavior of ZrO_2, Al_2O_3, and Al_2O_3-coated ZrO_2 in water without the use of surfactants. The dispersion behavior is different for the different oxides, which is due to the surface charge, and can be adjusted by the surface composition even on the level of submonolayers. The total force between individual nanoparticles can be calculated theoretically by considering the various attractive and repulsive interaction forces. From the calculation it is determined that ZrO_2 nanoparticles will agglomerate into larger particles, whereas Al_2O_3 nanoparticles will exhibit a total repulsive force. As can be seen from Figure 3, the experimental particle size distribution for Al_2O_3 confirms the theoretical calculation. The predictability of the dispersion behavior is essential in many industrial processes involving dilute and concentrated particle dispersion in solvents.

REFERENCES

Alof, C. 2000. GMR-Efekt in granularen metallischen Schichten für Sensoranwendungen. Ph.D. dissertation. Technische Universität Darmstadt, VDI Fortschritt-Berichte, Reihe 9, Nr. 326, 1–150.

Alof, C., B. Stahl, M. Ghafari, and H. Hahn. 2000. Interface contribution to giant magnetoresistance in granular AgFe studied with Mössbauer spectroscopy. Journal of Applied Physics 88(7):4212–4215.

Betz, U., G. Scipione, E. Bonetti, and H. Hahn. 1997. Low-temperature deformation behavior of nanocrystalline 5 mol% yttria stabilized zirconia under tensile stresses. Nanostructured Materials 8:845–853.

Gleiter, H. 1989. Nanocrystalline materials. Progress in Materials Science 33:223–315.

Möller, A. 2000. Modellierung der Dispergierbarkeit nanokristalliner Al2O3- und TiO2-Pulver. Ph.D. dissertation. Technische Universität Darmstadt.

Srdic, V., M. Winterer, and H. Hahn. 2000. Sintering behavior of nanocrystalline zirconia doped with alumina prepared by chemical vapor synthesis. Journal of the American Ceramic Society 83(8):1853–1860.

Science and Technology of Nanotube-Based Materials

Otto Z. Zhou
Department of Physics and Astronomy and
Curriculum in Applied and Materials Science
University of North Carolina
Chapel Hill, North Carolina

Carbon nanotubes (Iijima, 1991; Dresselhaus et al., 1996) are a specific type of one-dimensional nanomaterial that has sparked people's imaginations. A cover story in *American Scientist* magazine a few years ago noted that carbon nanotubes could be used to build a space cable connecting the Earth and moon. News from NASA indicates that, in the near future, spacecraft may be based solely on carbon material—powered by either fuel cells based on carbon materials or lithium-iron batteries based on nanomaterials. Carbon materials also were featured very prominently in the recent national nanotechnology initiative and were mentioned in the President's State of the Union address, in which he referred to carbon nanotubes as a thousand times stronger than steel.

Are carbon nanotubes useful at all in the near term—within the next 5 years, rather than 20 or 30 years down the road? Let us go back to first-year chemistry to think about the materials that can be made from carbon. Carbon certainly is unique in the periodic table; it is truly multifunctional because of its ability to form either sp^2 or sp^3 bonds. It can form allotropes with very different structures. For example, diamond with sp^3 bonding has a closely-packed structure, a large elastic modulus, and is electrically insulating. Graphite with sp^2 bonding has a layer structure that is very strong within the layer due to the strong covalent bonds, but weak between the layers due to the weak *van der Waals* bonds. As a result, graphite can be used as a lubricant and battery electrode because ions can be stored between the graphite layers.

Buckyballs can be thought of as zero-dimensional structures. They are a truly nanoscale material, because the diameter of the molecule is about 1 nm (10^{-9} m). Assembled into a solid with weak intermolecular bonding, buckyballs have very interesting properties. For example, if C_{60} molecules are charged, they become superconductors, with the highest superconducting transient temperature of any organic superconductor (Andreoni, 2000).

Carbon nanotubes come with different chiral angles. Calculations and experimental results have shown that they can be either metallic or semiconducting depending on the chirality. This is very different from what we typically know of semiconductors, which have to be doped to make either a P-type or an N-type. Here, then, is a structure that intrinsically becomes either metallic or semiconducting.

Why should we be interested in carbon nanotubes? For one thing, there are not very many materials that have structural perfection at a molecular level as ideal as a single carbon nanotube. One can think about using their aspect ratio and small diameter for imaging applications. Also, they have very good mechanical properties and thermal properties. Theoretical calculations and measurements performed on individual carbon nanotubes have shown that their elastic modulus is as high as that of diamond, on the order of one terapascal. Indeed, if we could make a defect-free cable—one as long as we wanted—then a cable to connect the Earth and the moon would be within the realm of possibility.

With carbon nanotechnology and nanotubes, theory is ahead of experimental data. Because a carbon nanotube has a relatively simple structure, its properties can be readily calculated. For example, one can ask questions such as, “If I start with a perfect carbon nanotube and remove one carbon atom from the graphite lattice, what property would result?” Or, “How would twisting or bending change the electronic properties?” The results of these calculations are very interesting and suggest that the carbon nanotubes can function as the smallest sensors, transistors, and so on. However, the difficulty here is how to control materials structure and properties at the atomic and molecular level in actual experiments. Right now, there are three main techniques for making carbon nanotubes: laser ablation, arc-discharge, and chemical vapor deposition (Dresselhaus et al., in press). Each has its own advantages and disadvantages. Although rapid progress has been made in terms of synthesis of carbon nanotubes with selected diameters and orientation, there is still no effective method to control the nanotube chirality that is a prerequisite for utilizing carbon nanotubes in electronic devices.

Nanotechnology is becoming a reality in no small part due to advances in imaging technology such as scanning probe microscopy and electron microscopy. Because of its sharp tip and large aspect ratio, a carbon nanotube attached to the tip of an atomic force microscope (AFM) can significantly enhance the microscope’s resolution as demonstrated by researchers at Rice University (Dai et al., 1996). In addition, these functionalized AFM tips can be used to measure the binding energy of larger molecules—that is, the point at which they break or dissociate from the backbone structure (Wong et al., 1998). There are start-up companies now attempting to commercialize these kinds of AFM tips.

Carbon nanotubes are being investigated for chemical detection and storage applications. For example, automotive companies are looking at carbon nanotubes because of their reported (but controversial) high hydrogen storage capacities (Dillon et al., 1997). By replacing the carbon materials used as the anodes in current lithium-ion batteries with carbon nanotubes, it is possible to increase the battery lifetime. The advantage of carbon nanotubes is that, if processed such that their inner cores are accessible for diffusion, the storage capacity for both Li and H_2 can be significantly increased.

Another application of carbon nanotubes is as an electron field emitter. There are two ways one can take electrons out of a metal: one is to apply heat and the second is to apply a bias voltage so that electrons can escape to the vacuum level, which is termed electron field emission. Because of its small diameter and large aspect ratio, a carbon nanotube is expected, and has been shown experimentally, to have a low threshold electric field for electron emission (Zhu et al., 1999). This means that devices such as field emission displays (FED) can operate under lower voltages if carbon nanotubes are used as the emitters. An additional advantage compared to conventional electron emissive materials is processibility and high current capacity. Recent research has demonstrated emission current density as high as $4A/cm^2$ from a macroscopic nanotube film. This far exceeds the performance of other materials such as diamond and enables their application in devices with high current requirements such as microwave amplifiers.

Because of the potential market size, the possibility of utilizing carbon nanotubes as cold cathode materials for field emission displays has attracted considerable commercial interest. Several major display companies have devoted substantial R&D effort in this area. A prototype 9-inch field emission display based on carbon nanotubes has recently been demonstrated. Compared to the lithography-based Spindt-type emitters, the fabrication process can be simplified significantly by using carbon nanotubes as the cold cathodes. However, engineering issues such as emission uniformity and stability have yet to be solved.

Although the development of carbon nanotubes as practical engineering materials is still at an early stage, there are already indications of how they might be commercialized in the near future, particularly in devices that do not require a significant amount of material. For example, we have recently demonstrated applications of carbon nanotubes in gas discharge tubes used to protect houses and telephone lines from overvoltage (Rosen et al., 2000). This is a relatively simple device composed of two metal electrodes in an inert gas environment. In parallel with the electronic device to be protected, it acts as an insulator. The electronic device carries all the current, and it works fine. However, if there is a lightning strike or a voltage surge, electrons emitted from the metal electrode strike the inert gas and cause plasma breakdown, so a short is created that carries all the current density. The advantage of this device compared to a solid-state device is that it can carry very high current density, but it is not very reliable.

This is okay for telephone lines, but for high-speed internet data lines with more sensitive electronic components, this variation cannot be tolerated. We have shown that by coating the electrodes with carbon nanotubes, we can make gas discharge tubes that improve stability by a factor of 20. They are very stable to about a thousand surges, without much degradation.

Carbon nanotubes are, indeed, very interesting in the sense that the material is not only an ideal system to study fundamental science in one dimension but also has promising properties that can lead to some very practical devices in the near future.

REFERENCES

Andreoni, W., ed. 2000. The Physics of Fullerene-Based and Fullerene-Related Materials. Boston: Kluwer Academic Publishers.

Dai, H., J. H. Hafner, A. G. Rinzler, D. T. Colbert, and R. E. Smalley. 1996. Nanotubes as nanoprobes in scanning probe microscopy. Nature 384:147–151.

Dillon, A. C., K. M. Jones, T. A. Bekkedahl, C. H. Kiang, B. S. Bethune, and M. J. Heben. 1997. Storage of hydrogen in single-walled carbon nanotubes. Nature 386:377–379.

Dresselhaus, M. S., G. Dresselhaus, and P. C. Eklund. 1996. Science of Fullerenes and Carbon Nanotubes. San Diego: Academic Press.

Dresselhaus, M. S., G. Dresselhaus, and P. Avouris, eds. In Press. Topics in Applied Physics, Vol. 80. Heidelberg: Springer-Verlag.

Iijima, S. 1991. Helical microtubules of graphite carbon. Nature 354(6348):56–58.

Rosen, R., W. Simendinger, C. Debbault, H. Shimoda, L. Fleming, B. Stoner, and O. Zhou. 2000. Application of carbon nanotubes as electrodes in gas discharge tubes. Applied Physics Letters 76(13):1668–1670.

Wong, S. S., E. Joselevich, A. T. Woolley, C. L. Cheung, and C. M. Lieber. 1998. Covalently functionalized nanotubes as nanometre-sized probes in chemistry and biology. Nature 394:52–55.

Zhu, W., C. Bower, O. Zhou, G. Kochanski, and S. Jin. 1999. Large current density from carbon nanotube field emitters. Applied Physics Letters 75(6): 873–875.

Nanoscale Materials: Synthesis, Analysis, and Applications

Rudolf M. Tromp
IBM T. J. Watson Research Center
Yorktown Heights, New York

ABSTRACT

With the national initiative on nanotechnology, the art and science of the nanoworld have broken out of the lab and into the limelight of the media as well as the consciousness of politicians, business leaders, and the lay public. As in many nascent areas of science and technology, research advances in the lab are heralded as major breakthroughs promising faster computers, cheaper technologies, and better lives. But before concerning ourselves with the wonderful things that nanotechnology will do, we should look at some of the nanomaterials that will be at the bottom of this "food chain." What is the stuff of the nanoworld? How do we make it, how do we look at it, and, then, what might we do with it? Here, I touch on some of the broader issues and then focus on a particular example: the self-assembly of quantum dots and nanocrystals, the study of their evolution and structure, and their possible use in a variety of technologies.

As technologies continue to scale to ever-smaller dimensions, with 100-nm features in sight in semiconductor technology (Packan, 1999), excitement about materials on the nanoscale is heating up. In semiconductor technology, for example, scaling has been the driving force of the industry for decades. But it now appears that several fundamental barriers will limit scaling. Gate oxides are approaching a thickness of a mere 2 nm, where quantum mechanical tunneling becomes a serious concern because of its negative effect on device lifetime. The number of dopant atoms in the active region of a transistor becomes so small that statistical variations become significant, giving rise to undesirable device-to-device variations in operation. Also, it becomes increasingly hard to make such small devices because of the intrinsic resolution limits of traditional optical lithography, even with wavelengths pushing well into the ultraviolet. Similarly, the size of a magnetic bit on a storage disk is rapidly approaching the so-called paramagnetic limit, where thermal fluctuations destabilize the stored information, and reliable long-term storage becomes impossible. This hardly means that we are approaching the end of the computer revolution, but the material underpinnings of this revolution may be ready for a revolution of their own.

Some materials experts are undeterred by the limits of scaling. If scaling no longer works, the nanometer world has properties and promises of its own that go beyond scaling and that may create entirely new technologies based on new physics, materials, and paradigms. If this nanometer world is too small to be assembled by man, then let it assemble itself—a practice long known, practiced, and treasured by chemists and biologists. Clouds have a certain range of temperatures and contain a certain amount of water vapor. Under the right conditions, snowflakes, wonderfully intricate ice crystals, come falling out. They were not put together by anyone: they self-assembled. The worldwide drive to invest in and develop "nanotechnology" reflects the optimism that such a technology will soon become reality and that it will spawn new capabilities, opportunities, and wealth.

It may be useful to ask which structures on the nanoscale display what new phenomena and where these may be used in present or future technologies. For instance, what are "quantum dots," what properties do they have, and how can they be useful in creating new devices (Materials Research Society, 1998)? How do we make them? What level of control do we have, and what level do we need? How do we even look at them? This approach works from an evolving body of materials knowledge and involves looking for new ways to apply this knowledge. It relies on the assumption that if we know enough, applications will offer themselves almost unavoidably.

Alternatively, the general structure and properties of quantum mechanics have been explored to establish the new field of quantum cryptography and quantum computing (DiVincenzo, 1995). "Consider the superposition of two quantum states," a typical paper may state. What does this correspond with in the material world? How do we create such a superposition, and how do we manipulate it? Is this quantum world a nanoworld (atoms or quantum dots; Loss and DiVincenzo, 1998) or a more familiar microworld (superconductivity; Mooij et al., 1999). Does it matter which? Is the microworld a convenient testbed for the near future and the nanoworld the goal for a more remote future? Here the application is well defined, but the material basis is not. Quantum computers would do certain tasks much faster than digital computers. Will nanotechnology make practical quantum computing a reality?

There may be some discussion as to what nanotechnology actually is. For the present time it may be most useful to consider it the science of things on the nanoscale—anywhere from Angstroms to tenths of micrometers, which includes large chunks of materials science, physics, chemistry, and biology. Here, I look at the science of quantum dots—a good example of small structures formed by self-assembly.

QUANTUM DOT SELF-ASSEMBLY

There are at least two very different ways to make quantum dots assemble themselves. The first method (Nishi et al., 1996) consists of exposing a surface to a flux of foreign atoms or molecules: Take an atomically clean silicon (Si) surface and expose it to a vapor of germanium (Ge) atoms. The Ge atoms will at first wet the Si surface, growing a thin Ge layer over the Si wafer. However, Ge atoms are about 4 percent larger than Si atoms, giving rise to a misfit strain in this thin layer. After a thickness of three atomic layers has been reached, this misfit strain becomes too large, and any additional Ge atoms will aggregate in small, three-dimensional clusters. These Ge clusters or dots are so small—only 10 nm or so—that quantum mechanical effects play a significant role in their properties. Hence the name "quantum dots." This method works not only for Ge on Si but for a broad range of elements and alloys, both semiconducting and metallic.

The other approach involves using wet chemistry (Murray et al., 2000): Take a beaker and dissolve one constituent in a suitable solvent, then rapidly add a second constituent so that a new phase can nucleate in the now supersaturated liquid. This makes it possible to fabricate, for example, CdSe quantum dots in macroscopic quantities. After nucleation the quantum dots can be left to "ripen" for some time. In the ripening process the average dot size increases with time. This can be compared with the condensation of steam onto a cold surface. Initially, the condensed droplets are small but with time the droplets increase in size, even if no more steam condenses. This is because larger dots or droplets have less surface area per volume and are therefore energetically favored. The ripening process allows a certain degree of control over the average dot size. In the case of quantum dot growth in a liquid, a separate "distillation" step allows particles to be selected with excellent size control. Different sizes have different properties. Figure 1 shows fluorescence data (Bruchez et al., 1998) from compound semiconductor quantum dots over a range of sizes, showing a continuous shift of fluoresence wavelength with size and increasing to longer wavelengths for larger size as expected.

Although size selection and control are relatively straightforward in the wet chemistry approach, things are not so simple on surfaces. One problem is that there is no equivalent to the size distillation step that worked so well for the CdSe quantum dots shown above. Second, nucleation can occur over a relatively long time window, giving rise to broad size distributions. Both growth methods have their own advantages, which makes each more suitable for specific purposes. If one wants to put quantum dots in a solid-state device, such as a laser or a memory element, then growth on a solid surface is convenient, since it

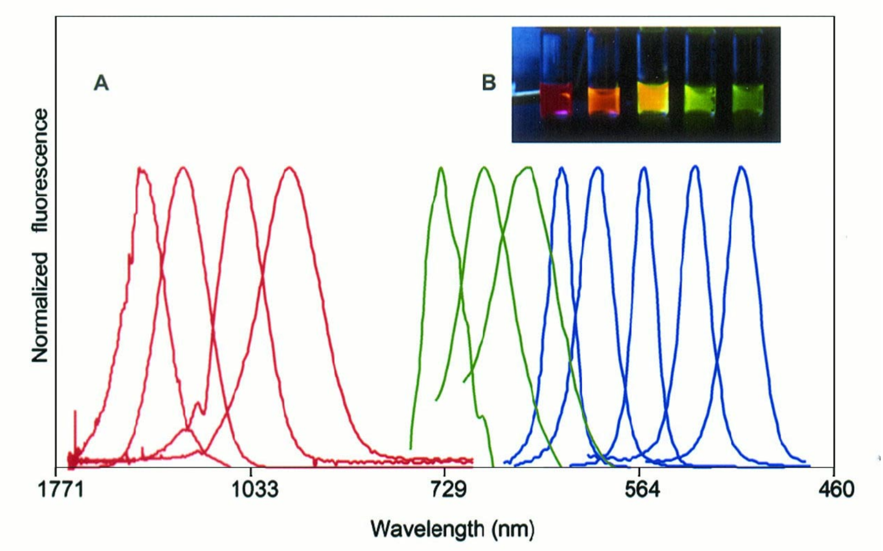


FIGURE 1 (a) Size- and material-dependent emission spectra of semiconductor nanocrystals. Blue: CdSe (diameters of 2.1, 2.4, 3.1, 3.6, and 4.6 nm, from right to left). Green: InP (3.0, 3.5, and 4.6 nm). Red: InAs (2.8, 3.6, 4.6, and 6.0 nm). (b) True-color image of silica-coated core (CdSe)-shell (ZnS or CdS) nanocrystal probes in aqueous buffer, illuminated with an ultraviolet lamp. This figure can be viewed in color by accessing http://www.cchem.berkeley.edu/~pagrp/publications.html (Bruchez et al.). SOURCE: Reprinted with permission from Bruchez et al. Copyright 1998 American Association for the Advancement of Science.

allows for direct integration of quantum dot growth in the device manufacturing process. However, one cannot, in general, remove these dots from the substrate on which they were grown and utilize them elsewhere. Wet chemistry methods allow for the growth of large quantities of quantum dots that can easily be "carried around" and used in a wide diversity of applications.

Other methods to create nanocrystals include physical vapor synthesis (in which a supersaturated vapor is seeded into a cold gas to nucleate nanocrystals), arc discharge methods (very effective in making buckyballs), and protein-mediated self-assembly methods, among others. Here I focus on wet chemistry methods and vapor deposition onto a surface.

HOW TO LOOK AT QUANTUM DOTS

What do quantum dots look like? Figure 2a shows a scanning tunneling microscopy image (Medeiros-Ribeiro et al., 1998) of Ge quantum dots grown on Si(001). Figure 2b shows a scanning electron microscopy image (Ross et al., 1999) of SiGe quantum dots. These images were obtained after the sample had been removed from the growth chamber. The dots come in two varieties: smaller dots that have a pyramid shape (P) and larger dots that have a dome shape (D). This is not a desirable feature. Ideally, all dots should have the same shape and the same size. Figure 2c shows an image of Ge quantum dots on Si(001) taken with an in situ transmission electron microscope (Ross et al., 1998) while the quantum dots were grown. The shape information is lost, but the size of the dots can be measured accurately, and growth can be followed in real time. This allows a detailed study of the time evolution of the quantum dot size distribution.

It has been found (Medeiros-Ribeiro et al., 1998; Ross et al., 1998) that the originally broad size distribution first bifurcates and then narrows significantly when, with continued growth, P dots disappear from the population and most dots take on the D shape. This process can be followed in even more detail with low-energy electron microscopy (Ross et al., 1999), as demonstrated in Figure 2d for SiGe quantum dots. Now both size and shape can be determined in detail, and in real time, during in situ quantum dot growth inside the microscope.

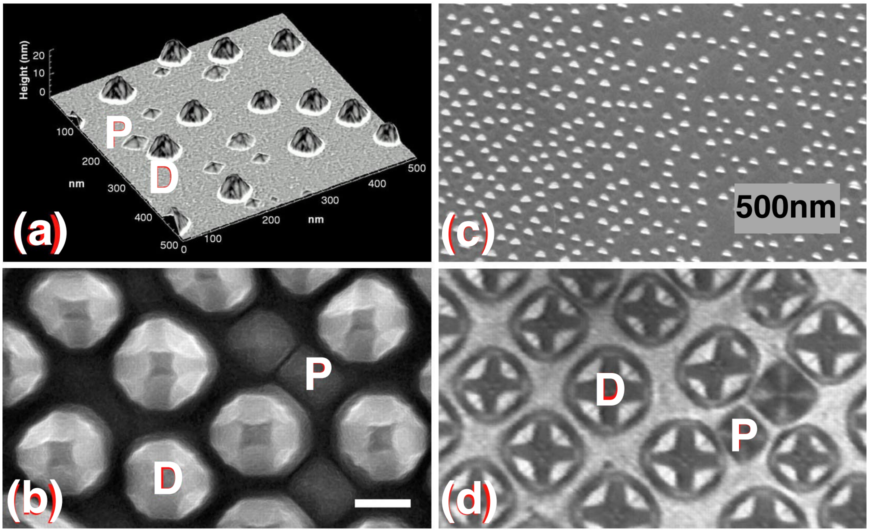


FIGURE 2 Images of Ge (a,c) and GeSi (b,d) quantum dots on Si(001): (a) Scanning Tunneling Microscopy, (b) Scanning Electron Microscopy, (c) *In-situ* Transmission Electron Microscopy, (d) *In-situ* Low Energy Electron Microscopy. Both pure Ge and GeSi alloy quantum dots show two distinct shapes, pyramids (P) at small volume and domes (D) at large volume. SOURCE: Panel a—Reprinted with permission from Medeiros-Ribeiro et al. Copyright 1998 American Association for the Advancement of Science. Panels b & d—Reprinted with permission from Ross et al. Copyright 1999 American Association for the Advancement of Science. Panel c—Reprinted with permission from R. Tromp (Ross et al., 1998). Copyright 1998 by the American Physical Society.

The change of the shape of quantum dots with increasing volume is not unique to SiGe. Theoretical studies (Liu et al., 1999) suggest that this is a rather universal phenomenon, although the detailed shapes are different for different systems. In the growth of quantum dots on a substrate, the shape is determined by the balance of surface and strain energies. For quantum dots grown in liquids, other factors may play a role as well. For instance, magnetic dots may change from spherical to ellipsoidal when, above a certain size, crystal anisotropies become significant.

Quantum dot properties can be studied by optical methods (Figure 1), by microscopic methods (Figure 2), and by a whole host of analytical techniques that are routinely applied to bulk materials. X-ray diffraction, magnetic measurements, and electrical transport measurements are examples of such techniques. Depending on the properties of interest, one chooses the appropriate analytical method. Some methods allow the study of single quantum dots. The microscopy methods shown in Figure 2 are examples, but optical studies of single quantum dots are also feasible (Nirmal et al., 1996). The electronic properties of single dots have been studied using electron energy loss in a special-purpose scanning transmission electron microscope (Batson and Heath, 1993). To date, magnetic measurements and X-ray diffraction measurements have only been performed on ensembles.

APPLICATIONS

The excitement about nanotechnology is based on the conviction that such novel materials as quantum dots have new applications that more conventional materials do not offer. Numerous applications have been proposed for quantum dots, some of which I review here, without any attempt to be complete.

Quantum dot lasers (Eberl, 1997) are among the earliest applications of quantum dots. The semiconductor laser inside a CD player uses a so-called quantum well sandwiched between Bragg reflectors for its laser action. The quantum well presents strongly confined electronic states in a two-dimensional sheet. Inversion can be obtained by injection carriers into the quantum well. The idea behind a quantum dot laser is that the quantum dots naturally provide strongly confined electronic states in zero dimensions. Quantum dot lasers have been successfully manufactured, although they have not displaced quantum well lasers. The problem is at least in part due to control. Molecular beam epitaxy and chemical vapor deposition, the leading methods for fabrication of quantum well lasers, give excellent control over the thickness and composition of the two-dimensional quantum well. Such control is more difficult for quantum dot growth. Relatively broad size distributions, as well as partial strain-driven intermixing with the substrate, lead to a broadened distribution of quantum dot electronic and optical properties and loss of laser efficiency. The development of quantum dot lasers continues to be pursued vigorously, and further advances might be seen in the future.

Color converters naturally come to mind when looking at Figure 1b. With the advent of solid-state ultraviolet lasers (GaN in particular), cheap sources of intense ultraviolet radiation are just over the horizon. Such lasers have the potential of becoming a "universal" light source when combined with suitable color converters. Display applications require red, green, and blue pixels. III-V quantum dots fluoresce in the full visible range and are a natural candidate for color conversion applications. Chemical self-assembly methods might enable bulk manufacturing at competitive cost.

Single Electron Transistors (SETs) (Matsumoto, 1996) use the extremely small electrical capacitance of quantum dots to create a novel type of field-effect transistor. Source and drain are coupled to the quantum dot by tunnel junctions. The quantum dot potential can be set by the use of a gate electrode in proximity to the quantum dot. As the gate potential is varied, the source-to-drain conductance oscillates with gate voltage intervals that correspond to the charging voltage of the quantum dot because of the addition of a single electron. The earliest SETs were often realized in two-dimensional electron gas structures in III-V semiconductors, but, more recently, a broad range of materials systems have been used, including Si. And although SETs initially only worked at cryogenic temperature, room temperature versions have now been reported. The temperature of operation is directly related to quantum dot size. With decreasing size the capacitance of a quantum dot becomes smaller, and the charging voltage due to the addition of a single electron increases correspondingly. This charging voltage must be significantly larger than the thermal energy, kT, if the device is to operate succesfully.

Much beautiful physics has been done using SETs, and both storage and logic circuits based on SETs have been proposed, although the practical development of such circuits has been slow. Of course, for commercial applications not a few circuits are required but hundreds of millions, even billions, on a single chip. Excellent uniformity and reproducibility of individual device operating characteristics are musts. Manufacturing control can be good in large devices (i.e., based on two-dimensional electron gas materials), but such devices work at cryogenic temperatures. Fabrication and control of room temperature devices with much smaller dimensions are much more difficult and have suffered from lack of control and device yield.

Quantum computing (DiVincenzo, 1995) is one of the hottest areas in information physics. For certain tasks, quantum computing promises to be far more efficient than digitial computing—at least on paper. At present, however, the theory of quantum computing far outstrips the practice. Experimentally, it is hard to construct the superposition of quantum states that quantum mechanics addresses with such ease. A number of different schemes are being pursued, including nuclear magnetic resonance-based methods (Gershenfeld and Chuang,

1998) and superconducting Josephson junction technology (Mooij et al., 1999). An alternative scheme (Loss and DiVincenzo, 1998) is based on the use of small ensembles of quantum dots. The requirements are quite daunting, particularly with respect to uniformity and control.

Nanocrystal semiconductor memory (Tiwari et al., 1996) is another application based on the small capacitance of quantum dots (or "nanocrystals," as they are often called in this context). In a conventional field effect transistor inversion is obtained by applying a suitable bias voltage to the gate. The incorporation of nanocrystals in the gate insulator would provide a means to apply a field offest by charging the quantum dots. Working devices have been successfully manufactured.

Patterned media (White et al., 1997) are widely discussed as the next step in magnetic storage. It has been proposed that, rather than use continuous magnetic thin films, researchers divide the film in spatially separated magnetic dots, where a single dot stores a single bit. This has certain advantages in terms of the ultimate storage density but is itself limited by the paramagnetic effect at a density of about 100 gigabits per square inch. When the bits become too small, the magnetic moment is subject to thermal fluctuations, and the stored information is subject to thermal decay. Recent attempts at fabricating magnetic nanocrystals have shown much progress, and studies of their magnetic properties are under way. As mentioned earlier, these studies represent ensemble averages over large numbers of particles, which significantly complicates matters. Even if the size distribution is very narrow, variations in orientation, together with crystal and shape anisotropies, may give rise to variations in magnetic behavior from particle to particle. For applications in patterned media, these particles must furthermore be placed in well-defined locations according to a precise architecture, over macroscopic areas. This entire research area is presently in a basic stage, with practical applications some years away.

Hard magnetic thin films (Sun et al., 2000) have been fabricated using magnetic nanocrystals, even though the use of such nanocrystals in patterned magnetic media is not imminent. The magnetic FePt nanocrystals are fabricated using wet chemistry methods discussed above. A substrate can be exposed to a solution containing these nanoparticles. As the solvent evaporates, the nanocrystals condense on the substrate, often in a crystalline network. Sun et al. (2000) recently demonstrated that such magnetic films, after suitable annealing, have magnetic properties that are superior to conventional magnetic films prepared by sputter deposition. These films represent an interesting application of quantum dots and nanocrystalline materials that is not to be overlooked: quantum dots as raw material in bulk quantities, used to assemble new nanostructural materials with superior mechanical, electrical, magnetic, or thermal properties.

Thin-film semiconductors are another example of such an application. Photoconductive properties of CdSe-based thin films have been explored, with potential solar cell applications in mind (Leatherdale et al., 2000). This may easily be broadened (at least conceptually) to other semiconductors. In thin-film transistor liquid crystal displays, amorphous Si is currently used for the switching elements. This material has rather poor mobility compared to crystalline Si, but deposition on a glass substrate is relatively straightforward and inexpensive. Alternatively, nanocrystalline Si deposited by, for instance, screen printing methods could be equally inexpensive, have competitive electrical properties, and may eliminate a few lithographic steps.

Nanostructural inks represent a generalization of the previous two applications. Manufactured in bulk, nanostructural inks can be applied to a wide variety of substrates using a wide variety of application methods, including inkjet printing, screen printing, spray application, spinning, immersion, and so forth. Inks can be mixed, layered, and patterned as applications require. Like the hard magnetic films and solar cells discussed above, new material properties might be realized with the resulting nanostructural solids, at low costs. Nanostructural inks might present the largest opportunity for the practical application of quantum dots and nanocrystals in the next two decades.

Tags are widely used in biology and medicine to allow visualization of specific structures or substances. Gold labeling is a classical example. Chemically prepared quantum dots usually have an organic coating that can be functionalized to interact with specific agents, and they have been used successfully for biolabeling (Bruchez et al., 1998; Chan and Nie, 1998), allowing optical observation by fluorescence. However, the application of quantum dots need not be restricted to biology. Tags are used everywhere for purposes of identification and detection, in areas ranging from research and development to manufacturing and to security. Quantum dot tags, manufactured cheaply and in bulk, offer an interesting alternative to conventional tags.

CONCLUSION

Nanoscale materials present many interesting opportunities and promises as semiconductor device scaling reaches fundamental limits in the near future. Such materials have novel properties that are of interest in a variety of applications—applications not limited to semiconductors. Over the last decade several methods have been developed for the self-assembly of quantum dots and nanocrystals, with varying degrees of control over size and properties. Chemical self-assembly is particularly promising for inexpensive, bulk manufacturing of such materials.

Unfortunately, but not surprisingly, surface and bulk energy terms strongly compete in determining the shape of quantum dots, and the shape may undergo significant changes with size. Therefore, size is not the sole characteristic that matters. Both in situ and ex situ microscopy methods have been developed that enable a detailed analysis of quantum dot size, shape, and evolution. Such studies have clarified some of the basic physics of quantum dot self-assembly, but it is hardly a field that has been explored exhaustively. Some properties (optical, electronic) can sometimes be studied for isolated dots, whereas other properties (magnetic) can only be studied in large ensembles. Many applications require ensembles of quantum dots with very narrow specifications.

Although novel devices have been proposed and some demonstrated in the lab, there are at present no devices, or even device concepts, that threaten to unseat Si. Quantum dot lasers have seen more practical progress but have not yet seen commercial use. Yet, there are numerous applications outside the semiconductor arena that are very promising. Color converters, nanostructural inks, and tags are examples of such applications. The use of nanotechnology on the more macroscopic scale may seem counterintuitive, but, on a short timescale, that might be where the largest opportunities are, with nanoscale devices winking at us from a more distant future.

ACKNOWLEDGMENTS

I gratefully acknowledge numerous colleagues at IBM and elsewhere for stimulating discussions, exchanges, and disagreements, as well as musings and philosophies on nanotechnology.

REFERENCES

Batson, P. E., and J. R. Heath. 1993. Electron energy loss spectroscopy of single silicon nanocrystals: The conduction band. Physical Review Letters 71(6):911–914.

Bruchez, M., Jr., M. Moronne, P. Gin, S. Weiss, and A. P. Alivisatos. 1998. Semiconductor nanocrystals as fluorescent biological labels. Science 281:2013–2016.

Chan, W. C. W., and S. Nie. 1998. Quantum dot bioconjugates for ultrasensitive nonisotopic detection. Science 281:2016–2018.

DiVincenzo, D. P. 1995. Quantum computation. Science 270:255–261.

Eberl, K. 1997. Quantum-dot lasers. Physics World 10(9):47–50.

Gershenfeld, N., and I. L. Chuang. 1998. Quantum computing with molecules. Scientific American. June. Online. Available: http://www.sciam.com/1998/0698issue/ 0698gershenfeld.html.

Leatherdale, C. A., C. R. Kagan, N. Y. Morgan, S. A. Empedocles, M. A. Kastner, and M. G. Bawendi. 2000. Photoconductivity in CdSe quantum dot solids. Physical Review B 62(4):2669–2680.

Liu, Q. K. K., N. Moll, M. Scheffler, and E. Pehlke. 1999. Equilibrium shapes and energies of coherent strained InP islands. Physical Review B 60(24):17008–17015.

Loss, D., and D. P. DiVincenzo. 1998. Quantum computation with quantum dots. Physical Review A 57(1):120–126.

Materials Research Society. 1998. Special issue on Semiconductor Quantum Dots. MRS Bulletin 23(2). Also available http://bloch.leeds.ac.uk/~ircph/maze/quantum-dot.html.

Matsumoto, K. 1996. Room temperature operated single electron transistor made by STM/AFM nano-oxidation process. Physica B 227(1–4):92–94.

Medeiros-Ribeiro, G., A. M. Bratkovski, T. I. Kamins, D. D. A. Ohlberg, and R. S. Williams. 1998. Shape transition of germanium nanocrystals on a silicon (001) surface from pyramids to domes. Science 279:353–355.

Mooij, J. E., T. P. Orlando, L. Levitov, L. Tian, C. H. van der Wal, and S. Lloyd. 1999. Josephson persistent-current qubit. Science 285:1036–1039.

Murray, C. B., C. R. Kagan, and M. G. Bawendi. 2000. Synthesis and characterization of monodisperse nanocrystals and close-packed nanocrystal assemblies. Annual Review of Materials Science 30:545–610.

Nirmal, M., B. O. Dabbousi, M. G. Bawendi, J. J. Macklin, J. K. Trautman, T. D. Harris, and L. E. Brus. 1996. Fluorescence intermittency in single cadmium selenide nanocrystals. Nature 383(6603):802.

Nishi, K., R. Mirin, D. Leonard, G. Medeiros-Ribeiro, P. M. Petroff, and A. C. Gossard. 1996. Structural and optical characterization of InAs/InGaAs self-assembled quantum dots grown on (311)B GaAs. Journal of Applied Physics 80(6):3466–3470.

Packan, P. A. 1999. Pushing the limits. Science 285(5436)**:**2079–2081.

Ross, F. M., J. Tersoff, and R. M. Tromp. 1998. Coarsening of self-assembled Ge quantum dots on Si(001). Physical Review Letters 80(5):984–987.

Ross, F. M., R. M. Tromp, and M. C. Reuter. 1999. Transition states between pyramids and domes during Ge/Si island growth. Science 286:1931–1934.

Sun, S., C. B. Murray, D. Weller, L. Folks, and A. Moser. 2000. Monodisperse FePt nanoparticles and ferromagnetic FePt nanocrystal superlattices. Science 287:1989–1992.

Tiwari, S., F. Rana, H. Hanafi, A. Hartstein, E. F. Crabbe, and K. Chan. 1996. A silicon nanocrystals based memory. Applied Physics Letters 68(10):1377–1379.

White, R. L., R. M. H. Newt, and R. F. W. Pease. 1997. Patterned media: A viable route to 50 Gbit/in/sup 2/ and up for magnetic recording? IEEE Transactions on Magnetics 33(1)(Part 2):990–995.

Dinner Speech

Bits Versus Atoms—The Future of Information Technology

Robert W. Lucky
Telcordia Technologies, Inc.
Red Bank, New Jersey

Earlier this year, I was asked to be on a National Academy of Engineering committee selecting the greatest engineering achievements of the last century. I thought it was going to be fun, but it wasn't. Everyone argued in favor of his or her own favorite thing, and it was fairly acrimonious in many ways. At first, the committee was not going to list the achievements in order, but I said, "You can't have a list and not order them. Nobody would care." So, in the end, we fought it out and did order them.

The criterion for making the list was the impact on society for the better, or how much the achievement bettered society. Number one was electrification. People pretty much agreed on that, because without electricity you really can't do much of anything. You take electricity for granted until there is a power failure, and then you're thrown back into the last century, and you realize how important it is. Number two was the automobile, which dramatically changed the way and where we live. Number three was the airplane. Some argued that a lot of people don't ride on airplanes and, therefore, they weren't something experienced firsthand by everyone. However, the airplane brought the continents together, and it was such a dramatic invention—the idea of men taking wing—that I think that aspect helped bring it up on the list. Number four was pure drinking water. Then came electronics, radio and television, the computer, telephone, air conditioning, the highway system, spacecraft, health technologies, petrochemical technologies, fiber optics, and so on down the list.

A couple of weeks ago I got a call from CNN, and they wanted to know why the Internet, which was ranked thirteenth, was so low on the list. CNN did a couple of interviews with people who said the committee didn't know what it was doing, and the Internet should have been ranked much higher. So, I had to do an interview and defend a decision that I didn't believe in, because I actually voted the Internet much higher. But the truth is that although there are certainly more than half a billion people in the world on the Internet, with six billion people in the world, less than 10 percent use the Internet, and the growth has slowed. For many years the number on the Internet doubled every year and then grew 60 percent per year for the past 2 years. Now, we track it every day, and it's growing about 50 percent a year. That's still very significant, but the truth remains that most of the world is not yet connected. However, I have thought about the Internet versus the other things on the Greatest Achievements list, and it has led into what I want to talk about today—that is, the nature of information technology.

Everything else on the list is a big physical thing that took 100 years to develop. The Internet is not a physical thing at all, and it's very recent. With electrification, for example, it took 50 years or so to bring electricity to everyone. The first electric power plant opened for business in 1882. The Rural Electrification Act goes back to the 1930s, when electricity was brought to rural areas. There were automobiles at the turn of the century, and highway systems—even the Romans had highways. What is the Internet after all? It's an idea. It's just a set of rules, protocols. It's not a thing at all.

Now what kinds of things will be on a list at the turn of the next century? Are there going to be big things like dams and great bridges? Last week I went through the computer museum, which is in Silicon Valley at Moffett Field. (They're building a new building, which will be open 3 years from now.) We were looking at all the old computers, such as the Sage system, the Stretch computer, the Enigma machine, and so on, and someone said, "Where's the software?" How do you show software in a museum? And yet that is what it's all made of. People gather round the old vacuum tubes and the old core memory, and they look at all this stuff and say, "Wow! That is neat." But the real stuff isn't shown there at all. Perhaps the list 100 years from now will be not big things at all but will be the virtual things—the ideas, the concepts, the software, whatever form those might take in the next 100 years.

Back when I was a kid in Pittsburgh, I grew up on a dead-end street in one of the suburbs. My dad went off to work every day, and I had no idea what he did. The only adults I saw working were the carpenters building houses along my street. At the end of the street was a big farm, and I saw the farmers working away and growing things, and I thought, "That's what adults do. They make things, they grow things." Then, one day after I grew up, I looked around and realized that I didn't know anybody who makes or grows anything. How do we get away with this? I walk down the halls of my company and other companies I visit, and I see silent rooms of people staring at CRTs and I say to myself, "What are they doing?" What is the nature of work today? Some of them look as though they're sleeping! The whole nature of what we do and the way we do it has changed.

Last fall I was chartered to do a paper for *The New Republic*. Apparently, the editors got into an argument and said, "Look. What makes you think information technology is going to last? There are lots of promises being made about information technology and how it's going to shape the economics of the future. But 100 years ago, those are the things they were saying about the railway. The railway was going to determine where commerce went, it was going to shape the nation, it was going to determine who made money and who didn't, and look, today, who cares about railroads anymore? What makes you think that information technology isn't going to be the same kind of thing?" Naturally, I had to defend this and say, "A hundred years from now information technology will still be big." But who knows? We don't have any clue of what it's going to be 100 years from now.

I thought about those trains. You see, a giant, powerful train is the epitome of the physical world. It throws out steam; the feeling of power, weight, force, and mass is there in the railroad. Information technology, however, is nothingness personified. It weighs nothing, is created from nothing, and is indestructible in many ways. The train is not that much different from what it was 100 years ago. How much better can it be? How much faster can it go? A train is subject to the rules of physics. But information technology—what rules is it subject to? We don't know. Is it infinitely expandable? It doesn't take any energy to create information. Information is nonrivalrous; I give it to you and I still have it, which is a troublesome thing I'll return to later.

A couple of years ago I visited Microsoft in Redmond, and one of their executives took me into their company store. It's about the size of a convenience store, like a 7-Eleven, and in it they have a rack that has all their products, which they sell to their employees for something like $19.00. I told the executive I was with, "I wish I was an employee so I could buy that product at this price." He nudged me and said, "We still make a profit." That is because those are empty boxes they are selling! This company at that time had the largest valuation of any corporation on earth, and they sold empty boxes. This is scary. In fact, I have to tell you that I met with Bill Gates, and he was predicting all the things that would happen, and I wrote them all down. I have them somewhere, and they are all absolutely wrong. I wish I could be wrong like that.

Here is another story that is indicative of the way the world works right now: Gates wrote the book *The Road Ahead*. One of my ex-bosses said that he visited a bookstore in Dallas, and there was a stack of Gates' book near the checkout counter. He noticed that affixed to the cover of every book was a sticker that read, "Recently updated to include the Internet." The road ahead.

This transformation of society and the way we work, from the world of atoms to the world of bits, is something that I have a hard time understanding. One of the things that epitomizes this to me is the death of the Heathkit. A lot of you are too young to have built one. I gave a talk in New Orleans to heads of about 300 electrical engineering schools, and I showed pictures of old Heathkits and asked how many had built one. Every single person had. There was something about working with your hands and building physical things. What happened to Heathkits? They just don't exist anymore. I've written a column for many years, and I've written a couple on Heathkits. I got more mail on those than anything else. People say, "They took it away from me." This sense of loss comes from having been able to create a physical thing with your own hands, understand it, and take pride in it, and now all that is gone. The beauty of how something works has sunk beneath the complexities. I hate it now when I have to go out and buy a new computer without caring what's inside it anymore. It's a tragedy to me that it's sunk into nothingness, it's disappeared in the misty veils of complexity and microcircuits, and it's no longer a physical, tangible thing.

Now, how does this new world work? There are a couple of laws I think about and try to understand. They're familiar to you, but you probably don't understand them either: Moore's Law and Metcalfe's Law. Now, everybody knows Moore's Law. It is incredible that progress in semiconductor technology has been at a constant exponential—a doubling of effectiveness every 18 months for 25 years. Why? Why is it exponential, and why is the period 18 months? To me this is critical, because progress is balanced on the precipice of chaos or stagnation. Think of a world in which the doubling period of Moore's Law was only 6 months. It's bad enough now that the computer you buy is obsolete in a couple of years, but imagine if it were obsolete so quickly that there was no stability in the world. Or, think of a world where Moore's Law stopped, and computers stopped getting faster. You can see, then, vastly different scenarios of how the world would work if Moore's clock ran at different rates. On the one hand, you would have stagnation, and your computer would be like an antique, perhaps coming in a mahogany case. On the other hand, it would be utter chaos; maybe the software wouldn't last because the platforms would disappear so quickly. Why that particular constant? Engineers and scientists all deal with exponentials all the time, but I really feel that none of us has an intuitive feel for an exponential. Many times in the past I have made serious management mistakes by underestimating Moore's Law and the power that doubling upon doubling upon doubling has.

There's an old parable about the inventor of chess. The emperor of China wanted to reward the inventor, so he asked him what he'd like for his reward. The inventor said, "I'd like a single grain of rice in the first square of the chessboard, and you can double it in the second and third and so forth. That's all I want." The emperor said, "Well, sure. No problem." Is it a problem? You all know mathematically that it's a problem, but do you know intuitively how much of a problem it is? I did a little web research on this. At the University of Maryland they have a physics experiment for freshmen, where they give them rice and say, "You do it. It'll give you a feel for how fast exponentials build up." It is quite easy in the first and second squares and so forth. In fact, the first half of the chessboard is really pretty easy. When you get to about square number 30, you need a wastebasketful of rice, and when you finish the first half of the chessboard at square 32, you need a closetful of rice. From then on, you're in trouble. By the time you get to the last square, the amount of rice you need would cover the entire area of the earth. Now here's the scary thought. So far, since the invention of the transistor in 1947, with doubling of power every 18 months, we have exactly covered the first half of the chessboard. The second half awaits us right now.

So, can this continue? All the papers that are written say that Moore's Law has to stop, because it runs into quantum limits. People have been writing these papers for 20 years, but they've all been wrong, because some fundamental assumption is found to be wrong. It could be that the drive for Moore's Law has such momentum, perhaps because the largest industry on earth depends on Moore's Law continuing, and the industry will make it continue. Moore's Law is not necessarily a law simply of feature sizes and semiconductors; rather, it is probably a law of functionality, which is an envelope of the succession of curves, each of which may top out. However, the envelope of all those curves continues to go up in functionality.

I have a secret thesis that all of technology is exponential in progress. The reason we first noticed it with Moore's Law is because we had a way of measuring the progress in that particular field by the feature sizes in circuits. It's hard to think of many other cases where there is a quantitative measure of the progress of technology, but there are some. Take the capacity of wireless communication or optics. There is a number. Each of those is also exponential, with different time constants. With optics it's 12 months doubling, and with wireless it's 9 months doubling. All measurable technologies, it seems, are exponential. Moore's Law may be some kind of law of nature.

There's an episode in Carl Sagan's book *Contact* where the character played by Jodie Foster in the movie goes to visit the aliens. I am paraphrasing here, but before she leaves them, the aliens say, "All right. We will answer one question. What one question do you want to ask to take back to earth?" She thinks and says, "Do you believe in God?" The aliens say, "Not as you define it. But there is something that bothers us. You have a number you call pi, 3.14159, and you've run it out to thousands and thousands of places. But our computers are much, much better than yours and we've run it out millions and billions of places and something strange happens. When you get far out the decimal digits turn to ones and zeros." She says, "What does it mean?" The alien says, "We don't know. But there's a message encoded in the structure of the universe." I told this to a mathematician friend, and he said, "Well, it's an irrational number. All sequences are in there, so it's to be expected." But I think there's a mysterious law here of exponential progress in technology.

I think Moore's law is so fundamental to what's happening in information technology that it's going to continue for another 100 years. How could it go past the next half of the chessboard? And, if it stops, what happens? What happens to the industry? What happens to all the edifices that we have built on the continuation of this growth? However, perhaps like the train, it stops getting better, and that's it. I don't think so. I think progress will continue, and I have no idea what this awesome exponentiation will realize as we go ahead.

Let me talk about the other great law, Metcalfe's Law. Economists know it as the law of increasing returns—of network externalities—but the idea is that the more people that are connected to a network, the more valuable the network is. Specifically, the value of a network grows by the square of the number of users. The value is measured by how many people I can communicate with out there, so the total value of the network grows as the square of the number of users. Now, what this means is that a small network has almost no value and a large network has a huge value. What it gives you is the lock-in phenomenon of winner takes all. You want to have the same thing as everybody else. Examples include the classic case of Betamax versus the VHS standard in VCRs, or Microsoft Word versus WordPerfect. You don't want to have something different from other people because it has less value. There are so many things that are like that in networks, where everybody wants to have the same thing because that's where the maximum value is. Whoever has what tips the balance and becomes what everybody wants gets a monopoly. This phenomenon stops a lot of applications. The videotelephone is a classic example, because if only one person has a videotelephone, it's useless. The videotelephone only becomes useful when a lot of other people have it.

I remember back in the 1960s, when the AT&T Picturephone was first developed. The mathematicians did a study of the rate at which the market would adopt the Picturephone, and they predicted a curve that was very much like Metcalfe's Law. They modeled it after the spread of the plague. The idea is that you don't want to be the first person on your block to get the plague. But when all your friends get it, you think about getting it. The more people have it, the more you're likely to get it, and suddenly there is this capture effect, where everybody has it. This law of network externality governs so much of the business and is at the heart of the Microsoft trial. Why does Microsoft have a monopoly? Is this a natural phenomenon that has to do with networks?

I heard a talk recently by Brian Arthur, an economist at the Sante Fe Institute, who is credited with writing the original paper on network externalities. He's got a new law that he unabashedly calls Arthur's Law, which is "Of networks there will only be a few." This really applies to the law of network externalities, but the kind of networks he's talking about are not AT&T versus AOL or anything like that. He's talking about customer association-type networks—for example, how many credit cards will there be? The answer is only a few, because you want to have the same thing that's accepted everywhere else. Arthur's Law doesn't apply to everything, but he gives two specific examples that would be familiar to all of us. One is Amazon. If Amazon doubles its network or number of users, it isn't worth a lot more to the individual consumer, because all the company may do is add a few more products to meet the needs of this expanded number of users. However, if eBay doubles it network, is it worth more to you? Absolutely yes. Yahoo is trying to compete with eBay, but the critical mass is with eBay.

David Reed coined another law—Reed's Law—that says there's something beyond Metcalfe's Law. There are three kinds of networks. First, there's broadcast, such as radio and TV, which we'll call a Sarnoff network. The value of that network is proportional to the number of people receiving the broadcast. Amazon would fall into this type of network, because people shop there but don't interact with each other. Then there's the Metcalfe's Law-type network, where people talk to each other—for example, classified ads. Reed says that the important thing about the Internet is neither of those. The Internet exhibits a third kind of law, where communities with special interests can form. The thing about communities is there are 2^n of them, so in a large network, the value of having so many possible communities and subnetworks is the dominant factor. He predicts a scaling of networks, starting with small networks having only the Sarnoff linear factor, larger networks dominated by the square factor, and giant networks dominated by the 2^n factor of the formation of communities.

Napster is another example of what's going on in information technology. First, it's an example of the kind of network in which winner takes all. Napster is where all the songs are, so that's where everybody else is. If Napster goes under, then all the little sites won't be able to replace it, because people won't find what they want there. Napster also brings up one of the other properties of information, which is troublesome and is going to shape our society in the coming years; the idea that information can be copied perfectly at zero cost. That flies in the face of so much of what we believe about commerce. As my friend Douglas Adams said to me, we protect our intellectual property by the fact that it's stuck onto atoms, but when it's no longer stuck onto atoms, there is really no way to protect it. He would like to sell his books at half a cent a page, the idea being that for every page you read, you pay him half a cent. If you get into the book 20 pages and say, "This book is really bad," you don't pay anymore. That would eliminate the copying of information at zero cost issue that he experiences as an author. He says people come up to him in the street and say, "I've read your book 10 times," and he says, "Yes, but you didn't pay 10 times."

So these are some of the things that trouble me about the future of information technology. What are its limits? Will the laws of network effects doom us all to a shared mediocrity? What will happen to intellectual property and its effect on creativity? Is it like the railroads, or is this something fundamentally different that will last through the next century?

Whatever the answers are, I'm sure that when we make a list 100 years from now, it's going to be dramatically different from the list that we made this year. The big things of the last century—spacecraft, highways, the automobile, and the airplane—may not characterize progress in this new century.

Appendixes

Contributors

DAVID BARAFF joined Pixar Animation Studios in 1998 as a senior animation scientist. Prior to his arrival at Pixar, he was an associate professor of robotics and computer science at Carnegie Mellon University. Dr. Baraff received his Ph.D. from Cornell University in 1992, where he was a graduate student in Cornell's Department of Computer Science and Program of Computer Graphics. Before and during his graduate studies, he also worked at Bell Laboratories' Computer Technology Research Laboratory doing computer graphics research, including real-time 3-D interactive animation and games. After receiving his Ph.D., he joined the faculty of Carnegie Mellon University. In 1995, Dr. Baraff was named an ONR Young Investigator. His research interests include physical simulation and modeling for computer graphics, robotics, and animation. *(deb@pixar.com)*

ERIC D. GREEN received his M.D. and Ph.D. from Washington University School of Medicine (St. Louis, Missouri) in 1987, studying the structure and biosynthesis of oligosaccharides on the pituitary glycoprotein hormones for his Ph.D. thesis work. During his residency training in clinical pathology, Dr. Green worked in the laboratory of Maynard Olson, where he developed approaches for utilizing yeast artificial chromosomes to construct physical maps of DNA. His work also included initiation of a project to construct a complete physical map of human chromosome 7 within the Washington University Genome Center—one of the first funded Centers in the Human Genome Project. In 1992, Dr. Green became an assistant professor of pathology, genetics, and medicine as well as a co-investigator in the Human Genome Center at Washington University. In 1994, he moved his research laboratory to the intramural program of the National Center for Human

Genome Research (recently renamed the National Human Genome Research Institute) at the National Institutes of Health, where he now serves as head of the Physical Mapping Section, chief of the Genome Technology Branch, and director of the NIH Intramural Sequencing Center. Dr. Green's research focuses on the mapping and sequencing of mammalian genomes and the isolation and characterization of genes causing genetic diseases. *(egreen@nhgri.nih.gov)*

HORST W. HAHN is vice chairman of the Department of Materials Science, professor of materials science, and head of the Thin Films Division at Darmstadt University of Technology, Germany. Previously, he was an associate professor of materials science with tenure at Rutgers University, a research assistant professor at the University of Illinois, Urbana-Champaign, and a research associate with the Materials Science Division at Argonne National Laboratory. Following the completion of his Ph.D., Dr. Hahn was the recipient of a fellowship from the Fritz-Thyssen Foundation and did a postdoc in the Department of Materials Science at the University of Saarland, Saarbrücken, Germany. Included in his professional activities are: associate editor for *Materials Letters*; past associate editor for *Nanostructured Materials*; member and chairman (since September 1999) of the Scientific Advisory Board of Hahn-Meitner Institute in Berlin; member of the Scientific Advisory Board of Sachs-Center at Technion, Haifa, Israel; and member of the International Committee for a series of nanotechnology conferences. Dr. Hahn is a member of the Materials Research Society, the German Materials Society, and the German Physical Society. He received his M.S. in materials science from the University of Saarland and his Ph.D. in metallurgy and materials science from the Technical University, Berlin. *(hhahn@tu-darmstadt.de)*

MARK D. JENKS is Boeing/MSFC International Space Station chief engineer. He is responsible for International Space Station engineering activities at Marshall Space Flight Center in Huntsville, Alabama, including design, development, and testing of the U.S. laboratory and joint U.S./Russian airlock elements; common hatches and berthing mechanisms; and payload racks. Prior to this assignment, Mr. Jenks was program manager for the ISS airlock element, and prior to that, managed the Integrated Product Development Teams responsible for the primary structure and external outfitting of the Station's first U.S. element, the Unity Node, launched in December 1998. Before coming to Huntsville in early 1996, Mr. Jenks managed Boeing's Helicopters Division Development Center in Philadelphia. He has also held positions in manufacturing technology, tool engineering, internal audit, project engineering, and aerodynamics research. A Boeing employee since 1983, Mr. Jenks was selected by Boeing for the MIT Leaders for Manufacturing Program in 1989 and received masters degrees in management and materials engineering. He also holds B.S. and M.S. degrees in aeronautical engineering from Rensselaer Polytechnic Institute. *(mark.d.jenks@boeing.com)*

RONNY KOHAVI is the director of data mining at Blue Martini Software in San Mateo, Calif., where he leads the engineering group responsible for the data collection and analysis modules in the company's Customer Interaction System (CIS). Prior to joining Blue Martini, Dr. Kohavi managed the MineSet project, Silicon Graphics' award-winning product for data mining and visualization. He joined Silicon Graphics after getting a Ph.D. in machine learning from Stanford University, where he led the MLC++ project, the Machine Learning library in C++ now used in MineSet and for research at several universities. Dr. Kohavi received his B.A. from the Technion, Israel. He co-chaired KDD 99's industrial track with Jim Gray and the KDD Cup 2000 with Carla Brodley. He co-edited (with Foster Provost) the special issue of the journal *Machine Learning* on Applications of Machine Learning and the special issue of the journal *Data Mining and Knowledge Discovery* on Applications of Data Mining to Electronic Commerce (to appear in 2001). *(ronnyk@cs.stanford.edu)*

DOUGLAS A. LAUFFENBURGER is the co-director, Division of Bioengineering and Environmental Health, and director, Biotechnology Process Engineering Center, at the Massachusetts Institute of Technology. His area of work is molecular cell bioengineering—the application of engineering approaches to develop quantitative understanding of cell function in terms of molecular properties for improved design of cell-based technologies. Dr. Lauffenburger is the recipient of an NSF Presidential Young Investigator Award, NIH Career Development Award, the Allan P. Colburn Award from AIChE, a J.S. Guggenheim Foundation Fellowship, the Curtis W. McGraw Research Award from ASEE, and the Food Pharmaceutical & Bioengineering Division Award of the AIChE. He is a founding fellow of the American Institute for Medical and Biological Engineering, currently chair-elect of the College of Fellows, and has served as president of the Biomedical Engineering Society. Dr. Lauffenburger received his B.S. from the University of Illinois and Ph.D. from the University of Minnesota. *(lauffen@mit.edu)*

ROBERT W. LUCKY is corporate vice president for applied research at Telcordia Technologies, Inc. in Red Bank, New Jersey. Dr. Lucky attended Purdue University, where he received a B.S. degree in electrical engineering in 1957, and M.S. and Ph.D. degrees in 1959 and 1961. After graduation, he joined AT&T Bell Laboratories in Holmdel, New Jersey, where he was initially involved in studying ways of sending digital information over telephone lines. The best known outcome of this work was his invention of the adaptive equalizer—a technique for correcting distortion in telephone signals that is used in all high speed data transmission today. The textbook on data communications that he co-authored became the most cited reference in the communications field over the period of a decade. At Bell Labs he moved through a number of levels to become executive director of the Communications Sciences Research Division in 1982, where he was responsible for research on the methods and technologies for future communication systems. In 1992 he left Bell Labs to assume his present position at Telcordia Technologies.

Dr. Lucky has been active in professional activities and has served as president of the Communications Society of the Institute of Electrical and Electronics Engineers (IEEE) and as vice president and executive vice president of the parent IEEE itself. He has been editor of several technical journals, including the *Proceedings of the IEEE*, and since 1982, he has written the bimonthly "Reflections" column of personalized observations about the engineering profession in *Spectrum* magazine. In 1993 these "Reflections" columns were collected in the IEEE Press book *Lucky Strikes . . . Again*. Dr. Lucky is a fellow of the IEEE and a member of the National Academy of Engineering. He is also a consulting editor for a series of books on communications through Plenum Press. He has been on the advisory boards or committees of many universities and government organizations and was chairman of the Scientific Advisory Board of the United States Air Force from 1986-1989. He was the 1987 recipient of the prestigious Marconi Prize for his contributions to data communications, and he has been awarded honorary doctorates from four universities. He has also been awarded the Edison Medal of the IEEE and the Exceptional Civilian Contributions Medal of the U.S. Air Force. Dr. Lucky is a frequent speaker before both scientific and general audiences. He is the author of the popular book *Silicon Dreams*, which is a semi-technical and philosophical discussion of the ways in which both humans and computers deal with information. *(rlucky@telcordia.com)*

MARK W. MAIER is senior engineering specialist, Engineering and Technology Group, Reconnaissance Systems Division, The Aerospace Corporation. In his current position, Dr. Maier leads the systems architecture training program and is the director of an advanced concepts analysis group for part of the intelligence community. Formerly, Dr. Maier was an associate professor with tenure in the Department of Electrical and Computer Engineering at the University of Alabama in Huntsville. From 1983–1992, he worked at Hughes Aircraft Company. Dr. Maier was the recipient of the Aerospace Institute Achievement Award in 1999 and 2000, and an International Council on Systems Engineering 1997 Professional of the Year award. He is a senior member of IEEE and active as a committee chair, working group member, and reviewer. Dr. Maier has a Ph.D. in electrical engineering from the University of Southern California, and M.S. and B.S. degrees from the California Institute of Technology. *(Mark.W.Maier@aero.org)*

PILAR N. OSSORIO is assistant professor of law and medical ethics at the University of Wisconsin at Madison. Prior to taking her position at UW, she was director of the genetics section at the Institute for Ethics of the American Medical Association. Dr. Ossorio received her Ph.D. in microbiology and immunology in 1990 from Stanford University. She went on to complete a post-doctoral fellowship in cell biology at Yale University School of Medicine. Throughout the early 1990s, Dr. Ossorio also worked as a consultant for the federal program on the Ethical, Legal, and Social Implications (ELSI) of the Human Genome Project, and in 1994 she took a full-time position with the Department of Energy's ELSI program. In 1993, she served on the Ethics Working Group for President Clinton's Health Care Reform Task Force. Dr. Ossorio received her J.D. from the University of California at Berkeley School of Law in 1997. She was elected to the legal honor society Order of the Coif and received several awards for outstanding legal scholarship. Dr. Ossorio is a fellow of the American Association for the Advancement of Science (AAAS), a past member of AAAS's Committee on Scientific Freedom and Responsibility, a member of the National Cancer Policy Board, and a member or chair of several working groups on genetics and ethics. She has published scholarly articles in bioethics, law, and molecular biology. *(pnossorio@facstaff.wisc.edu)*

LYNNE J. REGAN is a professor in the Department of Molecular Biophysics and Biochemistry and the Department of Chemistry at Yale University. After earning her B.A. from Oxford University, Dr. Regan received a Fulbright-Hayes Scholarship and attended the Massachusetts Institute of Technology, from which she received a Ph.D in biochemistry and molecular biology. She did postdocs at E.I. du Pont de Nemours & Company and the MRC Laboratory of Molecular Biology in Cambridge, England. She joined the Yale faculty in 1990. Dr. Regan is the recipient of an NSF Young Investigator Award, the Camille Dreyfus Teacher-Scholar Award, the Biophysical Society's Margaret O. Dayhoff Award, and most recently, the Herbert W. Dickerman Award. *(lynne@nero.csb.yale.edu)*

PETER SCHRÖDER is an associate professor of computer science at the California Institute of Technology. Prior to Caltech, he was a postdoctoral research fellow and lecturer at the University of South Carolina. Dr. Schröder received his Ph.D. in computer science from Princeton University. Prior to Princeton, he was a member of the technical staff at Thinking Machines, where he worked on graphics algorithms for massively parallel computers. In 1990, he received an M.S. degree from the Massachusetts Institute of Technology's Media Lab. He did his undergraduate work at the Technical University of Berlin in computer science and pure mathematics. He has also held an appointment as a visiting researcher with the German national computer science research lab (GMD) and its visualization group. Dr. Schröder is a world expert in the area of wavelet-based methods for computer graphics. He helped pioneer the use of fast wavelet solvers for illumination computations and developed with Wim Sweldens the first practical spherical wavelet transform. Multiresolution techniques have been the subject of many invited lectures and courses he has given in Europe and North America for academic and industrial audiences. His publications record ranges from *Wired* magazine to Siggraph conferences and special scientific journal issues on wavelets. In 1995 he was awarded an NSF Career Award and named a Sloan Fellow. More recently he was named a Packard Fellow. *(ps@cs.caltech.edu)*

MARVIN M. THEIMER is a senior researcher at Microsoft Research, exploring topics in peer-to-peer computing and Internet infrastructure. Dr. Theimer received his Ph.D. in computer science from Stanford University in 1986. He spent two years with the QuickSilver distributed operating system project at IBM's Almaden Research Center. Following that, he spent almost 10 years at Xerox's Palo Alto Research Center exploring the topics of ubiquitous computing and weakly consistent, replicated systems. Dr. Theimer joined the Systems & Networking group at Microsoft Research in late 1998. *(theimer@microsoft.com)*

RUDOLF M. TROMP is manager for analytic science at IBM's T.J. Watson Research Center and is currently on a 1-year assignment to the IBM Corporate Technology Council staff. After obtaining his Ph.D. in physics from the University of Utrecht, The Netherlands, Dr. Tromp joined the T.J. Watson Research Center in 1983. His scanning tunneling microscopy studies revealed the Si(001) dimer structure for the first time, as well as the spatial distribution of the Si(111)(7x7) electronic surface states, and their relation to the underlying atomic structure. Using medium energy ion scattering, he invented surfactant mediated epitaxial growth, a method which allows improved control over the morphology of epitaxial films and superlattices. More recently, his work has focused on the dynamics of surface and interface processes such as phase transitions, chemisorption and etching, epitaxial growth, and issues in nanotechnology. He has developed several *in situ* electron microscopy methods that allow detailed, real-time observations of such processes with high spatial resolution. These studies have shed new light on the thermodynamics of epitaxial growth, the dynamic evolution of the morphology of epitaxial films, the self-assembly of quantum dots, the spatio-temporal character of first- and second-order phase transitions at surfaces, and the growth of organic semiconductor films. Dr. Tromp is the recipient of the Wayne B. Nottingham Prize of the Physical Electronics Conference (1981), the Materials Research Society Medal (1995), and four IBM Outstanding Innovation and Technical Achievement Awards. He is a fellow of the American Physical Society and of the Boehmisce Physical Society. He is the holder of four U.S. patents. *(rtromp@us.ibm.com)*

OTTO Z. ZHOU is an associate professor of materials science and physics and the director of the North Carolina Center for Nanoscale Materials at the University of North Carolina at Chapel Hill. His research is focused on synthesis and solid state properties of nanoscale materials and their energy-storage, vacuum microelectronics, and nano-composite applications. He is the founding director of the North Carolina Center for Nanoscale Materials, which has 15 associated faculty members from the University of North Carolina at Chapel Hill, North Carolina State University, and Duke University. Dr. Zhou has been active in the field of carbon fullerenes and nanotubes since their discovery. He has played an important role in synthesis and characterization of new fullerene-based superconductors; one of the compounds he synthesized holds the current record high transition temperature among all C_{60} superconductors. His research on carbon nanotubes has demonstrated that these novel materials have promises in vacuum microelectronics, energy storage, and nanocomposites applications. His group has developed techniques for controlled fabrication and processing of nanotube materials; investigated their structure, mechanical and electronic properties; and fabricated nano-devices. He has published over 60 refereed journal articles (including 11 in *Nature* and *Science*) and 5 book chapters and holds 8 granted and pending U.S. patents on nanotechnology. He has given over 30 invited talks in national and international conferences, and has co-organized several symposia on nanomaterials and nanotechnology. Dr. Zhou received his doctorate in materials science from the University of Pennsylvania in 1992, and he was a post-doctoral member at the Bell Laboratories in Murray Hill, N.J. After spending a year at the NEC Research Laboratory in Japan, he joined UNC Chapel Hill as an assistant professor. *(zhou@physics.unc.edu)*