Defining the Mandate of Proteomics in the Post-Genomics Era: Workshop Report

Steering Committee for Defining the Mandate of Proteomics in the Post-Genomics Era, U.S. National Committee for the International Union of Biochemistry and Molecular Biology, Policy and Global Affairs Division, National Research Council

ISBN: 0-309-54226-X, 55 pages, 8.5 x 11, (2002)

DEFINING THE MANDATE OF PROTEOMICS IN THE POST-GENOMICS ERA

WORKSHOP REPORT

STEERING COMMITTEE FOR DEFINING THE MANDATE OF PROTEOMICS IN THE POST-GENOMICS ERA

U.S. National Committee for the International Union of Biochemistry and Molecular Biology
Board on International Scientific Organizations
Policy and Global Affairs Division
and
Board on Life Sciences
Division of Earth and Life Studies

NATIONAL RESEARCH COUNCIL
OF THE NATIONAL ACADEMIES

THE NATIONAL ACADEMIES PRESS
Washington, D.C.

www.nap.edu

THE NATIONAL ACADEMIES PRESS 500 Fifth Street, N.W. Washington, DC 20001

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

This study was supported mostly by industry contributions (see Acknowledgments) and by Contract/Grant No. 0222688 between the National Academy of Sciences and the National Science Foundation and Contract/Grant No. N01-OD-4-2139 between the National Academy of Sciences and the National Institutes of Health. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the organizations or agencies that provided support for the project.

This report is available online at http://www.nap.edu

Printed in the United States of America

THE NATIONAL ACADEMIES
Advisers to the Nation on Science, Engineering, and Medicine

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. Wm. 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. Harvey V. Fineberg 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. Wm. A. Wulf are chair and vice chair, respectively, of the National Research Council.

www.national-academies.org

.

STEERING COMMITTEE FOR DEFINING THE MANDATE OF PROTEOMICS IN THE POST-GENOMICS ERA

George Kenyon
University of Michigan, *Chair*

David M. DeMarini
U.S. Environmental Protection Agency

Elaine Fuchs
Rockefeller University

David Galas
Keck Graduate Institute of Applied Life Sciences

Jack Kirsch
University of California, Berkeley

Walter Moos
MitoKor

Gregory Petsko
Brandeis University

Dagmar Ringe
Brandeis University

Gerald Rubin
Howard Hughes Medical Institute

Contributing author: Thomas Leyh
Albert Einstein College of Medicine

National Research Council Staff
Board on International Scientific Organizations

Laura Sheahan
Program Officer

Scott Spaulding
Program Officer

Pamela Gamble
Senior Program Assistant

Wendy White
Director

Lois Peterson
Assistant Director

ACKNOWLEDGMENTS

The U.S. National Committee for the International Union of Biochemistry and Molecular Biology (USNC/IUBMB) and the Board on Life Sciences of the National Research Council (NRC) are grateful to the many individuals whose efforts made possible the symposium and the report. The symposium was supported by grants from the National Institute of General Medical Sciences/National Institutes of Health and the National Science Foundation, as well as by generous support from industry: ActiviX Biosciences, Applied Biosystems, Aventis Pharmaceuticals, DuPont Pharmaceuticals, Genentech, GeneProt, Large Scale Biology, Lynx Therapeutics, Micromass UK, MitoKor, Oxford GlycoSciences, Pfizer, Phylos, Prolinx, Proteome Systems., Structural Bioinformatics, and Structural GenomiX. The American Chemical Society also helped to support the symposium. People who were especially helpful in preparing for this meeting included Pamela Gamble, Laura Sheahan, Scott Spaulding, Lois Peterson, and Wendy White, all of the Board of International Scientific Organizations (BISO) of the NRC.

George Kenyon, chair of both the U.S. National Committee for the International Union of Biochemistry and Molecular Biology (USN/IUBMB) and the ad hoc symposium steering committee, conceived of the workshop in an effort to encourage discussion about the future of proteomics, similar to the preliminary meetings held for the Human Genome Project. The workshop was organized by an ad hoc steering committee derived from two committees of the National Research Council (NRC): the USNC/IUBMB and the Board on Life Sciences. The steering committee selected the speakers and some of its members wrote the workshop report. The committee members are George Kenyon (Chair), David DeMarini, Elaine Fuchs, David Galas, Jack Kirsch, Walter Moos, Gregory Petsko, Dagmar Ringe, and Gerald Rubin. Breakout session chair, Tom Leyh, Albert Einstein College of Medicine, was the only report author not part of the original steering committee.

This report has been reviewed in draft form by individuals chosen for their diverse perspectives and technical expertise, in accordance with procedures approved by the NRC's Report Review Committee. The purpose of this independent review is to provide candid and critical comments that will assist the institution in making its published report as sound as possible and to ensure that the report meets institutional standards for objectivity and evidence. The review comments and draft manuscript remain confidential to protect the integrity of the deliberative process.

We wish to thank the following individuals for their review of this report: Cheryl Arrowsmith, University of Toronto; Patricia Babbitt, University of California, San Francisco; David Baker, University of Washington; Samir Hanash, University of Michigan; John Quakenbush, Institute for Genomic Research, and Russell Thomas, Kalypsys, Inc. Although the reviewers listed above have provided many constructive comments and suggestions, they were not asked to endorse the content of the report, nor did they see the final draft before its release. The review of this report was overseen by Cynthia Beall, Case Western Reserve University, who was responsible for making certain that an independent examination of this report was carried out in accordance with institutional procedures and that all review comments were carefully considered. Responsibility for the final content of this report rests entirely with the authoring committee and the institution.

CONTENTS

BOXES

FIGURES

APPENDIXES

SUMMARY

Research in proteomics is the next logical step after genomics in understanding life processes at the molecular level. In the largest sense proteomics encompasses knowledge of the structure, function and expression of all proteins in the biochemical or biological contexts of all organisms. Since that is an impossible goal to achieve, at least in our lifetimes, it is appropriate to set more realistic, achievable goals for the field. Up to now, primarily for reasons of feasibility, scientists have tended to concentrate on accumulating information about the nature of proteins and their absolute and relative levels of expression in cells (the primary tools for this have been 2D gel electrophoresis and mass spectrometry). Although these data have been useful and will continue to be so, the information inherent in the broader definition of proteomics must also be obtained if the true promise of the growing field is to be realized. Acquiring this knowledge is the challenge for researchers in proteomics and the means to support these endeavors need to be provided. An attempt has been made to present the major issues confronting the field of proteomics and two clear messages come through in this report. The first is that the mandate of proteomics is and should be much broader than is frequently recognized. The second is that proteomics is much more complicated than sequencing genomes. This will require new technologies but it is highly likely that many of these will be developed. Looking back 10 to 20 years from now, the question is: Will we have done the job wisely or wastefully?

Introduction

Due to the rising interest in proteomics research worldwide, a symposium entitled “Defining the Mandate of Proteomics in the Post-Genomics Era” was held at the National Academy of Sciences on February 25, 2002, in Washington, D.C. Most of the attendees were invited because of their strong interest in proteomics, proteins, or drug discovery. They came from industry, both large and small, academia, and government. Most were from the United States, but an effort was made to invite people from outside the United States. Four of the 10 speakers came from outside of the United States. Six young scientists from around the world received travel fellowships to attend the meeting. The attendees heard about recent advances in the field that will greatly accelerate the process of accumulating and interpreting much of this additional needed data and information.

The planning committee selected speakers (see Box 1) and designed the symposium in the hope that one of the outcomes of the meeting would be helping to set the field on as wise a path as possible for the future. After the presentations attendees were involved in individual breakout sessions on a variety of topics, including

- protein separation and identification
- protein structure and function
- metabolic pathways and post-translational modifications
- implementation: necessary policy and infrastructure conditions for collaboration
- platforms: emerging technologies
- computational methods and bioinformatics
- clinical proteomics

The thoughts and ideas of the speakers and those expressed in the breakout sessions were captured by recorders to assist in the preparation of this report. While other organizations and meetings have addressed many of the issues facing proteomics, we hope that participants and readers of this report will look back on this meeting as the field progresses and find that it was of some help in defining the current efforts and applications, as well as providing direction to the advancing state of the art.

BOX 1

Symposium Speakers and Affiliations

- Ruedi Aebersold, Institute for Systems Biology, Seattle, Washington
- Cheryl Arrowsmith, University of Toronto, Canada
- Marvin Cassman, NIGMS, Bethesda, Maryland
- Julio Celis, Institute of Cancer Biology and Danish Center for Human Genome Research, Copenhagen, Denmark
- Brian Chait, Rockefeller University, New York, New York
- Francis Collins, National Human Genome Research Institute, Bethesda, Maryland
- Denis Hochstrasser, University of Geneva, Geneva University Hospital, Switzerland
- Joshua LaBaer, Harvard Medical School, Boston, Massachusetts
- Scott Patterson, Celera Genomics Corporation, Rockville, Maryland
- John E. Walker, Medical Research Council, Cambridge, United Kingdom

Proteomics

Now that the DNA sequences of the human genome and genomes of dozens of other organisms are essentially known, the biomedical and biological communities are placing increased emphasis on proteomics, the study of the proteins that are the gene products. Proteomics, a word derived from "protein" and "genomics," needs further definition, as do proteomics initiatives, especially since many in the scientific community are asking for a human proteome project.

Historically one can point back to meetings and articles over 20 years ago, when scientists began to think about mapping the entire set of human proteins (see, for example, B.F.C. Clark, "Towards a Total Human Protein Map[1]"). Indeed, Congress was considering a project called the "Human Protein Index" long before the Human Genome Project had been conceived. The Human Protein Index project was developed in the late 1970's by Norman G. Anderson and N. Leigh Anderson at the Department of Energy's Argonne National Laboratory[2]. Its objective was to enumerate the human proteins (what would now be called the human proteome) by separation on 2D gels and thus define their genes from the protein end, the only approach possible in those days before large scale DNA sequencing was possible. But this effort was perhaps ahead of its

[1] Clark, B.F.C (1981) Towards a Total Human Protein Map. *Nature* **292** (5823): 491-492
[2] Anderson, N.G. and Anderson, N.L. (1979) *Behring Inst Mitt.* **63**: 169-210

time given the lack of suitable technologies and shifting political sands. Instead, the rise of genomics took center stage. An Australian postdoctoral student, Marc Wilkins, is often credited with coining the term "proteomics" in 1994[3] at a time when only one proteomics company existed (Large Scale Biology Corporation).

Today many proteomics initiatives are underway in industry and otherwise, such as the Human Proteomics Initiative (HPI), an effort which began in 2000 by the Swiss Institute of Bioinformatics and the European Bioinformatics Institute. The goal of the HPI is to annotate each known protein, providing information that includes the description of protein function, domain structure, subcellular location, post-translational modifications, splice variants, and similarities to other mammalian proteins[4]. Another major proteomics effort is led by the Human Proteome Organization (HUPO), a group which has created a worldwide organization that engages in scientific and educational activities to encourage the spread of proteomics technologies and to disseminate knowledge pertaining to the human proteome and that of model organisms[5].

On which goals should these national and international efforts focus? Should they be limited to human proteomics or like the Human Genome Project, include key model organisms? Perhaps the proteomes of the human pathogens should be included as well (e.g., the malaria parasite and other infectious microorganisms), and if so, in what order of priority? Should development of more efficient instrumentation (e.g., mass spectrometers, X-ray diffractometers, nuclear magnetic resonance spectrometers) and improved computational methodologies (e.g., high-speed computers and software useful in bioinformatics) be emphasized? What should be the role of major federal funding agencies (e.g., the National Institutes of Health, the National Science Foundation, the U.S. Environmental Protection Agency, and the U.S. Department of Agriculture)? What should be the role of academic laboratories? Should projects be supported mostly by individual research grants or program project (group effort) grants? What should be the role of the private sector, particularly those companies large and small that have a major stake in exploiting the results of the various genome projects and proteomics initiatives? How can all of these stakeholders cooperate most effectively while still maintaining proprietary information where appropriate? Should the overall goal be to understand the structure and function of all known proteins or should only those known to be involved in diseases be emphasized? After all, one must first understand function if one is to fully understand dysfunction. Is enough emphasis being given to the functional aspects of proteomics? Are studies on post-translational modifications of proteins and subsequent functional aspects included in "proteomics"? Hence the interest in organizing the one-day symposium reported herein.

---

[3] < http://www.signalsmag.com/, November 2, 1999>
[4] <http://us.expasy.org/sprot/hpi/>
[5] <http://www.hupo.org/>

Discussion of General Topics Covered at Symposium

Beginning with a definition of the term "proteomics," Marvin Cassman, former director of the National Institute of General Medical Sciences, and now at University of California, San Francisco and the Institute for Quantitative Biomedical Research, was one of many speakers expressing an opinion on this subject and it was clear that proteomics means many (or at least different) things to different people. Some definitions include "high-throughput" and some do not. Obviously proteomics is not merely protein chemistry. Symposium chair and Dean of the University of Michigan College of Pharmacy, George Kenyon, commented, "Proteomics is not just a mass spectrum of a spot on a gel." Perhaps the most useful definition of proteomics for our purposes is the broadest: Proteomics represents the effort to establish the identities, quantities, structures, and biochemical and cellular functions of all proteins in an organism, organ, or organelle, and how these properties vary in space, time, or physiological state.

Somewhat limited operational definitions of proteomics were offered by some of the speakers. For instance, "In one sense it makes no difference at all why should you call something proteomics or call it something else?" Dr. Cassman continued, "What we call things often conditions how we organize our thinking and our efforts." He explained that genome-driven target selection coupled to high-throughput technologies is what he believes structural genomics means. "It means you are using the genomes as the primary source for target selection." However, structural proteomics uses these features "plus the additive feature of full coverage of protein space, that is, completeness" stated Dr. Cassman. The goal of completeness does not intend to suggest, however, that any smaller scale experiments, even including high- throughput analysis of specific tissues or subsets of proteins, would not be considered to be part of proteomics.

Of course there are many "-omics" along with proteomics including genomics, metabolomics, transcriptomics, interactomics and so on, which are collectively involved in the mandate of defining proteomics. However, we will restrain ourselves from commenting on other "-omics". Functional genomics and functional proteomics (which can encompass other 'omics' as mentioned) are closely juxtaposed on a continuum along the path of discovering the detailed secrets of life and life processes.

The general topics covered at the symposium included:

- Perspectives (including genomics perspective; relationship of proteome to genome)
- Source of proteins (including organism, sample storage)
- Protein separation (including purification if subcellular)
- Protein identification (largely mass spectometry)
- Protein function (including localization, protein:protein interactions, structure determination, structure-function, post-translational modifications)
- Applications (including drug discovery, diagnostics)
- Informatics (including homology modeling, databases, analysis software, standardization)
- Other topics (including international collaboration, ethical considerations, collaboratories[6])

[6] Collaboratories are distributed research centers in which scientists in two or more locations are able to work together with the assistance of various forms of communication and collaborative technologies.

Dr. Cassman defined proteomics as a set of related options: "the analysis of complete complements of proteins present in defined cell or tissue environments (i.e., context-dependent) and their variation in space and time" (with credit given to Stan Fields for his contributions to this definition). One example of a proteomic effort is the Protein Structure Initiative of the NIGMS, which has as a goal the generation of a complete complement of protein structures in nature through the combination of direct structure determination and homology modeling. Although it requires high-throughput technology and genomic data to use for target determination, the goal of "completeness" is what distinguishes the effort as proteomics, according to Dr. Cassman.

The second part of his definition is exemplified by the use of microarrays to identify characteristic markers for cancer progression in specific tissue samples. These studies involve image and pattern recognition tools, which yield large-scale visualization of specific cell-dependent, context-dependent proteomic outputs.

The third part of the definition involves examining proteomic outputs in time and space. This requires not only the application of bioinformatics tools but also computational biology, that is, the use of modeling and simulation. Complex systems analysis could be considered an important element in the larger picture of defining a proteome, and such analysis will require theoretical modeling of systems. Several examples of NIGMS initiatives that focus on mathematical modeling of complex biological systems were provided.

While we may be far off in terms of defining a complete human proteome, approaching proteomics on an organellar basis provides goals that are perhaps achievable in our lifetimes. Remember that the first DNA genomes sequenced were those of the bacteriophage, in the 1970s, followed in 1981 with the DNA sequencing of a human mitochondrial genome.

Consider also that the mitochondrion, which is estimated to be composed of about 2,000 proteins, presents a considerably more manageable problem and a microcosm of whole cell proteomics. With this in mind Nobel laureate Sir John Walker, head of the Dunn Medical Research Council Unit in Cambridge, UK, discussed his proteomic studies of mitochondria directed to resolving specific biological issues. Dr. Walker's work includes the definition of the protein complement assembled in the respiratory enzyme known as complex I, the identification of the biochemical functions of a family of transport proteins found only in mitochondria, and the discovery of phosphorylation-dephosphorylation pathways in mitochondria. These studies rely not only on mass spectrometric and bioinformatics tools but also on biochemistry and genetics. Such an integrated approach is proving to be quite rewarding in Dr. Walker's view, in terms of both understanding the biology of mitochondria and the technical development of new methods versus attempts to analyze the global complement of proteins in the organelle. It is also possible to focus on subcompartments of mitochondria, such as the inner mitochondrial membrane of so much interest to bioenergeticists.

In this report we have tried to avoid being constrained by a narrow definition of proteomics (e.g., merely quantitating protein levels) and have used the broad definition given earlier to allow a wide-ranging discussion of goals, techniques, opportunities, and challenges.

Lessons Learned from the Human Genome Project

Francis Collins, director of the National Human Genome Research Institute, spoke about lessons learned from the Human Genome Project that might be applicable to the discussion of a

public large-scale proteomics initiative (see Box 2). He began his presentation by taking issue with the term "post-genomics era." He queried whether this means that from the beginning of the universe until 2001 we were in the "pre-genome era," and then suddenly, "bang," we moved into the post-genome era (leading one to wonder what happened to the genome era). He suggested that it was presumptuous to say that the Human Genome Project is already behind us. He pointed out that proteomics is a subset of genomics, and genomics is more than sequencing genomes, which will be ongoing for decades to come. His comments are especially relevant given that the human genome was still only about 69 percent complete at the time of the meeting.

BOX 2
Lessons Learned from the Human Genome Project:
Comments from Francis Collins

- **High level planning process** with broad input from the scientific community is crucial to setting ambitious but achievable and realistic goals.
- **A focus on completeness** is important, even though this is extremely difficult when dealing with proteins. This is what distinguishes proteomics from the study of individual proteins, or the fields of biochemistry and physiology. Without completeness as a goal of proteomics much of the same research would be duplicated at a later time.
- **Technology must be developed and validated** before attempting to scale up. Technology development includes the range of activities from proof of principle, to pilot projects, to scaling up, to high-throughput. The Human Genome Project sequenced model organisms and generated the necessary infrastructure prior to actually sequencing the human genome, which did not start until six years into the project and was initiated first with pilot projects.
- **Public availability of data and resources** is absolutely critical if the benefits to the scientific community are going to be realized. The rapid release of pre-publication data was a key to the success of the Human Genome Project.
- **Interdisciplinary research** needs to be fostered, including the participation of experts in automation, chemistry, and bioinformatics.
- **International participation and coordination** is an essential component to bring the best minds to the problem, to avoid duplication, and for cost sharing.
- **Centralized databases** that allow for integration and visualization of the data are an essential resource and are needed to transfer all these data into the hands of those who want to use them. They are expensive and need to be nurtured.
- **Public-private partnerships** should be sought whenever feasible, especially for the generation of pre-competitive data sets. (Successful examples include the single nucleotide polymorphism consortium and mouse genomic sequencing.) Characteristics for successful public-private partnerships include a compelling scientific opportunity, pre-competitive data sets, simultaneous availability of data to all users, production facilities already in place, firm milestones and deliverables, affordability, and having well-defined endpoints.

Dr. Collins concurred with other participants in delivering the sobering message that a large-scale proteomics effort is orders of magnitude more complicated and difficult than the sequencing of the human genome. (As if 100 trillion cells making up an organism and billions of base pairs in genomes are not enough complexity already!) The concept of a complete dataset of

all human proteins is therefore very difficult to imagine. There are many challenges as stated below:

- Wide dynamic range of expression
- Protein modifications
- Physical handling of proteins is more difficult than working with nucleic acids
- Need for multiple technologies, many of which are not optimized or even invented
- Unlike DNA data, protein data are more analog than digital, making data integration and analysis very challenging
- Intellectual property rights and claims

Dr. Collins said that the most important area for investment in proteomics right now is technology development so that we can move these methods in the direction of being able to tackle a mammalian proteome without facing enormous costs and problems with quality of the data.

A number of resources for genomics research continue to be generated that may help inform a proteomics effort, including multiple coverage of certain genomes and more specifically:

- Multiple genomic sequences from mouse (6x coverage), rat (3x coverage), puffer fish, zebrafish, a sea squirt, and close relatives of *C. elegans* (10x coverage) and *D. melanogaster* will be forthcoming. Comparative genomics will be helpful in understanding gene models and gene function.
- Full-length human cDNA sequencing efforts are ongoing in Germany and Japan.
- Full-length cDNAs for human and mouse are being generated through the National Institutes of Health (NIH) Mammalian Gene Collection[7]. Multiple NIH institutes plan to support a central database of protein sequence and function through a new initiative[8].

Dr. Collins referred to one publication: "Global Analysis of Protein Activities Using Proteome Chips."[9] He finished his presentation with a particular recommendation, not from a scientist but from a famous athlete (hockey star Wayne Gretzky). When asked how it occurred that he was so good at playing hockey, and why it was that he always seemed to score the key goals, Gretzky said, "It is very simple. You have got to skate where the puck is going to be." In the field of proteomics Dr. Collins said he was not sure where exactly the puck was going to be, but there were a lot of "Wayne Gretzky's" at the meeting, and Dr. Collins was glad to get a chance to listen to them.

---

[7] <http://mgc.nci.nih.gov>
[8] <http://grants.nih.gov/grants/guide/rfa-files/RFA-HG-02-001.html>
[9] Zhu, H., et al. 2001. *Science* 293 (5537): 2101-2105

Sources of Proteins

By definition any proteomics effort aims at ‘completeness’ of information. This part of the symposium addressed primarily the comprehensiveness or completeness of any assembled library of proteins and the quality of the materials. It was noted that protein expression in a given cell varies from none to abundant. Historically, for practical reasons, the abundant proteins have been investigated most extensively; however, some of the rarely expressed proteins and proteins that appear only in disease states may be among the more interesting. Joshua LaBaer, Harvard Medical School, noted that the function of all proteins can be studied regardless of in vivo levels once a copy of the gene and adequate expression vectors are available. Ideally it would be desirable to have an available repository or library containing one clone for every spliced variant in the proteome. The size of that library will not be known for some time, but an intermediate realizable objective would be a repository consisting of one clone for every gene. These clones should be “expression ready”; that is, they should contain only the cDNA from the initiation site to the stop codons. It seems likely that we should have "some idea of all the different cDNAs" in the genome in the near future, stated Dr. LaBaer. The expressed proteins could be studied functionally and often identified by mass spectrometry. In general it is fairly easy to produce large quantities of proteins in insect cells or bacteria, but in certain cases it may be necessary to express them in their native cells in order to address such problems as localization or post-translational modifications. Dr. LaBaer compared the complexities of studying mammalian systems with those in yeast. There are approximately 6,000 genes in yeast compared to a much larger number in humans. Moreover, the genome in yeast is relatively simple; for example, there are only about 220 intron-containing genes in yeast, whereas a much larger fraction of mammalian genes contain introns and alternative splicing substantially increases the number of expressed proteins.

To this end Dr. LaBaer described the FLEX Gene repository, which is currently being assembled by a consortium of about 20 different public and private research laboratories. "FLEX" stands for Full Length Expression ready. This repository will enable scientists to move several genes simultaneously from the master vector to any expression vector, which will allow researchers to screen for function by high-throughput experimentation. It is the intention of this consortium to make this collection of all human genes broadly available without restrictions on their use. The four self-defined objectives of the consortium are (1) identification of the genes, (2) assembly of clones, (3) sequence validation, and (4) distribution to the scientific community. One example of the success of this effort resulted in the identification of two new genes that are likely involved in the migration of breast cancer cells through a membrane. The collaboration of public and private research groups raises certain legal issues, which include consideration of antitrust law.

Recombination-based cloning was presented as a high-throughput technology to enable the ready transfer of cDNAs from the supplied vector to one's own preferred expression vector. Dr. LaBaer described a protein purification scheme that was developed by a graduate student in his laboratory, Pascal Braun. “In the case of human proteins,” Dr. LaBaer explained, “where it is not easy to produce these proteins in human cells, [the availability of large numbers of purified proteins] will require the use of heterologous [expression] systems such as bacteria.” “To develop these methods,” continued Dr. LaBaer, “Braun transferred a collection of 30 cancer genes into four different expression vectors, each one adding a different epitope tag. [Braun]

then developed a two-hour automated protocol for purifying 96 proteins in parallel [and] has now purified over 330 different proteins using this approach." Braun and Yanhui Hu of the lab created a database that correlates the success of purification with various features of the proteins such as pI, GO annotation, subcellular localization, and domain structure. Dr. LaBaer said they found that the presence of certain domains such as SH2 domains or SH3 domains can predict success in purification.

Dr. LaBaer concluded with a description of a database derived from a computer program that searches the primary literature for abstracts that mention both a gene and a disease. The assumption is that a significant number of such occurrences may identify groups of genes associated with a given disease. This effort was presented as a task in progress, and interested scientists were invited to experiment with the database.[10]

Brian T. Chait from Rockefeller University described a proteomics approach to understanding cellular function. His group is interested in mechanisms by which materials enter and exit the nucleus, the isolation of multiprotein complexes and the determination of their cellular localization. The basic concept is to introduce a particular affinity tag to one of the proteins at its natural location in the chromosome, which is done by replacing the endogenous gene by a gene that will code for a protein with a tag on it or as he termed it, "a piece of molecular Velcro." So long as the multiprotein complex is stable, the tag allows isolation of the associated interacting proteins. An application to the nuclear pore complex, a group of proteins involved in nuclear trafficking, was described extensively. The complex as isolated has a molecular mass of 50 million daltons. Interestingly, in the initial purification experiments it contained about 180 interacting proteins, but upon further fractionation only around 50 were found to comprise the complex. The individual proteins are identified by mass spectrometry, which has the power to provide additional information about phosphorylation sites.

Preliminary experiments describing the use of this approach to follow proteins at different points in the cell cycle and in the regulation of chromatin were mentioned briefly. The genomic tagging and mapping approach can be used to gain analogous information about a number of other systems. Most importantly this approach can show where the protein is localized within the cell, how much is present, when the protein is present and for how long, with what it is interacting, and even something about the topology of the protein complexes.

Protein Separation

After more than a decade of effort in gene sequencing, reliable estimates of the number of human genes is still a matter of disagreement, speculation, and debate. From the point of view of proteomics, just the detection or enumeration of the numbers of expressed proteins defies prediction based on our current understanding of human cell-type protein composition and its modulation by myriad undefined post-translational modifications. Their actual identification or annotation of function remains a challenge. This entire situation is not significantly better for

[10] <http://hipseq.med.harvard.edu/MEDGENE/>

yeast. It is thus not surprising that a key problem in proteomics at a practical level is the simplification of protein mixtures to a state in which their characterization by physicochemical methods is experimentally tractable. There are no documented, reliable, or reproducible strategies for separation of classes of proteins or even individual proteins from very complex mixtures typically obtained in biological samples such as cell lysates. Clearly, not only does one wish to know which specific proteins are in a given sample but, ideally, one would wish to know whether specific proteins are part of a particular biologically significant compartment, complex, or subcomplex.

Denis Hochstrasser from the University of Geneva, a founder of GeneProt Inc., GeneBio SA, the Swiss Institute of Bioinformatics, and one of the pioneers in the identification of proteins in 2D gels, took the lead in dealing with the topic of protein separation. He stated at the outset that he wanted to play the role of "devil's advocate": to describe some of the excitement in proteomics but also to describe some of the difficulties. He outlined the scale of potential proteins one can look for in the millimolar (10^{-3}), micromolar (10^{-6}), nanomolar (10^{-9}), picomolar (10^{-12}), femtomolar (10^{-15}), attomolar (10^{-18}), zeptomolar (10^{-21}) and yoctomolar (10^{-24}) (which is less than one molecule per liter) ranges. When one considers human blood, for example, Hochstrasser noted, "typically you only see albumin, immunoglobulin, and transferrin," whereas cardiac markers such as troponin are present at nanomolar concentrations, and insulin-like growth factor or insulin are in the picomolar range. Parathyroid hormone is in the low picomolar range and Tumor Necrosis Factor is found in the femtomolar range (see Figure 1).

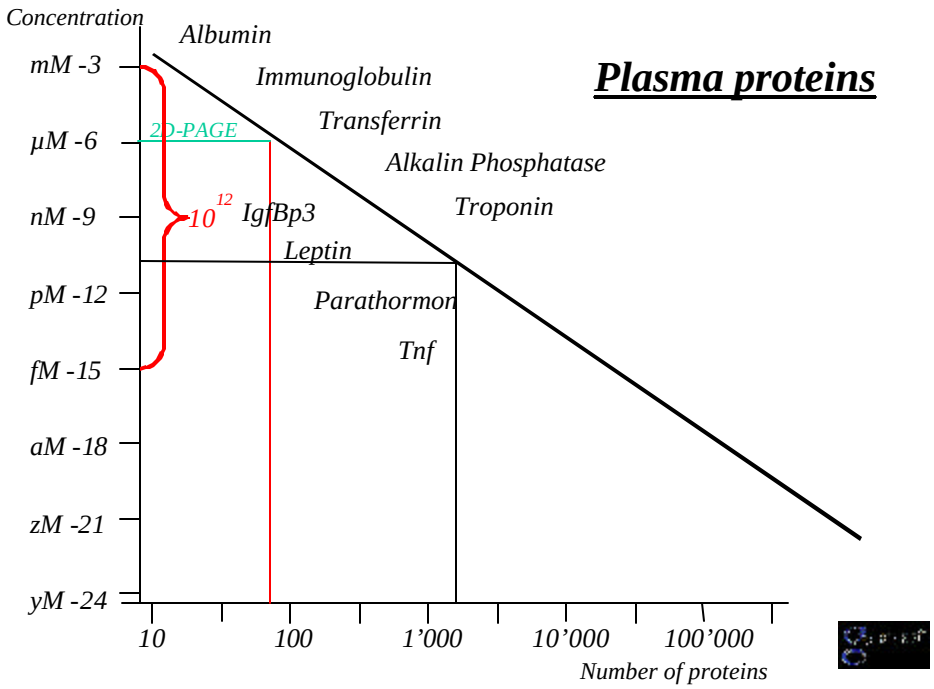

FIGURE 1. **Potential plasma proteins observable at various concentration ranges:** millimolar (10^{-3}), micromolar (10^{-6}), nanomolar (10^{-9}), picomolar (10^{-12}), femtomolar (10^{-15}), attomolar (10^{-18}), zeptomolar (10^{-21}) and yoctomolar (10^{-24}). *SOURCE*: Courtesy of Denis Hochstrasser, GeneProt Inc.

Hochstrasser speculated that there is "a linear logarithmic relationship between the concentration in blood and the number of proteins." He suggested that if there are about 300,000 proteins in the human body or five to six times the number of genes "you probably could find any protein you have in the body, maybe one in the total blood volume, which would be just below Avogadro's number (1 protein/L of plasma), because we have 6 or 7 liters of blood which makes about 4 liters of plasma, and if you have one in 4 liters, it is about at the yoctomolar (10^{-24}M) level."

For experimental studies the amount of starting material, such as blood, is considerable in order to have high enough levels of various protein material that can be detected by today's methods. Since a 2D gel has a dynamic range of only 10^4, Hochstrasser stated, "if anyone used [a] 2D gel from crude plasma, you never go below the micromolar range." Hochstrasser noted, for example, that starting with 1 mL of sample leads to roughly a nanomolar limit of detection. He further explained that starting with a much larger volume (e.g., 5-10 liters of plasma) is necessary to achieve detectability in the lower picomolar range. Clearly, prefractionation of proteins, individually, or as a subgroup is essential to reach the dynamic range of detectability required for both cell and tissue lysates, and plasma.

In subsequent discussion it became clear that even the best large-format 2D gels are inadequate for studies of the global range of expression, perhaps still inadequate by a factor of 10; therefore at least a 10-fold fractionation prior to large-format 2D gel separation would be required. Unfortunately, many membrane proteins do not enter 2D gels effectively. This presents a formidable challenge for the field.

In his presentation, Julio Celis from the Institute of Cancer Biology and the Danish Center for Human Genome Research in Aarhus, Denmark, also spoke about methods and challenges in the area of protein separation. He stated that "for the study of tissue biopsies the use of high-resolution 2D electrophoresis is the method of choice [for separations] as non-gel high-throughput technologies based on chromatography-mass spectrometry are not yet ready for the study of tissue samples." He stated that 2D gel technology in combination with mass spectrometry can be used to establish comprehensive databases of protein information that can be useful in the clinical setting. He also made the important point that data in a given cell type can be valuable to the study of other cell types since 80-90 percent of the proteins are believed to be shared by all cell types. While many structural and metabolic gene products may be the same between all cells, as one reviewer pointed out, cell-specific proteins will be important for understanding function and disease.

An afternoon breakout session, devoted to the topic of "protein separation and identification," was led by Julio Celis; Alain Van Dorsselear, Louis Pasteur University, CNRS; and A. L. Burlingame of the University of California, San Francisco. Most of the 16 discussants were experts in mass spectrometry. The discussants concluded that the issue of sample preparation and purification has been sadly neglected at most meetings dealing with proteomics. There was the impression among some of the discussants that protein biochemists were developing and using methods to purify proteins that were not being adequately defined compositionally by mass spectrometrists interested in proteins. They envisioned setting up "core centers of excellence" in proteomics where innovation, mobility of people and ideas, and training can all occur. These core centers might also lead to spin-offs for the development of new instrumentation. Resources required to support a broad proteomic effort could be in the form of sample collections, standardization of data across platforms, and ligands that allow assaying of individual proteins, to name just a few. These centers would complement the work of scientists

in individual, relatively small laboratories where more open-ended, curiosity-driven research can occur. Even when the advent of better strategies for protein mixture fractionation are in hand, new developments in mass spectrometry are needed to extend the dynamic range of detectability of protein samples, especially for proteins that are post-translationally modified.

Protein Identification

Until we can identify each expressed protein in a cell or target tissue we cannot fully define the proteome. Current, best practices have practical lower limits of protein-detection in the nanomolar or picomolar range, which is 10 to 15 orders of magnitude less sensitive than what is needed for complete proteome definition, which it was generally agreed, would require almost zeptomole or yoctomole sensitivity. The need for developing better and more sensitive methods of detection is pressing, and there are many opportunities to make significant technical advances. Insolubility is an important issue, membrane-associated proteins (which may comprise as much as 30 percent of the proteome) remain largely inaccessible experimentally, and the lack of sensitivity in protein detection and identification remains among our greatest limitations. Each order of magnitude increase in sensitivity brings important new insights into proteome composition and behavior.

Denis Hochstrasser brought into sharp focus the disparity between the sensitivity of current protein-detection methods and the proteomics community's expectations regarding the sensitivity required to identify the complete proteome of a target cell or tissue. Cell and receptor based assay systems can detect peptides in the femtomolar range. These methods define the lower limit of our detection ability but, unfortunately, are applicable only to small sets of proteins. 2D gel electrophoresis, still an experimental mainstay in the proteomics community, can detect protein concentrations as low as micromolar, a sensitivity sufficient to identify ~ 100 plasma proteins, not including modified forms. Under ideal conditions, including the sieving out of abundant proteins, mass spectrometry can extend sensitivity three orders of magnitude to the nanomolar level. Mass-spectral proteome screening is being carried out on an industrial scale by GeneProt Inc., one of the world's first large-scale proteomic R&D centers. The facility houses 40 Tandem Mass (MS/MS) spectrometers, each serving two High Performance Liquid Chromatography (HPLC) machines. With each spectrometer running two samples per hour, the facility is capable of performing 1,920 MS/MS-characterized HPLC profiles per day, remarkably with very little human intervention. The sensitivity can be extended to the picomolar range by preparing single, mass-spectrometry samples from 10-15 liters of the target.

Using a strategy that circumvents the need to detect a protein or to know its molecular function, Dr. Hochstrasser and colleagues are synthesizing proteins as large as 25 kDa in sufficient quantity and purity to immediately search for the effects of overexpression after injection in living systems, allowing them to move quickly from interesting protein candidates identified using informatics screens, to cellular or organism physiological response.

Dr. Hochstrasser underscored the well-known fact that mass-spectral identification is in general far more successful for peptides than proteins. He outlined a technological innovation that integrates the protein-separating resolution of 2D gels with the sensitivity of peptide-mass spectrometry. The method sandwiches a protease-impregnated membrane between a delivery membrane (that carries protein previously transferred from a 2D gel) and a capture membrane.

The proteins are cleaved into peptides during electrophoretic transfer to the capture membrane, and are then desorbed from small registered sections of the membrane, using a laser, and finally delivered directly into the mass spectrometer. Dr. Hochstrasser called this new technology "The Molecular Scanner" (see Figure 2).

In the "Emerging Technologies" breakout session, co-chaired by Ruth Van Bogelen, Pfizer Global Research, and Norman G. Anderson, Large Scale Biology Corporation, the need for specific new technologies was discussed at length by experts in the field. Dr. Van Bogelen commented that "detection and realizing that the dynamic range of proteins in cells is probably over six, seven, maybe even twelve orders of magnitude where we really want to be able to detect proteins that are present, at even less than one molecule per cell. We think those [proteins] are important, and we don't have the capabilities to do that."

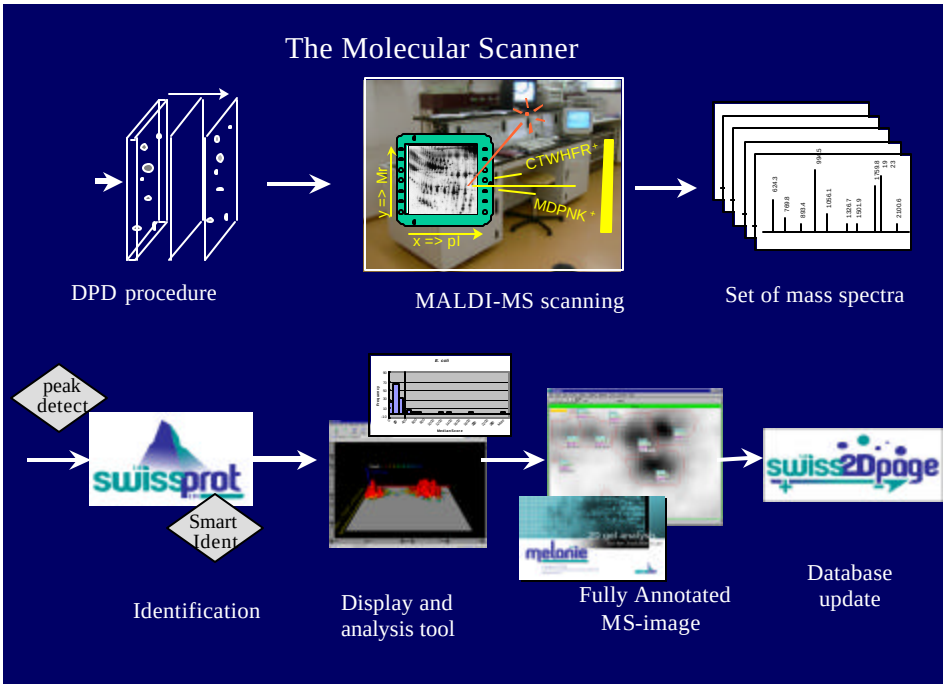

FIGURE 2. **The molecular scanner integrates the protein separating resolution of 2D gels with the sensitivity of peptide-mass spectrometry.** The proteins are cleaved into peptides and then desorbed using a laser and finally delivered directly into the mass spectrometer. *SOURCE:* Courtesy of Denis Hochstrasser, GeneProt, Inc.

The consensus of the group was that many areas are in need of development if the goal of defining the composition and behavior of proteomes is to become a reality. The perceived needs ranged from technologies that can determine the organization of the cellular matrix and the functions of the proteins in it, to single-cell proteomics, and the comprehensive analysis of post-translational modifications. Important practical issues were raised, like the need to standardize data across different technology platforms and how to organize the enormous volume of information being created daily. "The bottom line is, there is just a lot of work to be done," said Dr. Van Bogelen. "We need money invested into developing technologies. And we really need to have students in this area who are moving this field into the next generation."

Data Collection

"An essential element of proteomics is the intent to collect data on proteins systematically and, where applicable, quantitatively," said Ruedi Aebersold, co-founder of the Institute for Systems Biology in Seattle, Washington. "Systematic data collection," said Dr. Aebersold, "means that the measurements are made on all the proteins present in a sample, eventually all the proteins that constitute a proteome." It is expected that proteomic data will be useful for classification of cells and tissues in health and disease and, more ambitiously, for achieving a detailed understanding of biological mechanisms. Dr. Aebersold discussed the development of an automated quantitative approach by his laboratory to help achieve their goals.

The technologies to perform various types of proteomic measurement are not mature and thus are limited in capacity (see Figure 3).

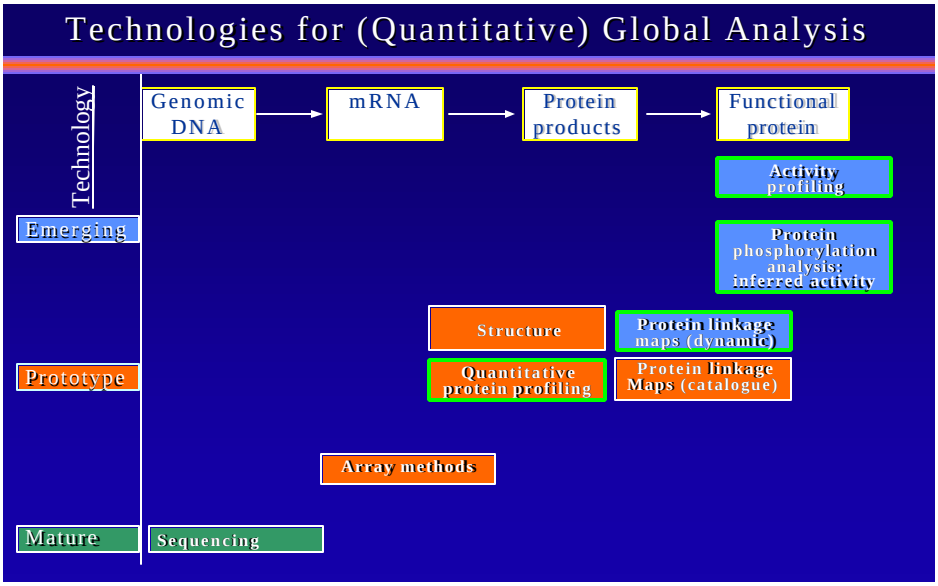

FIGURE 3. **Technologies for (quantitative) global analysis.** The technologies to perform different proteomic measurements have reached various degrees of maturity and none of them is a fully mature technology. It is unlikely that a single experimental platform will be able to collect all types of proteomic data. *SOURCE:* Courtesy of Ruedi Aebersold, Institute for Systems Biology, Seattle, Washington.

Dr. Aebersold's group has developed a general approach to quantitative proteomics based on automated tandem mass spectrometry, stable isotope dilution theory, and a suite of bioinformatics tools for data analysis.[11] Dr. Aebersold described the approach as follows: "Stable isotope signatures are introduced into proteins at specific sites by means of chemical reactions. Later these signatures are deconvoluted by a mass spectrometer and serve as the basis for accurate quantification of each labeled protein. The objective of the initial implementation of this technology has been quantitative protein profiling. The method is based on a class of reagents called 'isotope coded affinity tags' (ICAT reagents) (see Figure 4) and the method is schematically illustrated (see Figure 5.) By changing the specificity of the reagent, the approach becomes generic for different quantitative proteomic measurements. Work is underway to extend

[11] Gygi, S.P., Rist, B, Gerber, S.A., Turecek, F., Gelb, M.H., Aebersold, R. (1999). *Nature Biotechnology* **17**, 994-9.

this approach to determine profiles of enzyme activities (an area pioneered by Ben Cravatt from the Scripps Research Institute), and to protein linkage analysis, and protein phosphorylation profiles."

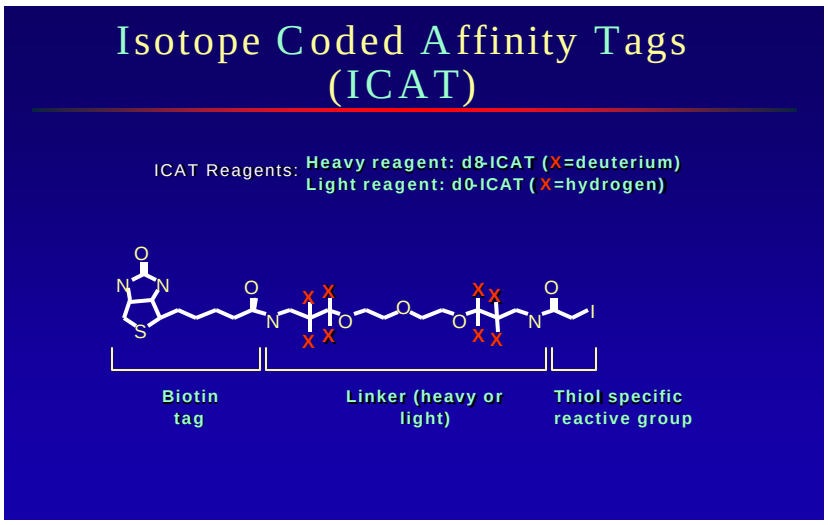

FIGURE 4. **Isotope-coded affinity tags (ICAT).** A class of reagents called "isotope coded affinity tags" (ICAT reagents) are used to perform quantitative proteomic analyses based upon automated tandem mass spectrometry, stable isotope dilution theory, and a suite of bioinformatic tools for data analysis. Stable isotope signatures are introduced into proteins at specific sites via chemical reactions. These signatures are later deconvoluted by a mass spectrometer and serve as the basis for accurate quantification of each labeled protein. The objective of the initial implementation of this technology has been quantitative protein profiling. The method is schematically illustrated in Figure 5. *SOURCE:* Courtesy of Ruedi Aebersold, Institute for Systems Biology, Seattle, Washington.

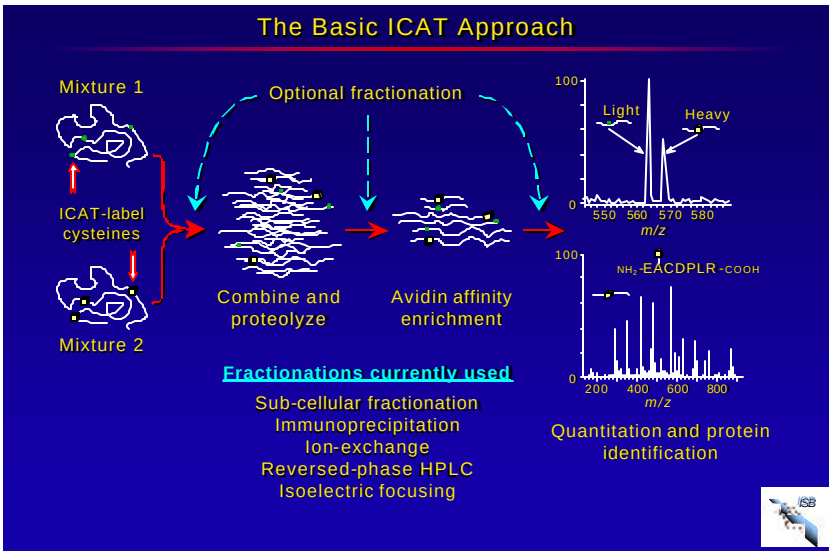

FIGURE 5. **The Basic ICAT Approach.** This figure schematically represents the ICAT approach to quantitative proteomics. By changing the specificity of the reagent, the approach becomes generic for

different quantitative proteomic measurements. *SOURCE:* Courtesy of Ruedi Aebersold, Institute for Systems Biology, Seattle, Washington.

Dr. Aebersold's group has also developed a suite of software tools that use statistical methods to identify and eliminate poor quality spectra often observed during validation of automated liquid chromatography-MS/MS experiments. The software assigns a numerical value to each database search result, which indicates the probability that the search result is correct. Dr. Aebersold believes that such tools save considerable amounts of time and that they are essential for the adoption of community-accepted standards for protein identification by mass spectrometry.

In collaboration with Sciex (a manufacturer of mass spectrometers) Dr. Aebersold's group has developed a mass spectrometry system that he refers to as "smart data acquisition." The system is based on a matrix-assisted laser desorption ionization (MALDI) quadrupole time-of flight mass spectrometer (QSTAR, Sciex) and is illustrated (see Figure 6.) This system allows one to quantify all the detected peptides first and then to selectively sequence only those that show an interesting quantitative change.[12] "By focusing the sequencing efforts on those peptides that show a change in quantity, the analysis is focused on those peptides that are relevant to the question asked, and the number of required sequencing operations is reduced by approximately an order of magnitude," stated Dr. Aebersold.

The systematic and quantitative analysis of the properties that define protein activity and function within a defined context (i.e., proteomics) is essential for biology and medicine. "It appears unlikely that a single experimental platform will soon emerge that can collect all the different types of relevant data," stated Dr. Aebersold, "however, improved bioinformatics tools and smart data collection (either by themselves or in combination) have the potential to significantly increase the sample throughput in proteomics."

[12] Griffin, T.J., Gygi, S.P. Rist, B., Aebersold, R., Loboda, A., Jilkine, A., Ens, W., Standing, K.G. (2001. *Anal Chem*: 73(5): 978-86.

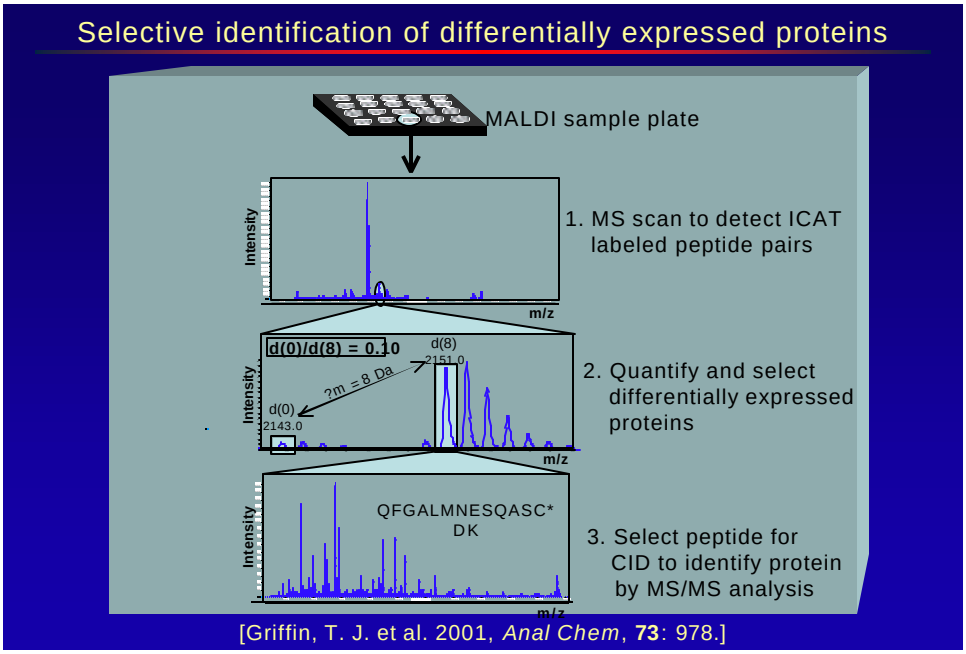

FIGURE 6. **Selective identification of differentially expressed proteins.** In collaboration with Sciex (a manufacturer of mass spectrometers), Ruedi Aebersold's group has developed a mass spectrometry system that allows one to quantify all the detected peptides first and then to selectively sequence only those that show an interesting quantitative behavior. The system is based on a matrix-assisted laser desorption ionization (MALDI) quadrupole time-of-flight mass spectrometer (QSTAR, Sciex) and is schematically illustrated in this figure. *SOURCE:* Courtesy of Ruedi Aebersold, Institute for Systems Biology, Seattle, Washington.

Proteomics and the Problem of Function

Proteomics represents an exploration of a great unknown. Some 40 percent or more of the sequences in the genomic databases represent open reading frames that code for proteins for which there is no assigned function or for which the annotated function is incomplete or incorrect. This presents an enormous challenge to the biological community.

The function of a protein can be defined in many different ways depending on the experiments being done and the questions being asked. It may be useful to preface the word "function" with an adjective that specifies the nature of the effect that the protein produces. "Chemical function" refers to the general type of reaction catalyzed in the case of an enzyme; for non-enzymes this term is not applicable. "Biochemical function" refers to the specific substrates used, the products produced and the mechanism of the transformation between them in the case of an enzyme; the specific molecules bound, and the response produced in the case of a receptor, scaffold, regulatory protein or channel, and so forth. "Cellular function" refers to the pathway(s) in which the protein operates. These pathways are created by the combined biochemical functions of the proteins involved. Note that there exists a hierarchy of functions in which a complete understanding of function at each level depends on the information from the previous levels. The hierarchy continues with functions defined at the level of a phenotype of, say, a

knockout: this may be manifest in the effect on a single cell or on an organelle or entire organism. Finally, it is possible to define function at the level of the effect of the loss or mutation of that protein on the development of a higher organism from embryo through to adult.

Function is not a fixed property for many if not most proteins. There are many ways that gene products can be altered to elicit modified or completely new functions. For example, there exist

- alternative splicing - which may affect as many as 1/4 or more of the genes in a higher eukaryote and can alter biochemical function either drastically or subtly, producing truncated proteins and proteins with different compositions
- post-translational modification, such as phosphorylation and glycosidation (which often occur on numerous sites on the same protein)
- pre-enzymes made for secretion and pro-enzymes that are activated by cleavage
- acylation and ubiquitination
- non-enzymatic modifications like oxidation, so a given protein exists in the cell in different oxidized states.

These modifications can modulate biochemical function either directly or indirectly by altering the pathway in which a gene product operates. Cellular function can be changed similarly. Cellular function, and in some cases even biochemical function can also be changed simply by changing the location where the protein is found in the cell or by binding it to another protein or small molecule. A proteomics study that aims at understanding function is incomplete without taking these aspects into account.

One breakout session addressed "Metabolic Pathways and Post-Translational Modifications," which defined "function" to the group participants, as reported by Edward Dennis, University of California, San Diego, and Eugene Bruce, National Science Foundation. It was noted that most speakers had emphasized inventorying, categorizing, high-throughput screening, methods, and qualities such as completeness in defining proteomics. While these were the goals of the genomics revolution, they should not be the goals of proteomics, stated some of the participants. In contrast with genomics, which is finite in scope, proteomics, especially when function is included, is essentially without limit. "Whatever is done, completeness will be very difficult, if not impossible, to achieve from the viewpoint of function in proteomics," remarked session co-chair Edward Dennis. "Proteomics is many orders of magnitude more complex than genomics. It has been suggested that there are about 300,000 human proteins, thinking only about splice variants and post-translational modifications."

The list goes on and on when trying to get one's hands around the number of discrete proteins that exist. Thus, instead of trying to count proteins some researchers suggested focusing on the life cycle of a protein. During its lifetime a protein undergoes phases of translation, maturation, regulation, and termination. Each of these phases involves numerous discrete proteins that interact with each other, and each phase involves protein modifications; so a given protein exists in an enormous number of discrete compositions and complexes, as stated previously in the report. There also exist many states imposed by protein:protein interactions, which change the nature of a protein, by ligands including a variety of metal ions, by activators, inhibitors, inducers, and this list goes on and on. There are even non-enzymatic modifications like oxidation that occur; so a given protein may exist in the cell in different redox states. Thus,

this is really a combinatorial problem, explained Dr. Dennis, with both transient, and one might call them somewhat permanent, changes that occur to the protein as it undergoes its life cycle. Structural changes in the protein conformation can also lead to the development of a disease state. Prion diseases are clear examples where modifications in the structural conformation of a benign protein can lead to changes in normal function. An understanding of all the structural and functional states for even a single gene product is a huge, complex task, but one that must be considered when annotating proteins.

A number of conclusions arise from these considerations. The first is that a complete description of the function of any gene product must include aspects of both spatial and temporal changes in the protein, including changes of state. Gerald Carlson from the University of Missouri, Kansas City, suggested that we are most interested in the steady-state proteins that exist at some point in metabolism, but we are also interested in looking at all other states on the way to and after steady state. To begin to handle proteomics conceptually one must integrate the experimental results with the enormous amount of data on the computational side, and it is a huge undertaking to even begin to figure out how to relate all those states that are so important.

An interesting example of a chemical function genomics program was given by Thomas Leyh, co-chair of the "Structure Function" breakout session, who outlined an initiative intended to provide a functional genomics counterpart to the structural initiative already under way. The core of this multifaceted program, the subject of a 2001 National Institutes of Health workshop, is to perform large-scale mutagenesis and protein functional studies to create a database that assigns catalytic, ligand-binding, or other functions to the highly conserved, non-structural core residues for every protein family. A compendium of molecular function annotation will be propagated across relevant databases to establish links and assign molecular function to specific biological phenomena. While the design of the program tightly couples it to the structural genomics initiative (whose mission is to provide a representative structure for each protein family) it also includes interfaces with programs dedicated to the identification of protein function and the development of bioinformatic recognition algorithms, among others. Such a compendium would be extremely valuable to the biochemical and other scientific communities, and the program would establish the classical structure-function equation on a genomic scale.

Proteomics is far more complex than a simple profiling of the protein content of a cell, even with potential modifications of the proteins and protein:protein interactions included. Profiling of gene expression or protein expression is a useful tool but in most instances gives little direct information about biochemical function, although sometimes cellular functions do emerge. Among other problems with these approaches the correlation between mRNA levels and protein levels is poor for all but the most highly expressed genes. The view of function presented here makes this complexity apparent. A final point was that the field needs more emphasis on what a protein does, not just which proteins exist under what conditions.

Bioinformatics

There are two ways that function is being determined at this time on a genome-wide scale. One is essentially bioinformatics driven and the other uses structural information. Bioinformatics involves, among other things, sequence comparisons and structure comparisons. These can be carried out on a genome-wide scale, as are comparisons of profiles of gene expression. Proteomics, as it is currently implemented in most instances, is geared towards

comparisons of datasets of profiles of protein expression, usually determined by mass spectrometry.

Sequence comparison can be powerful especially if families of related sequences are identified. However, it is becoming apparent that not only can function diverge markedly when two sequences differ by 50 percent or more, in some instances sequences that are more than 90 percent identical code for proteins that operate on completely different substrates and have no cross-reactivity. Assignment of biochemical function from sequence data alone should always be regarded as tentative without confirmatory experimental evidence. Most functional annotation errors in genomics databases probably arise this way.

Structural Proteomics

Among the possible experimental ways of approaching the problem of function determination on a large scale, the one that has received the most emphasis thus far is the use of structural information. Predicated on the assumption that the three-dimensional structure of a protein will often provide information about its biochemical and cellular functions, the structural approach is being applied on a genome-wide scale in a number of independent initiatives. Although in many instances at least the chemical function of an enzyme can be guessed from its overall fold, even that deduction is often problematic, and assignment of higher levels of function is practically impossible without additional information. This problem is exacerbated when membrane-associated proteins are considered. Between 25-40 percent of the proteins in the cell are estimated to be membrane associated (depending on the organism). The database of membrane protein structures is very small and the methods for determining those structures are very difficult and uncertain.

Cheryl Arrowsmith, a structural biologist from the Ontario Center for Structural Proteomics at the University of Toronto, discussed her group's research on structural proteomics. She emphasized the difference of structural proteomics from structural genomics because they work on proteins, not genes. The focus of her proteomics research is to use X-ray crystallography and NMR spectroscopy to determine the three-dimensional structures of proteins on a genome-wide scale. She is particularly interested in examining the extent to which protein structure can reveal protein function. The model system used is *Methanobacterium thermoautotrophicum*, whose sequence was completed at the time the project was initiated in 1998. Since that time, her laboratory has evaluated thousands of proteins by subcloning into bacterial expression systems, performing either NMR studies or X-ray diffraction on soluble and relatively clean purified protein. They have also evaluated hundreds of proteins from a number of different bacterial, viral, and yeast genomes. However, the number of proteins that give structural samples was low. "There is a huge attrition rate in going from cloned genes to those that can be readily expressed in bacteria, are soluble in bacteria, can be purified, give good crystals or promising NMR spectra, and these would be very good in terms of getting a structure." The attrition rate overall is about 85-95 percent of genes that are tried, in other words, approximately 5-15 percent of bacterial or archaebacterial genes can be processed straight through to three-dimensional structures using a single protocol (e.g., single expression conditions, single purification procedure), according to Dr. Arrowsmith.[13] The numbers are worse for eukaryotic systems.

[13] Christendat, D., et al. (2000. *Nat. Struct. Biol.,* 7(10): 903-909.

"Clearly one needs to try multiple procedures for protein expression, purification, and crystallization in order to improve the success rate for structures, " said Dr. Arrowsmith.

She has confirmed these difficulties in a number of other species and systems, and she reported that many of the other National Institutes of Health centers participating in the project are seeing these sorts of statistics as well. Only in a few cases have they had the opportunity and actually gone on to do functional studies of these proteins. Even with proteins of known function, such as spermidine synthase, the determination of structure can be useful in proposing an atomic model and thus a better understanding of the mechanism of enzymatic function. Dr. Arrowsmith's group was among the first to solve the structure of this protein. There are thousands of clones and proteins that have been prepared in the Ontario Center for Structural Proteomics and in many of the other centers; and these clones are available for further functional analysis. "I think this is a huge resource that is being generated, and it should be exploited through projects that emphasize [biochemical] functional analysis of proteins," said Dr. Arrowsmith.

Cellular Function

Protein location can be determined by such genome-wide techniques as green fluorescent protein (GFP) tagging, and protein:protein interactions can be determined by affinity chromatography, immunoprecipitation, and yeast two-hybrid experiments. Databases resulting from these methods are beginning to emerge, but they are of uncertain accuracy. Recent comparisons of independently obtained databases for yeast proteins suggest that location determination is fairly robust but protein:protein interactions are at best determined with less than 50 percent overall accuracy. Clearly more reliable methods are needed, and efforts to create protein chips for profiling of interactions with proteins and small molecules appear promising.

One useful addition to the available arsenal of function-finding tools would be a database of three-dimensional motifs of biochemical function. Such a database would contain those structural elements that participate in ligand binding and catalysis for proteins of known function. This database could be searched in a manner similar to sequence database searches whenever a new protein structure is determined. Another useful tool would be, for each protein family, a database of mutations with functional characterization. Essentially this database would provide a link between a mutation at a particular site, a genetic lesion, a metabolic lesion and even a phenotype such as a disease.

Once again it was stressed that proteomics should be considered as a much broader field than would be apparent from early efforts, which have focused on cataloging levels of protein expression. Ideally it should encompass efforts to obtain complete functional descriptions for the gene products in a cell or organism. Because of the complexity of functional description, clearly more than one technique is required and no one existing technique should be emphasized in preference to any others. This goal may be beyond the reach of existing technologies, even for small numbers of proteins, but it is the direction in which the field must go.

Applications

The application of proteomic technologies to clinical research and public health in general is an immediate goal of proteomics. A distantly related goal is the eventual application of proteomics to environmental, agricultural, and veterinary research, research areas that are far less developed than clinical applications. Thus, essentially all the applications discussed in the formal lectures and breakout sessions centered on clinical applications.

Clinical proteomics aims to discover proteins with medical relevance, said Alan Sachs, a director of R&D at Merck. Such discoveries can be defined broadly as those that identify a potential target for pharmaceutical development, a marker(s) for disease diagnosis or staging, and risk assessment, both for medical and environmental studies. Alan Sachs and Denis Hochstrasser co-chaired the "Clinical Aspects" breakout session and covered a wide range of issues: consent, samples, platforms, phases of diagnostic development, data analysis, and definition. (Note that there is a difference between developing biological insight and identifying clinically important diagnostic and prognostic protein-based assays, as one reviewer of this report has suggested: "By studying protein interactions, or splice forms, or abundance, one might be able to effectively distinguish between healthy and diseased tissues. One of the great promises of genomics, and one that has captured the imagination of the public, is the idea that we might move toward personalized medicine through broad genomic or proteomic surveys, what is often called 'pharmacogenomics')."

Samples

Julio Celis illustrated the potential of proteomics to the study of diseases during his talk of his research on bladder cancer. He stated, "one must take into consideration the set of samples you are going to use." "Biopsies," said Dr. Celis, "and other types of samples can be highly heterogeneous in terms of cell type, stage of pathology, etc., and this presents a challenge for proteomic analysis that must be faced." Experimental research also must consider the use of various types of cell lines, primary tissues, body fluids, and various animal models. Each of these may impose considerations on the types of techniques used for proteomic analysis.

As became apparent from several discussion participants, it is currently quite difficult to identify the best procedures for obtaining and storing samples for proteomic analysis, because the techniques used to analyze the samples are constantly changing, making it difficult to arrive at a consensus protocol for sample preparation that would be best for a particular analysis method. Thus, Dr. Sachs and others agreed that various strategies for handling samples or standard operating procedures, and long-term storage will need to be co-developed along with evolving protein detection methods. Dr. Hochstrasser raised the point that "we don't know how to store the samples if we don't know how we plan to use them later." This is important, especially considering that most proteins stored in the freezer at –20°C are useless for specific types of clinical research after a few months, according to Hochstrasser. The question of storage remains a problem because the technology for measurement in the clinics has not evolved, said Hochstrasser, "yet we need to start worrying about sample storage now."

Related to this is the nature of the samples. "Defining 'normal' is a major problem," stated Dr. Celis. As many researchers know, the pathology of samples can be open to interpretation,

and robust parameters must be delineated and adhered to when defining normal versus various stages of pathology.

Consideration of the various proteomic methods under development suggests that the size of samples required will be dictated largely by the constantly changing technology. As with all research the nature of the study will dictate the size of the sample available. Dr. Celis noted, "tissue biopsies will impose the most severe restraints, both in terms of size as well as the available clinical data to support the experimental work." Tissue epidemiological studies may provide blood or some other easily obtainable tissue that is not the target tissue of interest, whereas cancer epidemiology likely will provide tumors of different grades of differentiation. Each of these types of studies imposes complexities and limitations on sample size, number, and method of analysis.

The proteome itself has a large, dynamic range, depending on the cells being analyzed, and the location of cells within a tissue could influence its size and nature. Dr. Celis estimated that the dynamic range (i.e., the concentrations of proteins) spanned 12-13 orders of magnitude.

Given the limits of sensitivity of detection and the availability of a suitable amount of starting material, Hochstrasser stated, "I strongly believe that a combination of bioinformatics (dry lab) and chemistry (wet lab) is crucial to finding new diagnostic markers and therapeutic agents." Several participants expressed their belief that no single technology would be sufficient for proteomic analysis and that multiple approaches will be required, at least in the near future.

Ethical Considerations

In addition to the issues surrounding samples being obtained and stored properly certain consent requirements and sample limitations permit clinical samples to be used only once after patient consent has been obtained. In this case consent means both a clear description to the patient regarding how the samples will be used and a disclosure of who will have access to the samples. "Some samples will be anonymous, others will be 'de-identified', and yet others will have restrictions placed on their use," noted Dr. Sachs. For example, samples may have a limitation placed on the type of disease studied or the facility or institution at which the analysis may be performed. Thus, it is important that sample-tracking procedures are in place to ensure that only samples with appropriate consent from subjects are distributed to a specific site for a particular type of investigation.

Development of Diagnostics

Participants of the "Clinical Aspects" breakout session on diagnostics discussed the fact that although the experimental platform used in clinical settings to detect protein markers will change rapidly in the coming years, the underlying principles regarding the stages of going from the discovery of protein markers to their use as diagnostic tools in a community setting will remain reasonably constant. Consequently the criteria used to judge the quality of a marker or markers as diagnostics in a clinical setting are different from those used to evaluate the quality of a marker in the basic science setting. Discussion centered on the fact that the basic researchers developing protein markers, as well as reviewers evaluating such work, must consider the technical aspects of the application and development of such markers so that statistically underpowered or misinterpreted studies using such markers are not initiated or reported. Another reviewer pointed out an important variable to consider in clinical applications, which is

the impact of population or sample variability due to the heterologous nature of individuals. This point corresponds again to the idea of pharmacogenomics or personalized medicine.

Although data analysis (informatics) is addressed elsewhere in this report, several speakers noted that special consideration should be given to adequate data analysis when reporting something as significant as the association of protein markers with a disease. Participant Thea Kalebic from the National Cancer Institute (NCI) recommended publication criteria for reporting the use of marker and clinical samples. Criteria should be specified for the use and analysis of a particular method to avoid incorrect application of a technique or inadequate or wrong interpretation of the results, stated Dr. Sachs. Participant Izet Kapetanovic, NCI, further suggested that a paper be written for the lay audience to describe how algorithmic clustering methodologies are being used to do disease association studies. "I think a lot of physicians, as well as clinical researchers, are not bioinformatics or statistics people, and they would benefit from such a review," stated Dr. Sachs.

Clinical researchers will also need to consider the types of proteins that might be most relevant, noted Dr. Celis. "Because every modification has a functional meaning, [one must also consider] a protein-protein or protein-macromolecule interaction [as well as] cellular distribution, movement, or migration," added Dr. Celis. Regarding techniques, Dr. Celis believes the only available technique that provides a global picture of the cell proteome is high-resolution 2D gel electrophoresis, despite its obvious limitations in terms of the numbers of proteins resolved and the sensitivity of detection. The non-gel approaches based on chromatography and mass spectrometry allow for high-throughput, Dr. Celis noted, but he stated they are not yet ready for the study of complex tissue samples.

Scott Patterson, vice president of proteomics at Celera prefers the high-throughput approaches to clinical applications of proteomics research. "In our search for markers of disease or drug efficacy, and targets for small molecules, therapeutic antibodies, and cellular immunotherapeutics (vaccines), we employ a broad-based discovery approach," stated Dr. Patterson. His team uses chromatography and mass spectrometry as the basic tools in searching for protein diagnostic markers and therapeutic targets in specific diseases.

"Most of you will know Celera for sequencing genomes," commented Dr. Patterson. But as the company decided to embark upon drug discovery based upon its valuable genomics business, the first platform to be built was a proteomics platform. The proteomics component of that strategy is to discover diagnostic markers of disease and targets for therapeutic intervention, said Dr. Patterson. They are specifically focused on proteins that are differentially expressed in disease tissue compared with normal tissue. Contrary to Dr. Celis's approach of performing a high-resolution protein separation at the beginning of the analysis (as is the case for 2D gel electrophoresis), a very high-resolution peptide analysis is performed at the end of the process using chromatography and mass spectrometry. "In its simplest description," said Dr. Patterson, "protein-level analysis is accomplished through targeted capture of classes of proteins (or the depletion of abundant proteins) prior to proteolytic digestion, yielding peptides that are quantitated and identified by MS/MS using one of a variety of platforms [e.g., a MALDI-TOF-TOF-MS or the Voyager 4700 Proteomics Analyzer™']." The MS/MS spectra are identified using search algorithms for spectrum-to-sequence matching (using characterized protein sequence databases or a translation of the Celera human genome sequence). Automated identification can be achieved through spectral matching or spectrum-to-spectrum matching. This overall approach of peptide-level analysis can be employed with isotope dilution strategies

(such as ICAT™ for quantitation of the relative abundance of peptides and proteins from pairs of samples) and without if the fractionation of a series of samples is sufficiently reproducible.

Identification of early markers of disease is important for development of a reliable assay for tissue samples so as to help diagnose disease, provide insight into prognosis and identify risk for disease. These markers are especially important in identifying tumor stages such as with Dr. Celis's work. A therapeutic that derives from this information is also desirable. Proteomic data, in combination with microarray (gene expression) data, pathology, immunohistochemistry, etc., have the potential to identify novel markers for early detection, diagnostics, prognostics, and response to treatment, concluded Dr. Celis. Drug discovery and improvement in public health and environmental research will require a combination of all these and other technologies. Salvatore Sechi, National Institute of Diabetes, Digestive and Kidney Diseases, emphasized that although it is clear that the emphasis in the clinical community is on marker discovery, the technology needed for clinical assays and high-throughput proteomics has not evolved yet. It is important to recognize however, that developing clinically relevant diagnostic and prognostic tests is something separate from developing biologically relevant insight into disease. While these two goals are not mutually exclusive, they are not necessarily overlapping, notes one reviewer. It may be a relatively simple matter to identify patterns of gene expression that can be correlated with a clinical outcome but that does not provide an immediate insight into the underlying mechanism of the associated disease state. This does not mean that such a prognostic protein expression fingerprint is not useful. Any tool that can help improve and influence treatment has a great potential to affect patients' lives. This, it seems, defines a mandate for proteomics in the twenty-first century.

Computational Methods and Bioinformatics

Computation has become an essential component of biological research. The great quantity and diversity of the data being generated by different technologies is daunting and impossible to organize or oversee without computational assistance. In functional genomics, a great deal of effort has been devoted to developing community-based standards for reporting gene expression data to allow others to replicate experiments. The same will need to be done for proteomics to validate across the different technologies. Perhaps never before has a bioinformatics problem of this magnitude been approached. No one person can integrate and organize all the relevant information for even a single protein being studied without access to computational tools. Sequence, structure, expression profiles, functional assays, protein-protein interaction from yeast two-hybrid experiments or protein chip experiments, and other data all provide information on different aspects of proteins whose functions and roles we are only beginning to understand. Without effective and integrated databases to store and retrieve these data, and advanced computational methods such as pattern recognition and other machine learning approaches to analyze and interpret them, the full implications of these data will not be realized. A few years ago the typical biologist may have had little reason to turn to a computer for insights or information. Today the story is very different.

To paraphrase an old adage, "No protein is an island," and researchers who are unable (or unwilling) to use all available data do so at their own peril. Computation can provide powerful tools to enable the detection of subtle relationships between data and suggest hypotheses for

experimental validation. In addition to the traditional hypothesis-driven research to which we are all accustomed, computational methods provide a new paradigm: 'computationally assisted hypothesis generation'. Far from supplanting the biologist's intuition, understanding, and experimentation, computation can provide an added dimension enabling additional insight and understanding. Our ability to take advantage of the technological advances in genomics and proteomics will hinge greatly on our ability to integrate computation into the research and discovery process.

For instance, experiments performed on one protein often have relevance for other proteins, not only within the context of the organism from which that protein was derived (i.e., paralogues) but also within the context of other organisms containing orthologous proteins. Researchers investigating an individual protein, say, one involved in disease resistance in potatoes, would miss a wealth of information and experimental data if they were not aware of work being done on related plants, such as Arabidopsis, tomatoes, or rice. In fact, many disease-resistance proteins in plants have orthologues in insects and animals, and experiments on one will shed light on the function of another. Entire pathways in one organism have analogues in others. Genetic experiments in one organism will have implications for related organisms. Three-dimensional structures solved for one protein can be used to predict the structure of other proteins whose sequences are similar. Residues shown to be catalytic in one protein are likely to play a similar role for related proteins. Taken alone, proteomics data being generated (microarray, protein chip, structural, yeast two-hybrid, mass spectrometry) can provide important insights. Taken in concert and integrated, these data provide a context for understanding the complex interactions and roles of these biological molecules.

To take full advantage of the information contained in these data, computational development of two basic types is necessary: (1) database infrastructure enabling efficient and biologically intuitive storage and retrieval, and interface design to enable different databases to communicate with each other, and to allow investigators with disparate backgrounds to access the information in these databases; and (2) intelligent systems, agents, and software tools to discern relationships between data, and to generate hypotheses that can be tested experimentally. Such bioinformatics development may also be used to help answer fundamental questions in biology, which have never been posed. In addition, training the next generation of scientists to make the kinds of contributions that will be critical to discovery in this new century of proteomics must not be ignored. We discuss each of these issues separately.

The breakout group "Computational Methods and Bioinformatics," led by Kimmen Sjolander, University of California, Berkeley, and Dagmar Ringe, Brandeis University, discussed database issues.

Database infrastructure and interface design

For historical reasons most biological databases have been produced primarily by the biological community, while most computational tools have been produced by the mathematical and computational communities. This has resulted in databases that are often not easily amenable to automated data-mining methods, unintelligible to some computers, and computational tools that are often non-intuitive to biologists. Biological databases have inherent complications stemming from the nature of the information they contain and the dependence of computational methods on these data. Most biological data are not digital, making machine-readability of the data (for automated data-mining) impossible. In addition, the lack of standardized nomenclature

and ontology, the use of protein aliases (leading to ambiguity), the lack of interoperability across databases, and the presence of errors in database annotations have hindered and complicated the use of computational methods. While the computational biology community has begun dialogue in this area, a great deal of work needs to be done before access to information becomes routine and accessible to the computational non-expert.

Development of new methods

Computational methods are based on models, whether mathematical or biological. As biologists achieve new insights based on new data being generated by experiment existing models will be reevaluated and changed accordingly. New methods will need to be developed and existing methods refined. In all cases development of benchmark test sets is critical for the assessment of method accuracy and reliability. As more information becomes available in databases more robust tools and more intuitive methods for finding relationships within these data will be needed.

Training

Biology is being changed by computation. And computation, in turn, is being changed by biology. Researchers working at the interface of computation and biology are increasingly in demand. New degree-granting programs and departments are springing up around the world to train the next generation of scientists in this interdisciplinary field. To be effective in this new age of computationally assisted biological discovery biologists must receive training in statistics, mathematics, and computation, and become expert in the use and interpretation of the results of computational methods. It was suggested that life scientists learn at least some simple scripting language such as PERL, and a database language such as S2L. For computer scientists working in this area, training in life sciences is necessary. Both groups must learn to speak a common language.

The information explosion has presented an opportunity and a challenge to the biological and computational communities. The wealth of data being generated needs to be integrated in order to define a system from multiple viewpoints and to understand a system from different sets of empirical data. Such integration is possible only with computational tools that can find relationships within the data and use these relationships to create testable models. Such tools must also be user friendly.

Proteomics: A Coordinated International Effort

"It was a report by a National Academy of Sciences panel [in 1988] chaired by Bruce Alberts [president of The National Academies] that basically laid out the blueprint that became the Human Genome Project, and a wise blueprint, indeed," said Francis Collins. Dr. Collins stressed the success and importance of having a large international consortium of laboratories involved in the Human Genome Project. "Forming large teams and international teams [was critical] and this is the group that carried out the large-scale sequencing effort or at least the leaders of many of the labs that were involved in that six-country enterprise," he stated. Dr.

Collins hopes the same will be true for proteomics research. "I think it was very helpful that all of the groups had the capacity for large-scale effort, had an open door to come and join in and that this was an international enterprise, also something that I had hoped would happen for proteomics in the public sector because after all science is an international enterprise. That is one of the joys of the whole thing."

Protein Structural Initiative

In September 2000 the NIGMS initiated the support of seven centers to begin work on developing an approach to structural genomics in order to reap the benefits of the multiple genome projects being undertaken worldwide. Two more centers were subsequently added in September 2001, forming what is now known as the Protein Structural Initiative. The idea for the consortium resulted from a planning meeting in April 2000 jointly sponsored by the NIGMS and the Wellcome Trust, and there was wide representation from several countries. It was essentially a policy meeting to come up with ideas about how to consider structural genomics in a worldwide setting, said Marvin Cassman. He defined structural genomics as "the discovery, analysis, and dissemination of three-dimensional structures of proteins, RNA, and other biological macromolecules." However, the focus is primarily proteins.

Currently the NIGMS is helping to fund nine pilot projects to determine the best strategies for a large-scale production process. Each project is required to include all the components for the effort based on genome-driven target selection. These components include protein production, crystallization, structure determination, theoretical analyses for homology modeling, target selection testing approaches for full coverage of protein families, development of high-throughput methods, and best management practices. The consortium consists of industrial and international collaborators with 66 investigators and 24 institutions, according to Dr. Cassman. They all involve the development of technology.

Proteomics is generating novel requirements for scientific collaboration. Several drivers and barriers to collaboration were discussed by members of the breakout called "Implementation: Necessary Policy and Infrastructure Conditions for Collaboration," led by James Myers, of the Pacific Northwest National Laboratory, and Richard Morris, National Institutes of Health, Division of Allergy, Immunology and Transplantation. "As has been the case with genomics research, the promise of medical benefits is a major driver of proteomics research," stated Dr. Myers. Participants of the breakout session discussed how collaboration may create needed economies of scale in research, for example, by making it possible for groups of scientists to strive for completeness in critical description and annotation of proteins, an effort that would surpass the capacity of any individual scientist. It was suggested that success in this field would be based on the extent to which it selected and strived to characterize specific organs, pathways, and systems with completeness.

Research Collaboration

Proteomics research also poses unique challenges to collaboration. Due to its close tie with application, the field imposes barriers in intellectual property and authorship and other issues of attribution. As with genomics, tensions between collaboration and competition (as well as between government and industry) are also heightened in proteomics research. A global distribution of resources, expertise, and potential targets are necessary for collaboration and

success in the field. In terms of international focus the same problems arise here that arose in genomics. The proteomics techniques and data collection and expertise occur in one locale, but in developing countries there are important diseases and other health problems that are affecting millions of lives and should be studied. However, one has to address some of the differences in policies such as informed consent between countries if one is actually going to get some work to happen there.

Policy, organizational, and technology solutions were discussed as interdependent variables, and as essential future enablers of proteomics research. For example, the need to construct and use diverse data sets was examined. Proteomics poses unprecedented demands for integrating biological samples, electronic data, outputs from instrumentation, and expertise from multiple disciplines. Interactions across disciplinary boundaries that were crucial for the genome project may be even more so for proteomics due to the variety of expertise needed. The group examined how this might be coordinated and made accessible through shared user facilities. Such collaboration may also help overcome shortages of trained experts who can contribute to proteomics research through particular fields including mathematics, statistics, physics, chemistry, computer science, and engineering. The potential drawbacks of such shared facilities were identified, including the requirement to travel away from one's home institution and the overhead costs of dealing with multiple facilities. Such costs could be mitigated by the creation of virtual facilities providing aggregated capabilities over the Internet. Dr. Myers suggested that Internet-based collaboratories, or laboratories without walls, could permit researchers to easily share data, instruments, and expertise.

In addition, the group discussed the need to provide intellectual credit to developers of shared and reference resources (e.g., samples, instruments, software). The recognition by (and pressure for) scientific publication in academia leaves little room for scientists to work on reference resources and database construction. Since the best makers of such tools and databases are those who actually use them, scientists should be supported not only for the development of the tool (which often is hard to get through conventional funding means), but also for applications of the tool toward a biological problem. A scientist should be able to get funding for making a tool that will help with an important problem; clever grant writing should reflect this contribution toward their own work and toward the benefit of the scientific community. This problem could also be solved by adjustments to institutional tenure policies, whereby credit is given to those who take time away from the bench to develop critical databases, websites, and software for general use by the scientific community. The participants recommended that tenure committees and granting agencies should be able to recognize those kinds of contributions.

Information technology could also play a role by assisting managers to develop broad metrics of authorship or enabling them to track the pedigree or provenance of intellectual contributions at a finer grain than publication credit.

There are very few people trained in the multiple areas necessary for proteomics research, from computation to experimentation across disciplines of biology and chemistry. The need to train new researchers and to encourage practicing researchers to broaden their expertise could be met through support for additional fellowships and sabbaticals, which could be made more effective and less disruptive through remote collaboration capabilities. Difficulties in sharing and comparing data from different techniques and disciplines could be reduced through the promotion of standardization efforts and open, extensible data format standards.

Overall it was agreed that the pursuit of proteomic research needs to progress on an international scale with broad support from governments and industries alike. In turn, this

creates the need for international professional organizations, as well as software applications and standards to enable international collaboration. "There is the sense that just like the promise of the genomics revolution, there are many, many things that can be done that are of practical importance," stated Dr. Myers. "Collaboration to speed up that process is really a big driver here and driving it more than just the basic research kind of ideas."

Conclusion

The most useful definition of proteomics is likely to be the broadest: proteomics represents the effort to establish the identities, quantities, structures, and biochemical and cellular functions of all proteins in an organism, organ, or organelle, and how these properties vary in space, time, and physiological state. Proteomics is thus a huge, long-term task, much more involved than sequencing the genome.

At the time the Human Genome Project was begun the basic methodology for sequencing DNA, Sanger's dideoxy chain termination, had already been in place for five years and the task, while challenging, was essentially one of efficient scale-up. One of the main lessons from this symposium is that proteomics has not yet reached that stage. There is much work to be done in the technology sector. Perhaps the most important area for investment right now is in platform technology development for high-throughput systems. Other areas where emphasis might be placed in the short term include protein markers and clinical assays of disease, as well as the use of less complex model systems. Quality controls and annotations are needed at all levels. There are also several barriers that remain to translate proteomics results into clinical applications, but progress is being made as described in this report. There is room for both big and small science, stated George Kenyon. No one group, company, or government entity is going to solve these problems; there is a great need for interdisciplinary collaboration, locally, nationally, and globally.

REFERENCES

1. Anderson, N.G. and Anderson, N.L. (1979) *Behring Inst Mitt.* **63**: 169-210
2. Clark, B.F.C (1981) Towards a Total Human Protein Map. *Nature* **292** (5823): 491-492
3. Christendat, D., et al. (2000) Structural proteomics of an archeon. *Nat. Struct. Biol.* **7**(10): 903-909.
4. Griffin, T.J., Gygi, S.P., Rist, B., Aebersold, R., Loboda, A., Jilkine, A., Ens, W., Standing, K.G. (2001) Quantitative proteomic analysis using a MALDI quadrupole time-of-flight mass spectrometer. *Anal. Chem*. **73**(5): 978-86.
5. Gygi, S.P., Rist, B., Gerber, S.A., Turecek, F., Gelb, M.H., Aebersold, R. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnology* **17**: 994-9.
6. Zhu, H., et al. (2001) Global Analysis of Protein Activities Using Proteome Chips. *Science* **293**(5537): 2101-2105.
7. <http://grants.nih.gov/grants/guide/rfa-files/RFA-HG-02-001.html/>
8. <http://hipseq.med.harvard.edu/MEDGENE/>
9. <http://mgc.nci.nih.gov>
10. <http://us.expasy.org/sprot/hpi/>
11. <http://www.signalsmag.com/> (11/02/99)
12. <http://www.hupo.org/>

APPENDIX A

SPEAKER BIOGRAPHIES

RUEDI AEBERSOLD is a founding member of the Institute for Systems Biology in Seattle, Washington, where he leads the proteomics program of the Institute. The program is focused on developing new methods and technologies for quantitative proteomics and for applying this emerging technology to enhance our understanding of the structure, function, and control of complex biological systems. Current applications of quantitative proteomics technology at the Institute for Systems Biology are directed towards the discovery of proteins markers that differentiate cancer cells from their normal counterparts, to the investigation of the mechanisms of fundamental cellular processes by the comparative analysis of the gene and protein expression profiles in cells at different states, and to studies in the area of medical microbiology. Dr. Aebersold completed his undergraduate studies in biology at the University of Basel, Switzerland in 1979 and received a Ph.D. in cell biology at the Biocenter of the University of Basel in 1984. Holding fellowships from the Swiss National Science Foundation and EMBO he joined the California Institute of Technology as a postdoctoral fellow (1984-86) and remained at Caltech as a senior research fellow (1986-88). In 1988 he joined the University of British Columbia in Vancouver as an assistant professor in the Department of Biochemistry and Molecular Biology and as a senior investigator at the Biomedical Research Centre. In 1993, he moved to the University of Washington as an Associate Professor in Molecular Biotechnology and was promoted to full Professor in 1998. He served as the Associate Director for the Science and Technology Center for Molecular Biotechnology from 1994-2000. In 2000, he left the University of Washington and joined the Institute for Systems Biology as co-founder and full faculty member. Dr. Aebersold is a consulting editor for the journal *Physiological Genomics*, has been a member of the Editorial Advisory Boards of *Protein Science* (1992-'98*), Functional Proteomics* (1999- present*), Analytical Biochemistry* (1991-present*) Functional and Integrative Genomics (1999-present) and Electrophoresis* (1989-1993*) Journal of Proteome Research* (2001-present) and an Associate Editor for *Molecular and Cellular Proteomics* (2001-present).

CHERYL ARROWSMITH is a Senior Scientist at the Ontario Cancer Institute and Professor in the Department of Medical Biophysics at the University of Toronto. She received her B.Sc. degree in chemistry from Allegheny College and Ph.D. in chemistry from the University of Toronto. She carried out postdoctoral research at Stanford University where she applied NMR spectroscopic methods toward understanding protein structure and function. Dr. Arrowsmith's current research focuses on the use of NMR and biochemical methods for understanding the structure-function relationships of proteins. Most recently she has been involved in applying structural biology methods on a genome-wide scale and is co-founder of Integrative Proteomics Inc., a company that integrates multiple technology platforms in protein structure and function for use in pharmaceutical discovery.

MARVIN CASSMAN is the Director of the National Institute of General Medical Sciences at the National Institutes of Health. After serving as a NIH postdoctoral fellow at the University of California, Berkeley for two years and as an assistant professor at the University of California, Santa Barbara for seven years, Dr. Cassman joined the NIGMS in 1975. There, he has held several different positions, including chief of the Cellular and Molecular Basis of Disease Program and Director of the Biophysics and Physiological Sciences Program. Prior to becoming the director of NIGMS, he served as deputy director from 1989 to 1996 and acting director from 1993 to 1996. Dr. Cassman served on the staff of the House Subcommittee in Science, Research and Technology from 1982 to 1983. He also was a senior policy analyst at the Office of Science and Technology Policy from 1985 to 1986. Dr. Cassman received a B.A. in 1954, a B.Sc. in 1957, and a M.S. in 1959, all from the University of Chicago. He received his Ph.D. from Albert Einstein School of Medicine in 1965.

JULIO CELIS is the Scientific Director of the Institute of Cancer Biology and Danish Centre for Human Genome Research, Danish Cancer Society, Copenhagen, Denmark. Dr. Celis is also the Secretary General of Federation of European Biochemical Societies (FEBS). He is on the council of the European Molecular Biology Organization (EMBO) and President of the E-BioSci Committee of EMBO, which is creating a new database network involving seven European partners from four countries with expertise in providing access to and retrieval of information in the life sciences in digital form.

BRIAN T. CHAIT is the Camille and Henry Dreyfus Professor at Rockefeller and Head of the Laboratory for Mass Spectrometry and Gaseous Ion Chemistry. He is director of the NIH-funded National Resource for the Mass Spectrometric Analysis of Biological Macromolecules. His current research focuses on investigations of new techniques for volatilizing and ionizing proteins, designing and constructing novel mass spectrometers, and developing mass spectroscopic-based methodology to assist in the solution of challenging biological problems. Dr. Chait and his colleagues are applying these tools to the solution of biological problems that involve the rapid identification of proteins, the elucidation of posttranslational modifications that regulate the function of proteins, and the definition of sites of functional interaction between biomolecules. Dr. Chait received his B. Sc. (1969) and B. Sc. (Hons) (1970) from the University of Cape Town and D.Phil. (1976) in experimental nuclear physics from Oxford University. He carried out postdoctoral research at the University of Manitoba, where together with Professor Kenneth G. Standing, he constructed the first pulsed ion bombardment time-of-flight mass spectrometer. Subsequently, Dr. Chait moved to the United States where he joined the laboratory of Professor Frank H. Field, and constructed a number of mass spectrometers designed to measure biological macromolecules.

FRANCIS S. COLLINS is the Director of the National Human Genome Research Institute at the National Institutes of Health. He oversees the human genome project, a complex multidisciplinary scientific enterprise directed at mapping and sequencing the entire human DNA, and determining aspects of its function. A working draft of the human genome sequence was announced in June of 2000, an initial analysis was published in February of 2001, and the completed sequence is anticipated in the spring of 2003. From the outset, the project has run ahead of schedule and under budget, and all data has been made immediately available to the scientific community, without restrictions on access or use. Dr. Collins received a B.S. from the University of Virginia, a Ph.D. in Physical Chemistry from Yale, and an M.D. from the University of North Carolina. Following a fellowship in Human Genetics at Yale, he joined the faculty at the University of Michigan, where he remained until moving to NIH in 1993. His research led to the identification of genes responsible for cystic fibrosis, neurofibromatosis, and Huntington's disease. He is a member of the Institute of Medicine and the National Academy of Sciences.

DENIS HOCHSTRASSER is the director of the Clinical Pathology Department of the Geneva University Hospital, and President, Clinical Medicine, University of Geneva, Switzerland (www.hcuge.ch). He is also Head of the Central Clinical Chemistry Laboratory of the Geneva University Hospital. He is full Professor both to Geneva's Department of Pathology, Medicine Faculty and to the School of Pharmacy, Sciences Faculty. He was one of the founders of the Swiss Institute for Bioinformatics (www.expasy.org) and he is a scientific founder of GeneProt Inc (www.geneprot.com). Dr. Hochstrasser obtained his M.D. at Geneva Medical School, and after a visiting year at Duke, did a two-year residency and internship at UNC Chapel Hill. Dr. Hochstrasser oversees all computing activities in Geneva's Faculty of Medicine, as well as running an internationally renowned proteomics research group. Dr. Hochstrasser is known for his pioneering developments of what is now called proteomics. Dr. Hochstrasser's innovations in the methodology of two-dimensional gel electrophoresis, and his perception of the need for integration of the methodology with electronic data processing, have contributed decisively to the technique's becoming one of the main protein separation methods used in proteomics. In his own group, Dr. Hochstrasser had the foresight more than 15 years ago to initiate work on software development for what is now called proteomics. He put together a world class group, including Dr. Ron Appel, which produced the leading

2D gel analysis software package Melanie, widely used in both academic and commercial laboratories, and the standard database Swiss 2D-PAGE. Among other major developments, he was instrumental in bringing Ron Appel and Amos Bairoch together, which led to the creation of one of the first few web sites in the world, ExPASy, and the very first to be devoted to the life sciences. Most recently he and his team have been responsible for the Molecular Scanner, a technique that promises to enhance proteomic analysis in the future.

JOSHUA LABAER is the Director of the Institute of Proteomics at Harvard Medical School. He attended the University of California at Berkeley as an undergraduate where he was awarded the University Medal. His studies continued at the University of California, San Francisco where he attended medical and graduate school and where he studied steroid regulation of DNA transcription and protein-DNA interactions with Dr. Keith Yamamoto. Dr. LaBaer completed his clinical training at the Brigham and Women's Hospital in Boston, where he specialized in internal medicine and the Dana-Farber Cancer Institute in Boston, where he studied medical oncology. He also pursued research interests at the Massachusetts General Hospital in Boston in the areas of breast cancer, mammalian cell cycle regulation and cell cycle checkpoint genes. He is currently an Attending Physician at the Dana-Farber Cancer Institute and holds an academic appointment through the Department of Biological Chemistry and Molecular Pharmacology at Harvard Medical School. Together with Dr. Ed Harlow, Dr. LaBaer founded the Harvard Institute of Proteomics in the spring of 1999. The mission of the Institute is to use the information arising from the genome projects to revolutionize the study of proteins and their functions by enabling scientists to produce and study proteins hundreds or thousands at a time. The Institute has started an ambitious project to build a complete genome-wide collection of full-length genes in a recombinational cloning vector. The output will be a large repository of verified cloned genes that will allow the entire proteome to be used for any experimental needs. The recombinational cloning system allows genes to be transferred *en masse* into any expression vector essentially overnight. The goal of the Institute is to create the human repository first and then other appropriate model organisms. A fundamental principle underlying the repository is that the full-length clones in the repository, along with the technology to use them, will be broadly available without restriction to all scientists — academic, governmental or commercial — and will likely become a universal research standard. In this way, scientists everywhere will be able to obtain large collections of genes that can be transferred easily into the most relevant experimental systems. Since the Institute started in the spring of 1999, it has been developing the informatics and automation to begin constructing this large repository. The Institute now has a fully functional tracking database and workstation automation with the capacity to process >400 clones per week. At present, the Institute has completed a first pass of the budding yeast genome with ~95% success and there are over 3000 clones in the human repository. The Institute is now organizing a consortium of public and private entities to fund the completion of this important resource.

SCOTT PATTERSON is Vice President, Proteomics at Celera Genomics Corporation in Rockville, Maryland. He has been at Celera Genomics since November 2000 establishing an industrial-scale Proteomics facility for diagnostic marker and therapeutic target discovery and development as part of the evolution of Celera into a next-generation biopharmaceutical company. The Proteomics approach employs advanced chromatography-mass spectrometry based analyses he pioneered for therapeutic protein drug discovery at Amgen Inc. (Thousand Oaks, California) where he was from 1993. When he left Amgen Inc. he was Head of the Biochemistry and Genetics department and Proteomics Team Leader. During that time he developed methods for the identification of low-level quantities of gel-separated proteins by PSD-MALDI-MS, and complex peptide mixtures by LC-MS/MS (chromatography-based Proteomics) and has published papers on these topics including a number of review articles. The Proteomics Team at Amgen Inc. was formed in 1997. Dr. Patterson came to Amgen Inc. from Cold Spring Harbor Laboratory, New York (1991-1993) where he was a faculty member. At CSHL he supervised the 2-D Gel Laboratory Core Facility in addition to his own laboratory. His research interests included protein identification technology development and investigation of the molecular mechanisms of

apoptosis. Prior to moving to CSHL, Dr. Patterson held various positions (finally as Supervisor) in the Australian Equine Blood Typing Research Laboratory at The University of Queensland (1980-1990) in Australia where he obtained his B.Sc, and subsequently his Ph.D. in Physiology and Pharmacology.

JOHN WALKER has been engaged in structural studies of the ATP synthases from bovine heart mitochondria and eubacteria for more than 20 years. These studies resulted in a complete sequence analysis of the complex from several species and in the atomic resolution structure of the F_1 catalytic domain of the enzyme from bovine mitochondria. This structure suggested that ATP is made by a rotary mechanism and provided the means for direct demonstration of rotation. Recently, he and his colleagues have established the structure of part of the F_o motor domain, which generates rotation. He has also worked on other vital proteins in mammalian mitochondria. In the early 1990's his group established the sequences of 36 nuclear encoded subunits of complex I (NADH ubiquinone oxidoreductase). Recently they have taken steps towards unraveling its 3-dimensional structure. In addition, throughout the past 10 years he has collaborated with Professor Palmieri at the University of Bari in Italy to identify membrane proteins that transport a range of small molecules in and out of mitochondria. In 1998, Dr. Walker became the Director of the MRC's Dunn Human Nutrition Unit in Cambridge, with the aim of steering nutrition research towards the fundamental understanding of the processes involved in nutrition using molecular and genetic methods. Dr. Walker is a Fellow of the Royal Society, a Fellow of Sidney Sussex College, and an Honorary Fellow of St. Catherine's College, Oxford. Among his honors are the A. T. Clay Gold Medal (1959), the Johnson Foundation Prize by the University of Pennsylvania (1994), the CIBA Medal and Prize of the Biochemical Society (1996), the Peter Mitchell Medal of the European Bioenergetics Congress, and the Gaetano Quagliariello Prize for Research in Mitochondria by the University of Bari, Italy (1997). In 1997, Dr. Walker was awarded the Nobel Prize in Chemistry jointly with Dr. Paul Boyer for their elucidation of the enzymatic mechanism underlying the synthesis of adenosinetriphosphate (ATP). He became a Knight Bachelor in 1999.

APPENDIX B

SYMPOSIUM AGENDA

Monday, February 25, 2002

8:00 a.m. *Registration and Continental breakfast*

8:25 W*elcome*
George Kenyon, Ph.D, Symposium Committee Chair, University of Michigan, Ann Arbor, MI

Plenary Session

8:30 ***Proteomics at NIGMS: Why is Structural Genomics not the Same as Structural Proteomics?***
Marvin Cassman, Ph.D., Director, National Institute of General Medical Sciences, Bethesda, MD

Post Genomic Studies of Mitochodria
John E. Walker, Ph.D. Director of the Dunn Human Nutrition Unit, Medical Research Council, Cambridge, UK

Accelerating Drug Discovery by Targeted Proteomics
Scott Patterson, Ph.D., Senior Director for Proteomics, Celera Genomics Group, Rockville, MD

Large Scale Proteomics in a Clinical and an Industrial Setting
Denis Hochstrasser, M.D., President of Clinical Medicine, University of Geneva; Head of Clinical Chemistry Laboratory, Geneva University Hospital; founder, Geneva Bioinformatics, Switzerland.

10:15 Coffee

10.30 ***Proteomic Strategies in Health and Disease***
Julio Celis, Ph.D. Institute of Cancer Biology and Danish Centre for Human Genome Research, Danish Cancer Society, Copenhagen, Denmark.

Data Collection in Proteomics, What Data and How Much?
Ruedi Aebersold, Ph.D., Institute for Systems Biology, Seattle, WA

Scaling up Proteomics: Lessons Learned from the Human Genome Project
Francis Collins, Ph.D., Director, National Human Genome Research Institute, Bethesda, MD

12:15 Lunch

1:15 ***Structural Proteomics: Part of an Integrated Approach to Functional Genomics/Proteomics***
Cheryl Arrowsmith, Ph.D., Department of Medical Biophysics, University of Toronto, Canada

Manipulating the Proteome; Studying Protein Function in the Genomic Area
Joshua LaBaer, M.D., Ph.D., Director of the Institute of Proteomics at Harvard Medical School, Boston, MA

Proteomic Tools for Dissecting Cellular Function
Brian Chait, Ph.D., Rockefeller University, New York, NY

2:45 Break

3-5:00 pm Breakout Sessions [15 minutes prior to session end, each (co) chair will summarize main points to be presented to all meeting participants]

5:00 Summaries - 5 minutes per breakout session

6-7:30 pm Reception in Great Hall

SYMPOSIUM BREAK-OUT SESSIONS

1) ***Computational Methods & Bioinformatics***
This session will focus on the interface between computational and biochemical methods for the prediction and determination of the functions of gene products. Various experimental methodologies exist for the direct and indirect elucidation of protein function, including structure determination, expression and interaction profiling, knockout experiments for phenotype inference, and mutational analyses. We will examine current computational tools designed to analyze and integrate these disparate types of data into a functional picture of an organism, and discuss both what is possible today, and what we need for future research and development in this complex area.

Cochairs: **Kimmen Sjolander, Ph.D.,** University of California - Berkeley
Dagmar Ringe, Ph.D., Brandeis University, Waltham, MA

2) ***Platform/Emerging Technologies***
Interdisciplinary collaboration in computer science, engineering, and the biosciences has generated rapid advances in new technologies. This session will discuss some of the

technologies for global quantitative analysis of proteins from complex mixtures, and high throughput analysis of protein interactions. These new technologies include protein chips, microarrays, 2D gels, image and data analysis systems.

Cochairs: **Ruth Van Bogelen, Ph.D.**, Head of Genomics & Proteomics, Pfizer Global Research and Development, Ann Arbor, MI
Norman G. Anderson, Ph.D., Chief Scientist, Large Scale Biology, Rockville, MD

3) ***Protein Separation and Identification***

New uses are being developed for old technologies such as mass spectrometry and gel electrophoresis. These technologies provide the tools scientists use to identify proteins and multi-protein complexes. This session will address the use of mass spectrometry and related techniques to characterize proteins and protein-protein interactions, structure and folding.

Cochairs: **Alma Burlingame, Ph.D.,** University of California-San Francisco
Julio Celis, Ph.D., Institute of Cancer Biology and Danish Centre for Human Genome Research, Danish Cancer Society, Copenhagen, Denmark.
Alain Van Dorsselaer, Ph.D., Pasteur University, Strasbourg, France

4) ***Protein Structure and Function***

With the goal of creating an atomic description of each and every constituent of the cell, worldwide, large-scale structural initiatives are beginning to deliver what, over the next decade, will be an enormous wave of structural information to the shores of the biological community. Achieving this goal requires that many new and formidable challenges must be met. While the structural initiatives are underway, the functional/enzymological programs that will articulate the functions of these structures are, as yet, in the concept stage, and have recently been explored in workshops at the NIH/NIGMS. We invite you to help define the important issues that genomic-scale science has created for the structural and functional communities.

Cochairs: **Greg Petsko, Ph.D.**, Brandeis University, Waltham, MA
Thomas Leyh, Ph.D., The Albert Einstein College of Medicine, New York, NY

5) ***Metabolic Pathways and Post - Translational Modifications***

Proteomics research presents a much more elusive task than the mapping of the human genome. Protein modification during protein translation and various metabolic processes, creates a challenge for defining the mandate of proteomics research. This session will address how scientists might proceed in annotating proteins, while considering how protein modification and the metabolic products will affect function and proteomics research.

Cochairs: **Edward Dennis, Ph.D.,** University of California, San Diego
Eugene Bruce, Ph.D., Division of Integrative Biology and Neuroscience, National Science Foundation, Arlington, VA

6) ***Implementation: Necessary Policy and Infrastructure Conditions for Collaboration***

This session will address the unique/critical needs of proteomics research with respect to collaboration including education, funding, international cooperation, data sharing policies, and informatics infrastructure (e.g. software standards, scientific portals, collaboratories, computing grids).

Cochairs: **Jim Myers, Ph.D.,** Computational Science and Mathematics Department, Pacific Northwest National Laboratory, Richland, WA
Richard W. Morris, Ph.D., Division of Allergy, Immunology & Transplantation, National Institute of Allergy and Infectious Disease, NIH, Bethesda, MD

7) ***Clinical Aspects***

Proteomics research promises to help fundamentally change the practice of medicine in the 21st century. This session will focus on how clinical proteomics research can be used to define new molecular markers for risk assessment and disease diagnosis, and how it can be used with other molecular profiling techniques to identify new targets for pharmaceutical development. The use of protein chips in the drug development and clinical settings will also be discussed.

Cochairs: **Alan Sachs M.D,** University of California Berkeley; Director, Clinical Genomic Pharmacology, Merck Research Labs, Inc
Denis Hochstrasser, M.D., President of Clinical Medicine, University of Geneva; Director of the Department of Pathology and Head of the Central Clinical Chemistry Laboratory, Geneva University Hospital