reviews Journal of proteome research

Systems Biology, Proteomics, and the Future of Health Care: Toward Predictive, Preventative, and Personalized Medicine

Andrea D. Weston[†,‡] and Leroy Hood[*,†]

Institute for Systems Biology, 1441 North 34th Street, Seattle, Washington, 98103-8904

Received January 17, 2004

The emergence of systems biology is bringing forth a new set of challenges for advancing science and technology. Defining ways of studying biological systems on a global level, integrating large and disparate data types, and dealing with the infrastructural changes necessary to carry out systems biology, are just a few of the extraordinary tasks of this growing discipline. Despite these challenges, the impact of systems biology will be far-reaching, and significant progress has already been made. Moving forward, the issue of how to use systems biology to improve the health of individuals must be a priority. It is becoming increasingly apparent that the field of systems biology and one of its important disciplines, proteomics, will have a major role in creating a predictive, preventative, and personalized approach to medicine. In this review, we define systems biology, discuss the current capabilities of proteomics and highlight some of the necessary milestones for moving systems biology and proteomics into mainstream health care.

Keywords: proteomics • health care • nanotechnology • systems biology • microfluidics • predictive and preventative medicine

Introduction: Paradigm Changes in Health Care

As systems biology emerges as a discipline, it is becoming increasingly clear that it will catalyze fundamental changes in the future of health care. We predict that a paradigm shift in medicine will take place within the next two decades replacing the current approach, which is predominantly reactive, to one that can increasingly predict and prevent cellular dysfunction and disease. Within the next 10−15 years, a predictive medicine will emerge, capable of determining a probabilistic, individualized future health history. Over this time span, we will be able to sequence a human genome for less than $1000 in a fraction of an hour. Accordingly, we will have the ability to examine the variants (polymorphisms) of the 30 000 or so genes for each individual and make probabilistic statements about their disease likelihood.

Since environmental signals can greatly influence the onset of disease, it will be equally important to be able to measure the consequences of these pathogenic signals. We suggest that distinguishing health from disease is possible with the development of devices which enable multiparameter analyses of the blood. We can envision, for example, having a handheld microfluidics device capable of making thousands of quantitative protein or mRNA measurements that will permit one to detect emerging genetic mutations (as one sees with cancer) or pathogenic environmental stimuli such as infectious agents. These will be seen as a consequence of pathogenic perturbations of common protein and gene regulatory networks. Thus, predictive medicine will involve analyzing the individual genome for disease-susceptibilities and following pathogenic environmental exposures by multiparameter blood analyses. Given that individuals differ from one another by approximately 6 million DNA polymorphisms, each of us will be predisposed to differing combinations of disease. The environmental signals to which we are all exposed will also vary greatly. Accordingly, personalized medicine will be necessary for predicting disease.

Efforts toward creating such predictive approaches are ineffective if not accompanied by the development of suitable methods for preventing disease. This will hinge on our capabilities for characterizing biological systems in their normal states, and defining the molecular basis for pathology of these systems—tasks that will require an integrative, systems biology approach. As the pharmaceutical industry struggles to identify new drugs cost-effectively, we suggest that systems biology will play an essential role in the drug discovery process in the future. Here, we describe the importance of systems biology in health care, and highlight the increasing role of proteomics in predicting and preventing the onset of disease. In addition, a main objective of this review is to outline the myriad of challenges for integrating proteomics into contemporary medicine. These include, but are not limited to, technology (hardware and software) development, as well as challenging social, ethical, and legal issues.

What is Systems Biology?

Systems biology is the analysis of the relationships among the elements in a system in response to genetic or environ-

* To whom correspondence should be addressed. Tel: (206) 732-1201. Fax: (206) 732−1299. E-mail: lhood@systemsbiology.org.

† Institute for Systems Biology.

‡ Present Address: Pfizer Global Research & Development, Groton, Connecticut, 06340.

10.1021/pr0499693 CCC: $27.50

Published on Web 03/04/2004

mental perturbations, with the goal of understanding the system or the emergent properties of the system. A system may be a few protein molecules carrying out a particular task such as galactose metabolism (termed a biomodule), a complex set of proteins and other molecules working together as a molecular machine such as the ribosome, a network of proteins operating together to carry out an important cellular function such as giving the cell shape (protein network), or a cell or group of cells carrying out particular phenotypic functions. Thus, a biological system may encompass molecules, cells, organs, individuals, or even ecosystems.

Systems biology recently emerged as the result of five key advances: (1) The human genome project provided a genetics parts list of all the human genes and cis-control elements (genes are relatively easy to identify; cis-control elements are not). This project was the first large-scale discovery project in that it sought to sequence or "discover" the entire genome. This led naturally to discovery projects for individual organisms or cell types to quantitatively identify all mRNAs (the transcriptome) or all proteins (the proteome). Generating such parts lists is an important element of systems biology. (2) Cross-disciplinary biology has emerged, creating environments where biologists, chemists, computer scientists, engineers, mathematicians, and physicists all work together to develop new global technologies, integrative computational software, and mathematics and apply these to biology. (3) The Internet has given us the means for acquiring and disseminating large global data sets for genomes, RNAs, proteins, interactions and phenotypes. (4) The idea that biology is an informational science has played a key role in the emergence of systems biology.[1] In this regard, four points are important. First, biological information is of two distinct types—the digital information of the genome and the environmental cues that come from outside the genome—together they are responsible for the development of organisms as well as their physiological responses. Second, the digital information falls into two major categories: the genes and the cis-control elements which specify the behavior of the genes (when in time and space and to what amplitude they are expressed). The cis-elements, together with their cognate transcription factors, specify the architecture and linkage relationships of the gene regulatory networks that coordinate the behavior of groups or batteries of genes that govern the development of organisms and their physiological responses. One of the major challenges of systems biology is to determine the architecture of protein and gene regulatory networks and to understand how their behaviors are integrated to carry out biological functions. Third, biological information is hierarchical as one moves outward from the genome to ecologies (DNA → RNA → protein → biomodules or networks → cells → organs → individuals → populations of individuals → ecologies). The important point is that environmental signals change the biological information at each level of the hierarch; thus, to do systems biology, as many levels of information as possible must be gathered and integrated. Forth, biology is dynamic. Therefore, whether it is development or a physiological response, systems biology must gather data across the dynamics of the response if it is to be understood. (5) The fifth key advance that led to the emergence of systems biology was the development of high-throughput platforms for genomics, proteomics and metabolomics, which made possible the gathering of global data sets. These global data sets are an essential feature of systems biology. The development of high-speed DNA sequencers, DNA arrays, rapid methods for genotyping and a variety of improvements in proteomics, particularly in the area of mass spectrometry, has enabled systems biology.

In summary, systems biology is hypothesis-driven, in that systems approaches always begin with a model (descriptive, graphical or mathematical) and the model is tested with hypotheses that require systems perturbations and the gathering of dynamic global data sets. Different data types are integrated and compared against the model. At each turn of the hypothesis-driven process, the model is reformulated. This process is continued until the experimental data and the model are brought into juxtaposition.[2]

Systems Biology in Medicine

The value of systems biology in medicine will manifest itself in at least two major forms. First, systems biology will continually improve our capacity to understand and model biological systems on a more global and in-depth scale than ever before. This in itself is proving to be a daunting, but remarkably fruitful challenge. As researchers continue to gather new systems-level insights, an equally demanding task will be to apply this new knowledge in medicine in as timely a manner as possible. The second major impact of systems biology in medicine will be the continual spawning of new technologies, which will enhance the efficiency, scale and precision with which cellular measurements are made. This latter influence will facilitate all aspects of health care, including the detection and monitoring of diseases, drug discovery, treatment evaluation, and ultimately, predictive and preventative medicine. Moreover, as technologies mature, they will accommodate smaller sample volumes and will be more economical, in turn supporting personalized medicine. Many of these changes will come in the form of microfluidics and nanotechnology—both of which will transform most, if not all, analytical techniques in biology and medicine.[3]

A description of some of the emerging proteomics technologies applicable to health care and our predictions on which new technologies will revolutionize medicine, are discussed later in this review. First, it is worth emphasizing the importance of basic research using systems biology to understand both normal biological systems and pathological states. The ability to predict and prevent disease will always be dictated by our fundamental knowledge of the normal and diseased state of cells. Treating disease will require circumventing the limitations of specific genetic or protein defects. To do this, these defects, which include genetic mutations, inappropriate protein processing or folding, aberrant protein−protein or protein-DNA interactions, and protein mislocalizations must first be accurately placed within the context of disease. The best way to link these deficiencies to their respective diseases is to gain a comprehensive knowledge of the normal biological systems involved. Certainly, this is becoming increasingly possible with the continual development and improvement of new technologies that can profile global cellular changes in healthy and diseased cells.

Importance of Elucidating Cellular Networks

Understanding the root causes of complex diseases such as cancer is essential for developing the most effective detection methods and for defining the most appropriate treatment (and ultimately preventive) strategies. The detection of some of these diseases has been greatly facilitated by the identification of diagnostic biomarkers, but until very recently, this approach

focused largely on single molecules. In addition, a number of cancer therapies are targeted toward a specific molecule or signaling pathway, to inhibit tumor growth. These approaches reflect the traditional scientific approach of reducing cellular processes to their individual components and/or signal transduction pathways. However, the behaviors of most biological systems, including those affected in cancer, cannot be attributed to a single molecule or pathway, rather they emerge as a result of interactions at multiple levels, and among many cellular components.

Groups of interacting molecules that serve a specific function make up biomodules whose interconnections give rise to networks. Understanding the design principles of biomodules and protein and gene regulatory networks during normal physiology and disease will lead to more rationalized and efficacious treatment strategies, as the actual nodal points or direct underlying causes of diseases will be pinpointed. More straightforwardly, drugs and other therapies can be better directed at re-engineering the behaviors of malfunctioning networks. This may mean, for example, modifying the activity of a transcription factor to prevent abnormal expression of a whole subset of genes, as opposed to inhibiting only one or a few of the molecules that act downstream of that transcription factor. Such approaches rely on characterizing cellular modules and networks using systems biology. Here we outline the application of systems biology to two model systems: galactose utilization in yeast and endomesoderm specification in the sea urchin, with the purpose of highlighting the power of systems biology for gaining new insights into network systems.

The Power of Systems Biology: Galactose Utilization in Yeast

The systems biology approach has provided a wealth of new information even for the relatively simple system whereby yeast utilize galactose as a carbon source—a system that has been intensely studied for decades and which represents one of the best-characterized systems of gene regulation. Cells use galactose as a primary energy source by employing a series of enzymes that convert galactose to glucose-6-phosphate, as well as a regulatory network that controls expression of all of the genes required by this process (for review see refs 4 and 5). Galactose is transported into the cell via a permease (Gal2p), then converted to glucose-6-phosphate by the enzymes galactokinase (Gal1p), uridylyltransferase (Gal7p), epimerase (Gal10p), and phosphoglucomutase (Gal5p/Pgm2p). Tight transcriptional control of these enzymes and the permease is achieved primarily through the actions of three regulatory proteins, Gal3p, Gal4p, and Gal80p, which, to a certain extent, also regulate their own expression.

Until recently, many have regarded galactose utilization as a simple regulatory network. Gal4p, a transcription factor, binds to specific sequences upstream of the *GAL* genes to potently activate transcription, whereas Gal80p, a co-repressor protein, binds to and inhibits Gal4p in the presence of glucose, but is removed and tethered to Gal3p under galactose-inducing conditions (reviewed in refs 4,5). Further studies, however, have established additional regulatory roles for events such as the transport of Gal80p from the nucleus to the cytoplasm,[6,7] the phosphorylation of Gal4p,[8,9] and the recruitment of chromatin remodeling complexes such as SAGA to the *GAL* promoters.[10,11] All of these events take place during galactose induction. Despite these additional insights, however, it was not until the galactose system was interrogated using a large-scale, systems biology approach, that the complexity of this system and its interconnections with other cellular functions became apparent.[12]

Soon after the emergence of global technologies for studying cellular processes, Ideker et al., (2001) in an initial attempt to examine the feasibility of systems biology, focused on the galactose utilization pathway as a benchmark system. This study employed nine strains of yeast, each with a different galactose gene knocked out, and the wild-type, and monitored changes in the levels of ~6200 yeast genes using DNA arrays with the system on (in the presence of galactose) and off (in the absence of galactose) for each of the genetic perturbations (knockouts).[12] Nine hundred and ninety-seven mRNAs were changed in one or more of these perturbations. In addition, the quantitative changes in protein expression for 300 proteins of the wild-type yeast with the system on and off were determined using the ICAT approach (a method for quantifying changes in patterns of protein expression for different cellular states—see below). Finally, expression changes of the 997 mRNAs were integrated with all known protein−protein and protein-DNA interactions, and the information was displayed in a physical interaction map, which displays the interrelationships between different functional modules within the cell (Figure 1).

The systems biology study of galactose utilization provided a number of new insights. First, this was the earliest study to report, on a global level, a poor correlation between changes in mRNA levels and changes in protein expression. This suggests that posttranscriptional regulatory mechanisms are important for changing patterns of protein expression. Second, it was demonstrated, unequivocally, that although the galactose pathway itself involves a well-characterized transcriptional network controlling the genes required for galactose utilization (Figure 1B), the cellular response to galactose extends well beyond the activation of these genes. The global nature of the cellular response to galactose suggests that a network of biomodular interactions exists in the cell, and that many different biomodules are affected during galactose induction. A good example for this is the change in mRNA expression patterns observed for a number of genes important for amino acid biosynthesis, many of which are known to be regulated by the transcription factor, Gcn4p (Figure 1C). Finally, results from this study also suggested that the accumulation of galactose-1-phosphate causes a reduction in the expression of some *GAL* genes. These findings underscore the regulatory influence of metabolites, adding another layer of complexity to the system, and emphasizing the need to collect and integrate data at many levels.

Since the publication of studies by Ideker et al., we have continued to use systems biology approaches to understand the gene regulatory networks underlying the galactose response. Our ability to efficiently identify the elements (proteins and cis-elements) and interactions (protein−protein and protein−DNA) that characterize gene regulatory networks has advanced substantially over the past few years. For example, a technology known as genome-wide binding analysis was established to identify, on a genome-wide scale, the DNA targets of any given regulatory protein.[13] This genome-wide location analysis combines a modified chromatin immunoprecipitation method with microarray analysis using microarrays containing all yeast intergenic sequences. From this method, the critical linkages between cis-control elements and their cognate transcription factors in gene regulatory networks

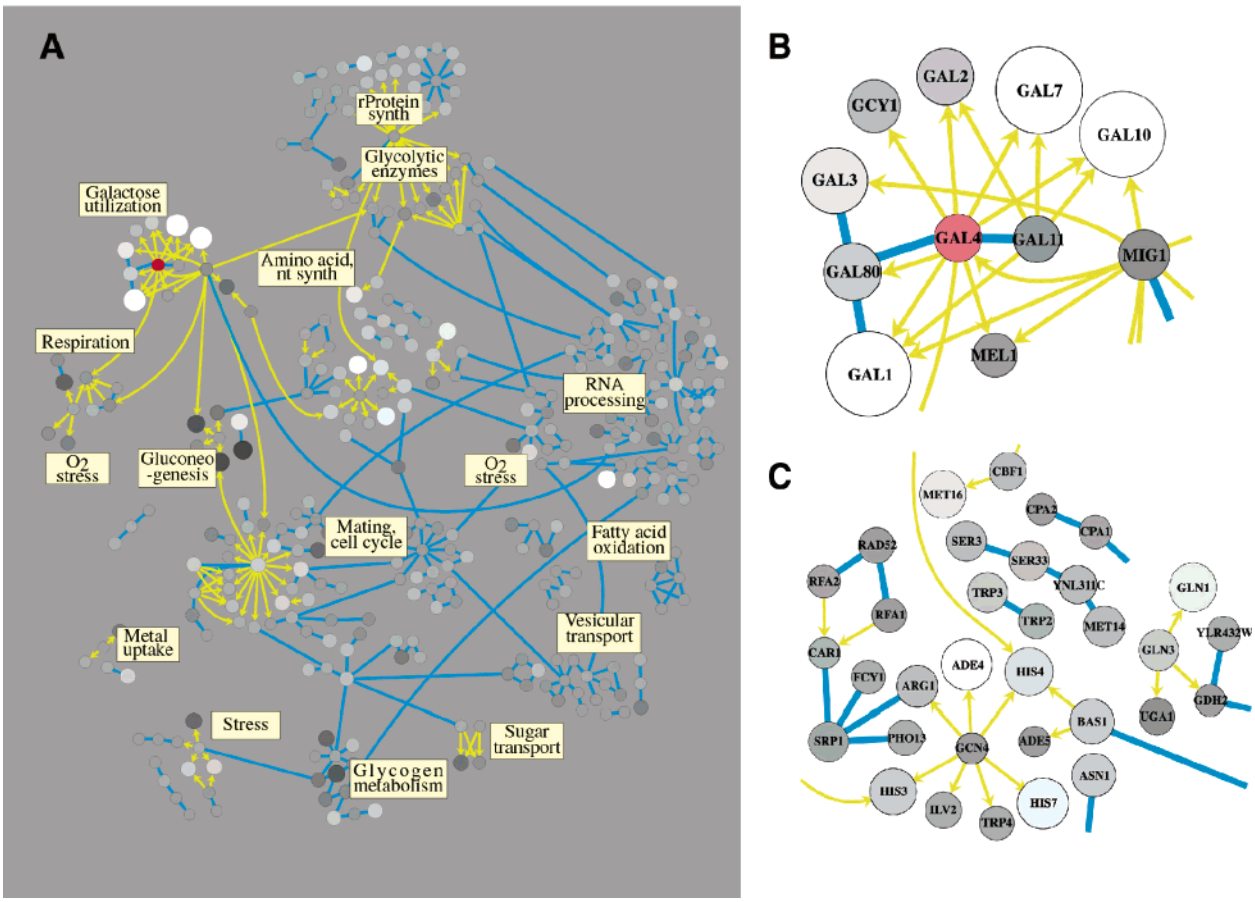

Figure 1. Integrated physical interaction network. (A) A network of protein−protein and protein-DNA interactions is displayed with the changes in gene expression (caused by Δ*gal4*+galactose perturbation) superimposed on the network. Each gene is represented as a node, protein−DNA interactions are indicated by a yellow arrow and protein−protein interactions are represented by blue lines between nodes. The changes in gene expression are represented using a gray scale for each node (white represents a decrease in expression and black denotes an increase), and Gal4 itself is red. Highly interconnected groups of genes tend to have common biological functions and are labeled accordingly. Regions corresponding to the galactose utilization module are shown in (B) and those corresponding to the amino acid synthesis biomodule are shown in (C). Reprinted with permission from Ideker et al. [12] Copyright 2001, American Association for the Advancement of Science.

can be ascertained. Another advance for studying regulatory networks in yeast was the recent release of high-quality draft sequences for three species of yeast that are related to *S. cerevisiae* (*S. paradoxus*, *S. mikatae*, and *S. bayanus*).[14] The availability of these sequences opens the door for comparative genomics, a particularly useful approach for the identification of regulatory motifs.

The information from microarray studies, genome-wide location analyses, and comparative studies of different sequences can all be integrated using computational approaches to generate accurate models of gene modules in which the targets of a transcription factor are defined, as are the cis-elements to which these factors bind. We have used such an approach to identify previously unknown targets for Gal4p. These include: *MTH1*, *PCL10*, *FUR4*, (also identified by ref 13), as well as *NAR1*, *YPL066w*, *YEL057c*, and *YPS3*. Expression of all of these genes is altered in response to galactose, and they were all shown to be direct targets of Gal4p using the genome-wide location analysis protocol. Moreover, each of these genes have, within their promoters, the Gal4p binding site which is conserved across all four species of yeast for which sequence information is available. The identification of these targets for Gal4p reveals potentially new functions for this transcription factor. *MTH1* encodes a repressor of the hexose transport (*HXT*) genes (incidentally, a number of *HXT*s are repressed in response to galactose), *PCL10* encodes a cyclin-dependent protein kinase important for glycogen biosynthesis, *FUR4* encodes a uracil permease, and *YPS3* encodes an aspartic-type endopeptidase. The functions of *NAR1*, *YPL066w*, and *YEL057c* have not yet been characterized. The additional insights into Gal4p targets resulted from the integration of four key approaches: microarray expression analysis, genome-wide binding analysis, the use of search algorithms on a defined list of sequences, and comparative genomics.

The integrative approach for defining Gal4p targets represents just one example of how we have combined the information from multiple, high-throughput studies, to derive useful information. We chose this example to emphasize the benefits of data integration and to demonstrate the power of discovery science. Each large data set on its own contains sufficient noise as to preclude a similar identification of Gal4p targets, whereas combining the information to consider only those genes that satisfy more than one condition, provides a filter for noise, and attaches confidence to experimental findings. It is interesting to note that the methods used to generate a list of Gal4p targets are all discovery-based approaches, including the sequencing of additional yeast genomes. With the resultant list of targets,

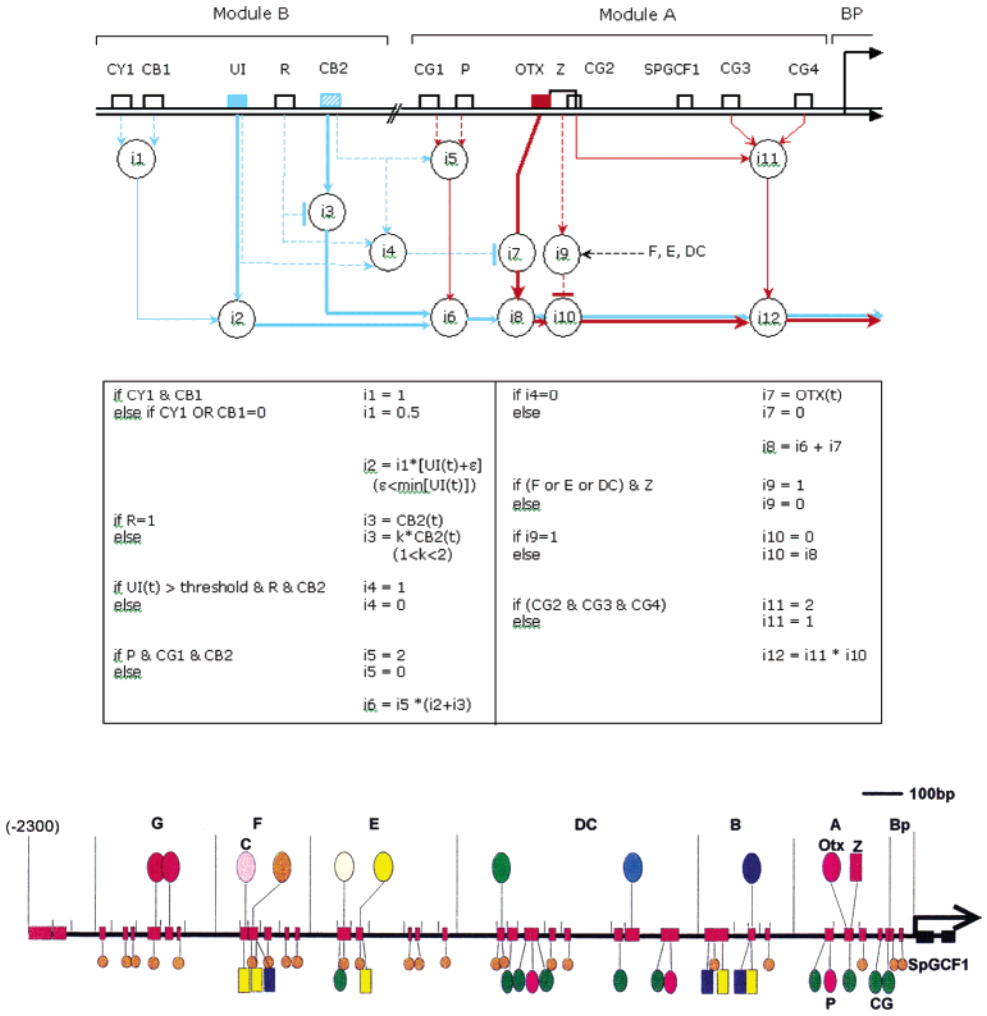

Figure 2. Cis-regulatory network controlling *endo 16* expression in sea urchin. The protein−DNA interactions within the 2300 bp *endo 16* cis-regulatory region are shown (bottom), with each protein denoted by a different color. The letters above represent the different modules controlling different expression features of *endo 16*. The control logic model for modules A and B are shown above the cis-regulatory region. Boxes above the line indicate protein binding sites, and circles below the line indicate logical operations. Influences by module A are shown in red and those of Module B are shown in blue. Dashed lines represent interactions that can be modeled as boolean inputs, thin solid lines represent scalars, and thick lines represent time-dependent quantitative inputs. Arrowheads indicate a positive input, perpendicular bars indicate a negative input. Statements below the model are defined as individual logic interactions. Reprinted with permission from Davidson.[78] Copyright 2001, Academic Press.

we can now generate systems-level hypotheses to test these connections and to further understand the widespread function of Gal4p.

The Power of Systems Biology: Endomesoderm Specification in Sea Urchin

The analysis of the galactose utilization system in yeast displays a systems approach to understanding a simple physiological response. The studies carried out by Eric Davidson and colleagues, to understand endomesoderm specification in sea urchin larva, demonstrate the power of a systems approach to understanding developmental processes. The sea urchin is an excellent model for studying development because it exhibits a simple mode of development and the fertilized egg divides within a day or so into the five territories that constitute the major tissues in the larva. The larva emerges within just 72 h of fertilization. Davidson and co-workers have extensively analyzed the regulatory gene network underlying endomesodermal specification in sea urchin embryos.[15] In one approach, they focused on the cis-regulatory system of the developmentally regulated *endo 16* gene—a marker of endoderm cell fate specification. First, they defined the cis-control elements for *endo 16* whose expression is specific to the endomesoderm. This gene has 34 cis-control elements and is regulated by 17 transcription factors (Figure 2). The complicated temporal and spatial expression pattern of *endo 16* during development is recreated by expression of a reporter linked just to the regulatory region of 2.3 kilobases (just 5′ to the *endo* 16 gene). Indeed, the regulatory region behaves as if it were a modular computer chip—integrating the environmental signals reflected in the changing concentrations of various cognate transcription factors across larval development (Figure 2). Moreover, a logic diagram can be constructed to explain relatively completely the expression patterns of the *endo 16* gene across the 72 h of development.[16]

In addition, Davidson and colleagues constructed a gene regulatory network for endomesodermal development in the larva (Figure 3). About six different genetic and environmental

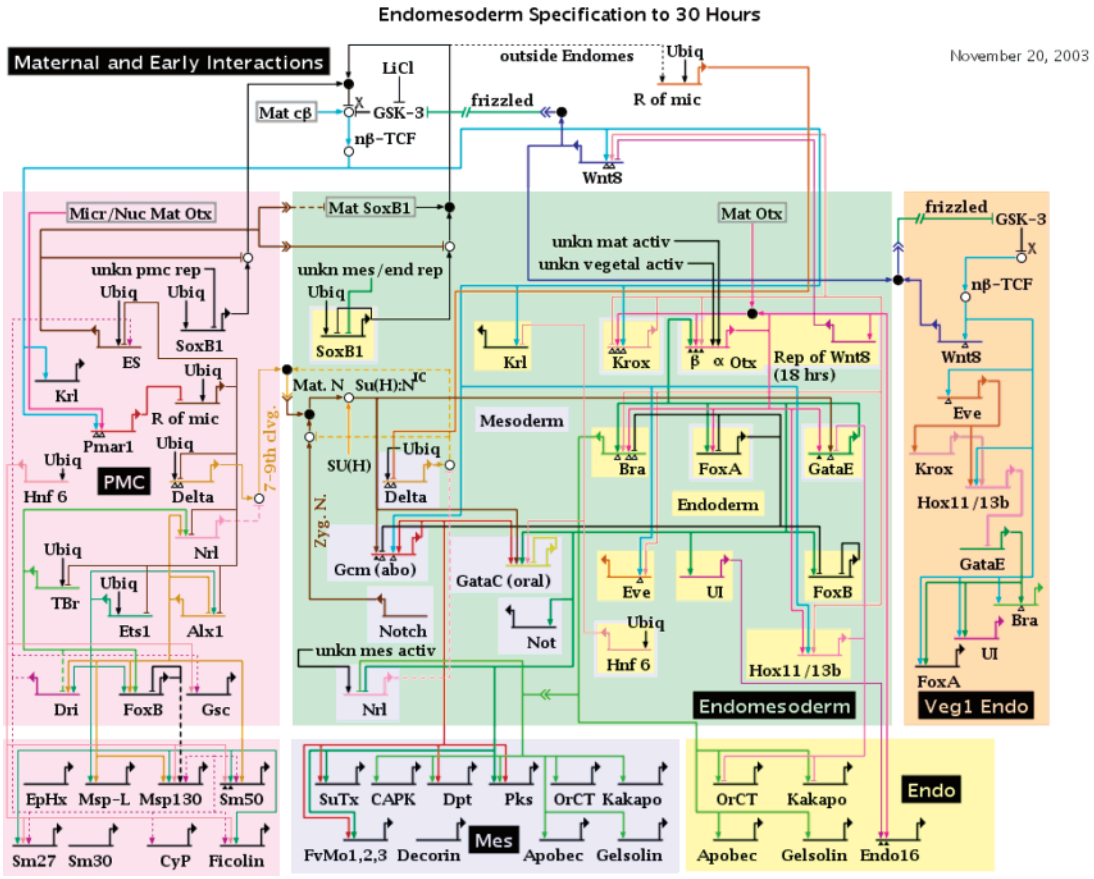

Figure 3. Regulatory gene network model for endomesoderm specification in the sea urchin. The current form of a provisional regulatory gene network is shown. This updated network (http://sugp.caltech.edu/endomes/) is based on an ongoing analysis of the gene regulatory mechanisms controlling endomesodermal specification in the sea urchin embryo.[15,77] Extensive experimental evidence on endomesoderm specification is depicted. Much of the network architecture is based on perturbation and expression data and on cis-regulatory studies of several genes. Bent arrows indicate transcription, and each short horizontal line from which these arrows extend represents the cis-regulatory element controlling expression of that gene (gene name is below the line). The arrows and barred lines indicate activation and repression, respectively, of the downstream genes. Triangles indicate known or putative cis-regulatory elements. Details on this network and the underlying data can be found at http://www.its.caltech.edu/~mirsky/. Reprinted with permission from Davidson and Bolouri (http://sugp.caltech.edu/endomes/). Copyright 2001−2003, Hamid Bolouri and Eric Davidson.

perturbations were employed to construct this network,[15] which now contains about 60 genes of which 50 are transcription factors. Several important conclusions can been drawn from an analysis of this network. First, there appear to be a variety of subcircuits similar to those found in engineering (feed-forward, feed-backward, positive feed back loops, negative feedback loops, etc). It is worth noting that similar subcircuits, underlying transcriptional regulation in *Escherichia coli*, were described by Shen-Orr et al., [17,18] and by Lee et al., in their global analysis of transcription factor binding sites in yeast [19]. Thus the hope is that a lexicon of these subcircuits can be determined and used to analyze the higher order functioning of these networks, and possibly to design new networks or redesign old ones. Second, the network is designed to move development forward inexorably, in keeping with the fact that development is, under most conditions, irreversible. Finally, a careful examination of the network suggests perturbations that may change fundamental emergent properties of the system. Indeed, one such perturbation has been carried out to generate a larva with two guts. The important point is that in preventive medicine the problem will essentially be identifying protein and gene regulatory networks and changing their behaviors with drugs. Thus, these model systems are providing fundamental new strategies for thinking about drug and drug target discovery.

Proteomics and Systems Biology

Proteomics has recently come to the forefront as an area that promises to transform biology and medicine. As informative as DNA microarray expression studies are, it is becoming increasingly apparent that changes in mRNA expression often correlate poorly with changes in protein expression.[12,20−22] Proteins also present several striking analytic challenges. First, proteins can be expressed across enormous dynamic ranges—l in $l0^6$ in cells and perhaps greater than l in 10^9 in blood. We do not yet know how to measure proteins across these dynamic ranges, especially because there is no protein equivalent of the polymerase chain reaction (PCR). Second, proteins change enormously in patterns of expression across developmental and physiological responses. The dynamic nature of these changes requires a way of taking global snapshots of patterns of protein expression that does not now exist. Finally, proteins may be altered by many environmental perturbations—each changing their information content. Protein function can be regulated by a number of post-transcriptional events, further limiting the inferences one can make based on mRNA expression studies

alone. In short, proteins are the actual effectors driving cell behavior, and they cannot be studied simply by looking at the genes or mRNAs that encode them, thus warranting the establishment of a field, now termed proteomics, devoted entirely to their study.

The goal of proteomics research is to understand the expression and function of proteins on a global level. More than simply cataloguing the proteome—a quantitative assessment of the full complement of proteins within a cell—the field of proteomics strives to characterize protein structure and function, protein−protein, protein-nucleic acid, protein−lipid, and enzyme−substrate interactions, post-translational modifications, protein processing and folding, protein activation, cellular and sub-cellular localization, protein turnover and synthesis rates, and even alternative isoforms caused by differential splicing and promoter usage. In addition, the ability to capture and compare all of this information between two cellular states is essential for understanding cellular responses. Achieving the goals of proteomics is no small feat, but advances are certainly improving the rate at which proteins can be characterized and their functions determined. Adding to the complexity of this field is the need to integrate proteomics data with other information such as gene, mRNA and metabolite profiles, to fully understand how systems work.

Current Proteomic Technologies

Proteomics has gained steady momentum over the past five years, with the development of several approaches, some of which are new and others that build upon traditional methods. Mass spectrometry-based methods and protein microarrays are the most common technologies currently being used for the large-scale study of proteins. Here, we highlight some of the new approaches in these areas that are already proving to be applicable in the clinic, and which will largely impact both basic research and medicine.

Quantitative Protein Profiling and the Impact of Mass Spectrometry

Much effort has been spent on the development of quantitative methods of protein profiling. There are currently two mass spectrometry-based approaches in use for global quantitative protein profiling. The more established and most widespread method uses high-resolution, two-dimensional electrophoresis (2DE) to separate proteins from two different samples in parallel, followed by staining and selection of differentially expressed proteins to be identified by mass spectrometry. This proteomics approach has evolved over the years, with improvements in 2DE separation, and protein detection, and indeed, a number of advances have led to increased reproducibility of proteome patterns between different laboratories. Despite the advances in 2DE and its maturity, which have made 2DE a reliable method of choice for many, there are limitations. The major concern is an inability to resolve all the proteins within a sample, given their exceptional range in expression level and differing properties. Although the extent of these limitations has been debated,[23,24] it can be agreed upon that alternative methods are needed to study protein expression in a more comprehensive and high-throughput manner.

A second quantitative approach, which is gaining in popularity, uses stable isotope tags to differentially label proteins from two different complex mixtures. In this method, proteins within a complex mixture are first labeled isotopically then digested to yield labeled peptides. The two differentially labeled peptide mixtures are then combined, peptides separated by multidimensional liquid chromatography (LC) and analyzed by tandem mass spectrometry. Peptides are identified by automated database searches, and relative protein abundances are obtained from the mass spectra. Isotope-coded affinity tag (ICAT) reagents are the most commonly used isotope tags. The ICAT reagent consists of three main components: (1) a reactive group with specificity for cysteines; (2) a linker labeled with either heavy hydrogen (d0) or light hydrogen (d8) isotopes; and (3) an affinity tag (biotin) for the solid-phase capture and isolation of labeled petides. In this method, cysteine residues of proteins are covalently attached to the ICAT reagent, reducing the complexity of the mixture significantly by omitting analysis of all noncysteine-containing peptides. These peptides do not necessarily have to be excluded, however, as alternative chemistries have been developed to label other amino acid residues.[25−27]

Quantitative proteomics using stable isotope tagging is becoming an increasingly useful tool, as several improvements in isotope tagging, mass spectrometry, and data analysis have been made (reviewed in ref 28). Following its initial description 5 years ago, the ICAT method has been applied to a number of problems previously less tractable with existing technologies. First, alternative chemistries have been developed to label other (noncysteine) amino acid residues,[25−27] enabling different coverage or (if used in combination with the ICAT approach, more complete coverage of) the proteome. Chemical reactions have also been used to introduce tags into specific sites of peptides or proteins, for the purpose of probing specific functionalities of proteins. For instance, the isolation of phosphorylated peptides has been achieved using isotopic labeling and selective chemistries to selectively capture this fraction of proteins among a complex mixture.[29−33] Reversible phosphorylation of proteins has been known for some time to control many biological functions, and this and other post-translational modfications have been widely implicated in disease. The ability to compare relative abundances of specific post-translational modifications between healthy and diseased cells, will therefore profoundly impact the study, diagnosis, and treatment of those diseases.

The ICAT technology was recently used to differentiate between the protein composition of purified, or partially purified, macromolecular complexes such as the large RNA polymerase II (Pol II) preinitiation complex, and the proteins complexed with the yeast transcription factor, Ste12.[34] In addition, ICAT labeling was recently combined with chromatin isolation to identify and quantify chromatin-associated proteins.[35] These latter techniques extend the use of the ICAT technology to the analysis of protein−protein and protein−DNA interactions as they pertain to transcriptional regulatory networks, which are integral to cellular functions. Finally, ICAT reagents are useful for proteomic profiling of cellular organelles and specific cellular fractions. For instance, the abundance ratios for almost 500 proteins associated with the microsomal fractions were obtained for naive and in vitro-differentiated human myeloid leukemia (HL-60) cells.[36] Essentially, the marriage of stable isotope tagging and mass spectrometry enables researchers to profile changes in: protein expression, protein−protein or protein−DNA interactions, post-translational modifications, and the constituents of cellular fractions or organelles. With such versatility, and ongoing improvements in the preci-

sion and throughput, this proteomics platform will have a lead role in transforming biology and medicine.

Diagnostic Protein Biomarkers

Single molecule biomarkers have been useful for the diagnosis of certain diseases, for treatment monitoring, and for the evaluation of new drug candidates. An example is prostate specific antigen (PSA), which is elevated in men with prostate cancer. Unfortunately, despite their usefulness, there has been a decline in FDA-approved diagnostic biomarkers over the past decade, and there are multiple diseases for which no useful indicators have been identified. Consequently, a large number of diseases are detected at an advanced stage, limiting prognosis, and treatment options. The identification of biomarkers is an area in which proteomics will undoubtedly have a significant impact—a prospect that has not gone unnoticed by the proteomics community.

As researchers seek to find new biomarkers for various diseases, there are two growing concerns. First, of the biomarkers routinely used to diagnose disease, most are capable of detecting the onset or advanced progression of disease, but have little, if any, predictive power. Part of the solution to this problem will be to collect and test samples retrospectively. Since researchers cannot go back and test patient samples prior to the onset of their disease, this will require the establishment of large, long-term studies, in which individual samples are collected over the span of at least a decade. It will be possible, with the necessary technological advances, to identify the early signs of disease or even the markers of predisposition. It is reasonable to envision, for instance, that the most efficient biomarkers for predicting cancer are not tumor-derived, but are those that are indicative of a microenvironment that typifies the pre-tumor state. If these markers are identified, then individuals could be warned of their condition prior to the development of any damaging tumors. If retrospective samples were available for groups of individuals diagnosed with cancer, then these biomarkers could be identified. In addition to retrospective sampling, identifying biomarkers in general will be expedited with a more detailed understanding of the human plasma proteome (see below).

The second concern with respect to the use of single molecule biomarkers is that it is based on the expectation that an increase in the concentration of a single protein can unambiguously specify disease—a dangerous and unrealistic assumption. Diseases are characterized by heterogeneity between individuals; the same disease can be initiated by numerous factors and can cause a range of molecular changes. Ideally, biomarker assays should be used alongside individualized DNA sequencing, DNA microarray analyses, metabolite studies, etc. Thus, the key to diagnostics in the future will be multiparameter analyses—the measurement of 10s, 100s, 1000s, or even 10 000s of mRNA, protein, or small molecule components in the blood. Just as normal physiology and disease arise from protein and gene regulatory networks, normal and perturbed, and these require analyses of all the elements in the system, diagnostics will also require the analysis of multicomponents to reflect the true complexity of the disease process. Moreover, multiparameter analyses will be able to (1) predict the onset of disease, (2) stratify disease (e.g., prostate cancer is probably three or four diseases and not just a single one), (3) indicate the progression of the disease, (4) follow the course of treatment, and (5) make predictions about the effectiveness of a drug or adverse reactions, etc. By this view, multiparameter analyses of the serum or blood will provide a window into health and disease. The term biomarker will come to mean 10s, 100s, or even 1000s of informative markers identified in the context of a much larger sample of measurements. For each type of disease, stratification, degree of progression, etc. the informative set of markers will be different.

In addition, for multiparameter analyses of the blood (or other body fluids) to serve as a window to distinguish health from disease, it is necessary to correlate the multiparameter data sets from large numbers of individuals with different physiological and disease states. This is classical discovery science. The challenge will be to organize these efforts so that the global analytic capacities of systems biologists are applied to the 1000s or 10 000s of clinical samples from physicians that are necessary for these correlations. In a subsequent step, systems biology can begin to explain the differences in the multiparameter sets in terms of perturbations of protein and gene regulatory networks. These explanations may provide important insights into how to approach disease prevention or therapy.

New Promise for Biomarkers: Serum Proteome Pattern Diagnostics

An exciting approach has emerged which replaces single molecule discovery efforts with serum proteomic pattern diagnostics. The concept behind pattern diagnostics is that the blood plasma proteome reflects tissue and organ pathology, causing patterns of protein changes that have diagnostic potential without even knowing the identities of the individual proteins. Since MS-based approaches provide a pattern of peaks, the idea is that these patterns can discriminate certain diseases. The diagnostic becomes the pattern or "signature" of the proteins rather than their identities (reviewed in ref 37). To apply this approach, researchers have used surface enhanced laser desorption/ionization time-of-flight mass spectrometry (SELDI-TOF−MS). SELDI is similar to MALDI (matrix-assisted laser desorption/ionization), in that target molecules are laser desorbed and ionized for analysis by MS. With SELDI, however, molecules are present on the surface of protein chips, which have an active surface chemistry (hydrophobicity, cation exchange properties, anion exchange properties, or metal affinity) to retain proteins with complementary properties. For serum proteome pattern diagnostics, samples from affected and unaffected patients are individually applied to a chip and retained proteins are subsequently ionized and detected by TOF−MS. Sophisticated bioinformatics software is then used to compare spectra and determine discriminatory patterns of peaks within samples of unhealthy individuals.

In the first proof-of-principle study, a new computer-based artificial intelligence algorithm was used to identify patterns among a "training set" of mass spectral data, generated by SELDI-TOF−MS, starting with serum from 100 females, of whom 50 were diagnosed with ovarian cancer.[38] The algorithm generated a proteomic pattern that was then used to identify ovarian cancer in individuals from a second independent group of 116 individuals, of whom 50 were affected. The positive predictive value of this approach was 94%, the specificity was 95%, and the sensitivity was 100%.[38] Improvements in the resolution, sensitivity, and mass accuracy of mass analyzers, an increase in the sample size used as the initial training set, and the combined use of multiple SELDI chip surfaces can all be expected to improve the specificity and predictive value of this approach. Moreover, this technique can be used alongside

any number of other indicators such as genetic defects, or histopathological findings, to make more accurate diagnoses. In any case, the initial findings using this method are encouraging, and to generate mass spectra requires only a small volume of serum and is relatively fast, taking as little as 30 min.[38] Equally encouraging is the opportunity to apply this approach to a broad spectrum of diseases. In this regard, a proteome "signature" has already been developed to discriminate individuals with prostate cancer from healthy individuals.[39] A note of caution should be sounded however, in that far more patient and disease analyses will have to be done before the approach can truly be evaluated.

The method of serum proteome pattern generation more comprehensively exploits the information contained within the serum proteome compared to single molecule identification, and the approach exploits a major characteristics of complex diseases, namely, that there is a whole cohort of molecular changes ranging from different protein levels to changes in protein cleavage and other modifications. One can envision applying a similar approach to the problem of disease stratification, which will be useful for developing the most appropriate treatment strategies. Whether or not the use of SELDI−MS and bioinformatics for proteome pattern diagnostics will live up to such expectations any time soon, however, is the subject of an ongoing and lively debate.[40−43] Some researchers contest that SELDI is not sensitive enough, and captures only high-abundance proteins, and therefore is not suitable for measuring true cancer biomarkers. Of equal concern is the reproducibility of the technique. If important clinical decisions, such as whether to proceed with tissue biopsies or even major treatments (including surgeries), are to be based on a diagnostic approach, reproducibility is critical. In addition to these concerns, the concept of using a pattern of MS peaks to diagnose disease without knowing the identities of the proteins responsible for those peaks is a foreign one and a major point of contention for many researchers.[40−42] Although the diagnostic in this procedure is an actual mass spectral pattern, and can be made without any knowledge of the actual protein identities, understanding the nature of these proteins would clearly impact cancer research, and should be made a priority. Currently, one can only speculate on the nature of these proteins, and it is not known if they are tumor-derived, or if they are a general epiphenomenon of cancer. Again, we stress the need to obtain more data on this approach to evaluate its predictive power.

Although the importance of obtaining the identities of discriminatory proteins is highly recognized, it has been suggested, nonetheless, that efforts to use proteome pattern diagnostics to detect disease should proceed independently of efforts to obtain the relevant protein identities (which incidentally are underway).[43] Given the desperation for early detection methods for certain cancers (e.g., pancreatic, breast, and ovarian cancer), we too believe that this method could impact treatment outcomes even before the discriminatory proteins are known and characterized—at least for certain cancers which advance rapidly but for which no early detection measures exist. Regardless of the amount of work still to be done to address a number of concerns, there is a general consensus that the marriage of mass-spectrometry-based methods with pattern-recognition algorithms offers unprecedented promise for diagnosing cancer.[40,42,43] In addition, the approach has implications that extend beyond proteomics. Signatures for blood metabolite profiles, generated by MS or by NMR for instance, would also provide important diagnostic information which, if analyzed alongside proteome signatures, would offer a more comprehensive diagnostic tool. Combine such approaches with other methods such as gene expression analysis, sequencing etc., and the power of systems biology in medicine becomes clearer.

Protein Chips

Complementing the use of mass spectrometers in proteomics and in medicine is the use of protein microarrays. Using similar technologies already in place for the production of DNA microarrays, the goal behind protein microarrays is to print thousands of protein-detecting features, for the interrogation of biological samples. An example is antibody arrays (also referred to as protein profiling arrays), in which a host of different antibodies (e.g., monoclonal, polyclonal, antibody fragments) are arrayed to detect their respective antigens from a sample of human blood. In another variation of this approach (functional protein arrays) multiple protein types are arrayed for the study of a number of properties such as post-translational modifications, protein−DNA, protein−protein, and protein−ligand interactions, and biochemical functions such as enzyme activities (for recent reviews see refs 44−49). Ideally, these functional protein arrays would contain the entire complement of proteins from a given organism. The first version of such whole-proteome arrays consisted of 5000 purified proteins from yeast (*S. cerevisiae*) deposited onto glass microscope slides, and their utility for studying protein−protein and protein−lipid interactions on a global scale was demonstrated.[50] Despite the success of the first whole-proteome chip, the implementation of protein arrays is a much greater challenge than DNA arrays for a number of reasons. Proteins are inherently much more difficult to work with than DNA, their solubility varies widely, they have a broad dynamic range, they are much less stable than DNA, and their structure can be difficult to preserve on a glass slide, but is essential for most assays (unlike DNA, in which only the sequence order needs to be maintained). Finally, there is no technique, analagous to PCR, that exists for amplifying proteins, and thus the starting material is much more of a limiting factor.

It should be emphasized that the global ICAT technology described earlier will have striking advantages over protein chip technologies for the quantification of large numbers of proteins. Whether one uses, antibodies, proteins, or protein capture agents—the generation of the reagents and their disposition on the chip is time-consuming and expensive and the capacity for accurate quantification is highly problematic. A recently proposed adaptation of the ICAT technology (described below) can be expected to circumvent all of these challenges.

Reverse Phase Protein Microarrays

Here, we highlight a relatively newer microarray application that is particularly promising for the study, diagnosis, and treatment of complex diseases such as cancer. The technology merges laser capture microdissection (LCM) with microarray technology, to produce "reverse phase protein microarrays". In contrast to conventional protein microarrays, which contain immobilized probes, and are interrogated with protein samples (e.g., lysates from patient serum), in the case of reverse phase protein arrays, the whole collection of proteins themselves (e.g., from an individual patient sample) are immobilized with the intent of capturing various stages of disease within an indi-

vidual patient. When used alongside LCM, reverse phase arrays can monitor the fluctuating state of the proteome among different cell populations within a small area of human tissue. This will be particularly useful for profiling the status of cellular signaling molecules, or post-translational modifications, among a cross-section of tissue that includes both normal and cancerous cells.

In the initial studies, the feasibility of this approach was demonstrated by monitoring the status of key factors (e.g., factors indicative of pro-survival, mitogenic, and apoptotic pathways) in normal prostate epithelium, prostate intraepithelial neoplasia, and invasive prostate cancer tissues.[51,52] The tissues were dissected by LCM, and their protein lysates were arrayed onto nitrocellulose slides which were subsequently probed with specific anitbodies. The transition between the different tissues (reflective of cancer progression) was shown to be associated with a change in the status of signaling components such as Akt, GSK3B, PKC-α, p38, etc., providing important information about the specific molecular changes that accompany cancer progression.[51,52] This method can track all kinds of molecular events and can compare diseased and healthy tissues within the same patient, enabling the development of individualized diagnosis and treatment strategies. The ability to acquire proteomic snapshots of neighboring cell populations, using multiplexed reverse phase microarrays in conjunction with LCM, will have applications in a number of areas beyond the study of tumors. The approach can provide insights into normal physiology and pathology of all tissues, and will be invaluable for characterizing developmental processes and anomalies. It should be emphasized, however, that beyond reverse phase microarrays, the marriage of LCM with any refined proteomics platform offers great promise for extracting information from pure cell populations, in turn decreasing some of the limitations imposed by tissue heterogeneity.[53] In this regard, it is interesting to note that a quantitative analysis of the proteomes of hepatocellular carcinoma was recently achieved by coupling LCM with ICAT and 2D-LC−MS/MS.[54]

Emerging Trends in Proteomics

A number of emerging concepts have the potential to dramatically improve current capabilities in proteomics. Obtaining absolute quantification of proteins and monitoring post-translational modifications are two tasks that top the list in terms of their potential impact on our understanding of protein function in healthy and diseased cells. The ICAT approaches used thus far (described previously) offer the ability to obtain only relative protein quantifications. Aebersold has recently proposed a global approach (e.g., quantitation in principle of all proteins) to obtain absolute quantification of proteins from a single sample such as a particular cell type or tissue.[55] The idea is to chemically synthesize a cysteine-containing peptide from each of the 30 000 or so proteins encoded by the human (or mouse) genome, and label each with tags of a heavy stable isotope (admittedly not all proteins have these residues, but other ICAT specificities could be used for these proteins). A carefully quantified standard mixture of the 30 000 peptides (each labeled with the heavy ICAT reagent) could then be constructed and mixed with the light ICAT-labeled peptides from the proteins of a cell whose quantitative proteome is to be determined. In a typical vertebrate cell, there may be a total number of different proteins that could upon digestion with trypsin generate 1.5 million tryptic peptides, yet this newly proposed global approach could employ software that could instruct the mass spectrometer to analyze only those peptides displayed as pairs (the light unknown and its heavier counterpart from the standard mixture—that is just 30 000 pairs). This would minimize the computational requirements for this approach compared to standard analyses and absolute amounts of the unknown proteins could be determined. Moreover, this approach is far more global than the existing protein chips where 200 element features is a good-sized chip. An obvious drawback to this approach is that peptides are expensive to synthesize, but several companies are working on methods that could reduce the costs by one or 2 orders of magnitude.

Advances in quantitative proteomics would clearly enable more in-depth analyses of cellular systems. However, for many cellular events, protein concentrations likely do not change significantly, rather their function is modulated by posttranslational modifications (PTMs). Over 400 PTMs have been described, many with important influences on cell function. Methods of monitoring PTMs are sorely needed in proteomics, but to date, this remains an underdeveloped area. Although many researchers are optimistic that improvements in mass analyzers will give way to whole-protein proteomics, we are skeptical that modifications will be detected from whole proteins with sufficient accuracy and reproducibility. In the immediate future, the most efficient means for studying PTMs will likely be the selective capture of modified proteins followed by relative comparisons using the ICAT labeling method or a comparable labeling method.

Selecting a particular subset of proteins for analysis will substantially reduce sample complexity making this approach particularly advantageous for diagnostic procedures for which blood is the starting material. There are several general challenges in blood proteome diagnostics. For instance, about six proteins are present in high concentrations and constitute about 80% of the serum proteins (e.g., albumin alone constitutes about 51% of the blood protein). This makes it virtually impossible to visualize proteins that are present in the blood at lower levels—and many of these lowly expressed proteins are likely to be key diagnostic parameters. Aebersold and his colleagues have developed an approach, which solves these problems—by rendering the dominant proteins virtually invisible and by selecting only a subset of blood proteins for analysis.[56] The approach is to oxidize proteins containing N-linked sugars and covalently link the resulting proteins to beads. Only proteins linked to the beads are retained thereby eliminating nonglycosylated proteins (including albumin). The bead-linked proteins are subsequently digested with trypsin, further reducing sample complexity, as only the N-linked tryptic peptides remain covalently linked to the beads. The peptides can then be labeled at their amino terminus with isotopically light (normal blood) and heavy (disease blood) isotopic labels. Then the normal and pathologic sample beads are mixed, the peptides released from the beads and then analyzed by mass spectrometry just as with the ICAT procedure.[56]

This procedure was applied to an inbred strain of mice in which some animals had been treated with a chemical carcinogen to induce a malignant skin tumor (R. Aebersold, H. Zhang, E. Ye, X−J. Li, and P. Mallik, personal communication). Starting with only 50 μL of blood, N-linked protein abundance was compared between healthy and diseased animals as described above. About 3000 peptides were identified representing approximately 1000 proteins. About 100 of these

proteins (peptides) were collectively diagnostic of the normal and disease states. Thus, this procedure identifies peptides (proteins) from the blood, even those present a lower levels. In addition, because of the isotope labeling it also determined, quantitatively, the relative concentrations of the proteins from the normal and diseased bloods. It renders invisible the dominating effects of certain proteins (by removing most of those proteins after tryptic digestion and by assigning their few N-linked peptides to a very small fraction of mass spectrometry space). This technique can easily be applied to virtually any disease state. In collaboration with Dr. Aebersold, we are attempting to convey this procedure into a microfluidics format. Once again many more analyses will be need to verify the usefulness of the technique—but it does appear very promising.

In addition to absolute protein quantification and analysis of post-translational modifications, another important aspect of proteomics, not yet addressed, is that ultimately, proteomics methods should focus on studying proteins in the context of their environment. This is not typically done with the current use of protein arrays or with current approaches in mass spectrometry. However, the increasing use of chemical cross-linkers, introduced into living cells to fix protein−protein, protein−DNA and other interactions, immediately prior to harvesting, may ameleoriate this problem in part. The challenge in this respect is to identify suitable methods of preserving relevant interactions but not at the expense of fixing transient interactions that are not physiologically relevant. Another goal for studying proteins in vivo is to develop more sensitive and sophisticated methods to image proteins and other molecules in living cells and in real time. A particularly exciting advance in this area, which is detailed later, is the use of semiconductor quantum dots because they offer increased sensitivity of detection and a greatly increased potential for generating a multiplicity of reporter groups (greater than 50).

The Human Plasma Proteome

Characterizing the human plasma proteome has become a main goal in the proteomics arena. The plasma proteome is undoubtedly the most complex proteome in the human body; consisting not only of the resident, hemostatic proteins, but also immunoglobulins, cytokines, protein hormones, secreted proteins, and foreign proteins, indicative of infection. In addition, blood circulates through almost all tissues of the body, and therefore contains tissue leakage proteins, including those released from damaged or dying cells. The blood should, as noted above, therefore contain information on the physiological state of all tissues in the body. This, combined with its accessibility makes the blood proteome invaluable for medical purposes. With a detailed understanding of the plasma proteome, ultimately it will become possible to relate individual serum proteome profiles to the genomes, environments, and lifestyles of those individuals. As discussed above, these types of integrative studies will open the door to predictive and preventative medicine. However, even with recent advances in proteomics, characterizing the proteome of blood plasma is a daunting challenge. In addition to the immense repertoire of proteins present, the dynamic range of these proteins is on the order of 10^9, with serum albumin being most abundant (30−50 mg/mL) and low-level proteins such as interleukin-6 present at 0−5 pg/mL (reviewed in ref 57). Identifying proteins at each end of this spectrum in a single experiment is not feasible with current technologies.

Further complicating the study of the human plasma proteome are temporal and spatial dynamics. The turnover of some proteins is severalfold faster than others, and the protein content of the arteries may differ substantially from that of the veins, or the capillary proteome may be specific to its location, etc. All of these differences make even the most simple proteomics task of cataloging the proteome seem out of reach. Factoring in the importance of understanding post-translational modifications, protein interactions and other aspects, and the challenge becomes overwhelming. To tackle this problem, priorities need to be established. Capturing the most meaningful subset of proteins (multiparameter analyses) among the entire proteome to generate diagnostics tools is one such priority, and relating this information to other data types should be another. In this regard, the focus should be directed at the less abundant fraction of proteins in the blood plasma, as it is this fraction that best reflects tissue physiology and pathology, as these are the nonresident proteins that are either actively or passively released from tissue into the blood stream (reviewed in ref 58). Second, since cancer is associated with enhanced glycosylation of proteins, methods that focus on this fraction of proteins will also be useful. It should be stressed again, that multiparameter anaylses will best reveal a pathological state. For instance, the algorithms described above, for generating discriminatory patterns of diseases, should be used with mass spectra from different proteome fractions (e.g., low abundance proteins and glycosylated proteins), and gene expression changes revealed by microarray analysis. We discussed earlier a new blood proteomics technique that focused on proteins with one or more N-glycosylated sites.[56] As proteomics techniques improve, the disease profiles generated should be continually related to the respective gene expression changes, the genome sequence information, etc. One can imagine in the future, the existence of large archives against which all of this information is compared to reveal predisposition to, or onset of, disease.

Merging Biology with Nanotechnology and Microfluidics: Shrinking Medical Tool Kits

Even the most fundamental task of proteomics—identifying the individual proteins within a complex mixture—is a multistep process involving sample preparation from whole cells, protein separation and digestion, peptide fractionation, and peptide detection. Defining the most efficient ways to carry out these steps has been a remarkable challenge even with large quantities of cultured yeast available. To use similar methods in the health care arena, where the starting protein material will be derived from small tissue samples or small volumes of body fluids, miniaturization of analytical instruments is critical. Developing any technology intended for clinical use will require the minaturation, integration and automation of the procedures for sample analyses. This in turn will lead to more sensitive and cost-effective analyses. The fabrication of micro- and nanoscale devices for assessing biological processes—an area of development that has exploded over the past five years—will be the key to achieving these goals. Biological systems are made up of individual molecules operating on a nanoscale, whereas current tools used in medicine are much larger and thus inadequate for fully characterizing cellular function at the molecular level. This, combined with the technical issues of dealing with small starting samples, makes "nanobiotechnology" an area that will undoubtedly revolutionize biology and health care.

The Promise of Quantum Dots

Though the integration of biology with nanoscale science is still in its infancy, nanostructured materials are emerging which will provide new and powerful tools for biology and health care. For instance, semiconductor nanocrystals, also referred to as quantum dots, promise to transform in vitro imaging. Quantum dots are semiconductor crystallites (in the size range of 2−8 nm (a scale comparable to many cellular macromolecules) that are highly light absorbing over a broad spectral range. They can be linked to biological molecules such as peptides, proteins, or nucleic acids. Quantum dots are emerging as a preferable class of biological label with properties that are much more desirable compared to traditional dyes and fluorescent proteins (reviewed in ref 59). For instance, the longevity of quantum dot florescence far exceeds that of other fluorophores, and quantum dots can be made in a multiplicity of colors according to their size (perhaps 10 different colors are now available and 10s more are expected soon), making them preferable to the use of common reporters such as green fluorescent protein or luciferase.

Recent studies underscore the in vivo potential of quantum dots for biological studies and for therapeutics. First, they were encapsulated in special micelles to make them biocompatible, and when conjugated to DNA, nanocrystal-micelles hybridized to complementary DNA and acted as fluorescent probes. When injected into individual cells of *Xenopus* embryos, nanocrystal-micelles were cell autonomous and, given their stability and a lack of photobleaching, they were visible until the tadpole stage, enabling lineage tracing experiments,[60] and, in the future, comparative embryology. In another study, quantum dots were coated with different "homing peptides", that is, peptides that recognize specific tissue markers or addresses. As a result, the peptides, delivered intravenously in mice, targeted the quantum dots to the appropriate sites (in this case lung, or vascular tissue) with remarkable specificity.[61] Although the use of nanocrystals in vivo is still in its infancy, the demonstrated ability to specifically target nanocrystals suggests a future in which nanomachines are used for disease detection and drug delivery. Finally, more recently, quantum dots were used to track individual glycine receptors, monitoring their dynamics within neuronal membranes of live cells.[62] The real-time, in vivo imaging of protein, DNA, and lipids has always presented a major challenge for medicine.

Microfluidics

A form of miniaturization that promises to minimize the time and cost of biological assays is microfluidics. Microfluidics collectively refers to technologies and tools that enable controlled transport of tiny volumes of liquid in glass, silicon or plastic molds. Microfluidic devices take advantage of microfabrication technologies that are commonly used in microelectronics. An ultimate goal of microfluidics is to create small devices that can carry out multiple experiments and to integrate together a series of procedures starting with small volumes of liquid. Currently, only a few microfluidic devices have been made commercially available.

Steve Quake and his group at Caltech have pioneered the development of a unique and flexible class of microfluidics technologies. They have developed large-scale integrated microfluidic circuitry that is based on multilayer soft-lithography to fabricate micro-molded elastomer valves, channels, chambers and pumps. These microfluidics platforms have been deomonstrated for cell-sorting applications, gene expression profiling (PCR-based) at the single cell level, the production of protein crystals and much more. These microfluidic devices offer the ability to integrate complex biological procedures in precisely the same manner that silicon chip technology has permitted the integration of complex information technology procedures.

Sensing Biomolecular Interactions: Microcantilevers and Nanowire Sensors

A second, exciting concept emerging in biology is the use of microcantilevers to study molecular interactions. Studies have demonstrated the ability of microcantilevers to detect the nanomechanical changes that take place during biomolecular interactions.[63−66] Specifically, when biomolecular interactions take place on one surface of a microcantilever beam, the nanomechanical forces of the interaction cause the cantilever to bend. When molecules bind to the surface of cantilevers, they induce movement of only 10−20 nanometers. These movements can be detected by lasers, which are capable of detecting deflections as small as a fraction of a nanometer. The use of microcantilevers for diagnostics and drug discovery is far from reality, but since the initial reports that cantilevers bend upon molecular binding, the performance and capabilities of microcantilevers has become of great interest.

In theory, the use of microcantilevers would far surpass current methods of measuring biomolecular interactions given the inherent precision and the lack of a labeling step, which can alter the properties of various biomolecules. Some recent demonstrations in particular provide encouraging results for the use of this technology for diagnostics and drug discovery. First, using cantilevers, researchers were successful in reproducibly measuring the deflections caused by a 12-mer hybridizing to a complementary sequence immobilized on the cantilever. More importantly, however, the change in deflection caused by a single base-pair mismatch was also detectable.[63] In the same study, the binding affinities of protein A for the constant region of immunoglobulins (IgGs) from rabbit and goat were distinguishable. The higher affinity of protein A for rabbit IgG, which was previously known, was detected by greater bending of protein A-coated cantilevers in response to the rabbit IgG.[63] A more recent study took the technology one step further by demonstrating the successful detection of two forms of prostate-specific antigen (PSA) over a range of concentrations, in a background of 1 mg/ml human serum albumin and human plasminogen.[67] PSA, found at higher levels in the blood of men with prostate cancer, is an extremely useful biomarker for not only detecting primary prostate cancer, but also for monitoring progression of the disease, and for evaluating treatment efficacy. The results from this study indicate that the technique of measuring nanomechanical changes with microcantilevers is sensitive enough to detect levels of PSA that are 20-fold lower than levels that considered to be clinically relevant.[67] Importantly, this technology is also specific enough to measure PSA levels among high background levels of other proteins such as albumin,[67] a technical hurdle that is common to the use of blood for diagnostics. With these capabilities, it is conceivable that many diseases, for which biomarkers have been identified, could be assayed using microcantilevers.

Researchers can envision a generation of arrays or "microcantilever chips" that are capable of assaying multiple proteins or other protein- or DNA-binding molecules in a single experiment or diagnostic test. Such devices would be compa-

rable to the use of DNA microarrays, but would have much more widespread application and, with sub-nanometer precision, be much more sensitive and require much less starting material. As exciting as this prospect is, however, moving the technology from the initial proof-of-principle stage, in which one cantilever is used at a time, to an array format for which several hundred cantilevers are represented on a single chip, is not trivial. Nonetheless, there is a great impetus for creating such arrays and early studies are hopeful.

For the past few years, there have been literature reports of chemical sensors based on single-walled carbon nanotubes or semiconductor nanowires. The idea is that a capture molecule (antibody or single strand of DNA) may be attached to the nanowire so that upon binding of its cognate molecule, measurable changes in the conductivity of the nanowire occur. Such a detection device has the potential to be highly sensitive (in principle down to single molecules) and one can imagine constructing parallel arrays of nanowires (1000 in the diameter of the typical cell—10 μm). Thus, one can envision functionalizing each of these 1000 nanowires with a different capture molecule, such that the mRNA or protein molecules from a single cell can be captured and quantitatively measured. Moreover, the measurements are taken in real time and do not require any modifications or reporter groups so that rapid physiological processes (approximately 0.1 s) can be captured. Jim Heath and his colleagues at Caltech are pioneering this approach. The Institute for Systems Biology, Caltech and UCLA have recently formed the NanoSystems Biology Alliance with the primary objective of using use the needs of systems biology to drive the design and development of nanolab chips that have the ability to make five kinds of measurements on single cells: phenotypic assays (single cells placed on nanopores and the individual cell behaviors monitored either optically or electronically); mRNA and protein concentrations (measured by functionalized nanowires); and protein/protein and protein/DNA interactions (measured by functionalized nanocantilevers) (Figure 4). The idea is that l00s of nanolabs (each capable of analyzing single cells and their contents) can be integrated together with a microfluidics device that will bring in cells from the outside world—thus integrating the worlds of biology and nanotechnology. Clearly, these advances in measurements will revolutionize both biology and medicine.

Potential Impacts of Miniaturization in Health Care

As mentioned at the outset of this review, mutiparameter analyses of the blood will provide a window into the differentiation of health and disease. These measurements will undoubtly be carried out by hand-held devices that integrate microfluidics and nanotechnology, the so-called laboratory on a chip, similar to the one that is being developed by the NanoSystems Biology Alliance described above. With these devices, one can eventually imagine analyzing 100s, 1000s, or even 10 000s blood elements. In addition, we predict that individuals will have their genomes sequenced relatively inexpensively within the next l0—15 years, making it possible to provide each individual with a probabilistic future health history. Thus, the predictive medicine will assess the digital information of the genome and the pathological cues of the environment.

Another area for which nanotechnology has an application is that of drug delivery systems. It is conceivable that in the future, drugs will be delivered to specific targets in the body via biodegradable devices. Implantable biosensors can also be foreseen, which can monitor sugar levels in the cells of diabetics, and release insulin as needed, resulting in much more precise control of blood sugar levels than is currently attainable in diabetics. Finally, an exciting possibility is the use of microrobots and probes, which can target and destroy tumors. The potential for these technologies to reach fruition is met with as much if not more skepticism than support. However, with each new breakthrough, new measurements become possible and the quality of data dramatically improves. The immediate challenge will be to push these technologies beyond the proof-of-principle stage to a point where they are at least providing novel information. With the appropriate infrastructural changes, and with partnerships between academia and industry such as the NanoSystems Biology Alliance, these technologies will find a place in health care.

Bioinformatics and Proteomics

Proteomics has reached a state of generating voluminous data sets, necessitating computational biologists, mathematicians, and statisticians to help deal with the overwhelming amount of data generated. As described above, tandem mass spectrometry has become the method of choice for protein identifications. In this approach, proteins are identified from peptide fragments by matching each MS/MS spectrum to a database of sequences. To do this requires the use of database search programs such as SEAQUEST,[68] MASCOT,[69] ProFound (http://prowl.rockefeller.edu/cgi-bin/ProFound), MS—Tag[70] and Sonar.[71] These programs are used determine the amino acid sequence and thus the protein(s) corresponding to a given mass spectrum, but in many cases they generate a large number of incorrect assignments. Improvements to the current capabilities for tandem-MS identification are continually being developed.[72] To validate peptide assignments in a more automated, unbiased manner, Nesvizhskii and colleagues in Reudi Aebersold's group recently developed PeptideProphet, which uses a statistical approach to validate peptide identifications made by tandem MS and database searching.[73] This approach is based on the expectation maximization algorithm, and computes probabilities, which accurately apply confidence measures to protein identification. The availability of database-matching programs and for statistical methods to comb through datasets, which are growing in size and complexity, emphasize the importance of bioinformatics for exploiting proteomic data. The development and use of programs, like PeptideProphet raise the bar on formats for publishing mass spectrometry data, ultimately reducing the noise present in ever-growing databases.

High-Quality, Unified Databases Are Essential

In addition to developing more effective ways to filter and interpret the output from tandem-MS, a major task is to assemble protein identification information into databases that present the data in as useful and comprehensive, format as possible. Thus far, there has been an overwhelmingly large amount of data generated, but these data are fragmented in different databases, with high error rates estimated. In most cases, there are no metrics for evaluating the quality of the global data sets (especially the global protein data sets such as mass spectrometry data, protein/protein interactions and protein/DNA interactions). This is an extremely important objective for the near future, as we discuss below. The key to

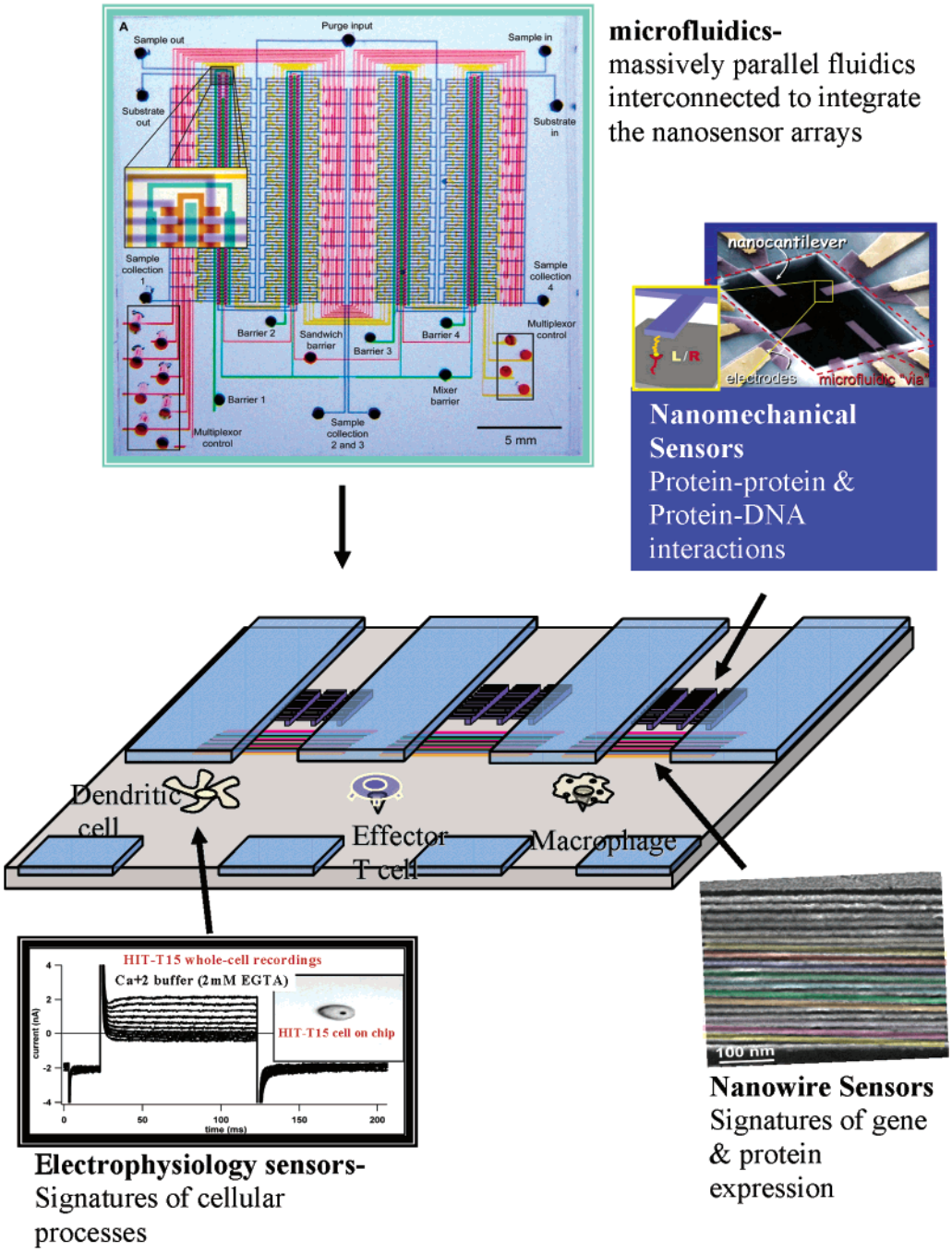

Figure 4. Illustration of a lab-on-a-chip for taking multiple measurements from single cells. Microfluidic circuits can be designed to accommodate a number of analytical biochemical applications and to support parallelized, high-throughput screening. Given the dimensions of these devices, they will be suitable for single-cell analysis. Shown at the top is an optical micrograph of a microfluidic system generated in Steve Quake's lab (California Institute of Technology), in which the various inputs were loaded with food dyes to show the valves and flow structure (this portion of the figure reprinted with permission from Hong et al.[78] Copyright 2003, Nature Publishing Group. Also shown are representative applications that can be carried out at the single cell level to obtain multiple measurements using a single device. These include protein and/or gene expression analyses, protein−protein or protein−DNA interaction studies (using cantilevers), and electrophysiological recordings.

the wide-spread accessibility of DNA sequence data emerging from the genome project was a direct consequence of the fact that a widely accepted metric was available for evaluating the quality of any DNA sequence data produced.[74] For example, the various large datasets of protein−protein interactions vary enormously in their error rates—and there is no simple way to compare different interaction data sets. Synchronizing these databases will facilitate efforts to exploit this information, but is no small task. Fortunately, the Human Proteome Organization (HUPO) (http://www.hupo.org/information/mission.htm), which was formed to coordinate worldwide proteomic efforts, has taken on this challenge and through the Proteomic Standards Initiative (PSI) group (http://psidev.sourceforge.net/), established in 2002, is developing a common data standard which will enable users to retrieve data from different sites and perform comparative analyses of different data sets.

Aside from curating data, an essential task, as noted above, is to set data quality standards and insist that data meet these standards prior to their addition to databases. In a similar manner, a standard format has been adapted by the Microarray Gene Expression Data Society (MGEDS) for depositing microarray expression data. As a result, MAGE-ML (MicroArray Gene Expression Markup Language) was designed to describe and communicate information about microarray experiments, incorporating the principles outlined by an earlier standard, MIAME (Minimum Information About a Microarray Experiment). Similar to MAGE and MIAME, standards for depositing proteomic data will undoubtedly evolve as the field of proteomics does. Although tens of thousands of protein−protein interactions have been described for yeast, and deposited into widely used databases, there are believed to be extremely high error rates, in turn slowing the process of generating accurate biological models. Although error is inevitable, especially given the current technological limitations, a minimum set of standards needs to be agreed upon, and should be continually upgraded as the technology and bioinformatics applications improve. This challenge also tops the list of priorities for PSI, which will first focus on two areas: protein−protein interaction data and mass spectrometry data. Moving forward on this initiative will require the establishment of a common data format, applicable to a range of analytical platforms that are currently in use for carrying out proteomics. True to its mission, PSI has already initiated efforts to develop a standardized general proteomics format.

It is important to point out that error or signal-to-noise ratios in global data sets arise from two distinct sources. There is noise arising from the measurements of the data (instrument noise) and there is noise arising from the biology itself (biological noise). Each of these noise sources must be considered separately. The errors in global data sets are clearly a combination of the two. To give an example, when measuring the proteome of a population of cells, it is obvious that individual cells may be at very different stages in the cell cycle, may be at different stages in responding to a physiological signal, etc. Hence, the population measurement is an average of the different states of many different cells. Very misleading conclusions could arise from the analysis of heterogeneous populations of cells. Clearly, the ability to move to single cell measurements will, at least in part, rectify this type of biological noise.

The challenges posed by a need for unified databases are immense, but must be an immediate priority in the field, as the ability to derive meaningful information from protein studies hinges entirely on the quality of the data generated, the appropriate curation of that data, and the accessibility. The availability of high-quality protein databases will be also be essential for predictive and preventative medicine. Ideally, information from basic research should be related to data from pre-clinical and clinical trials. Additional data that exists such as genotyping information, patient history, disease, response to treatment, biomarker levels, etc., also need to be accurately preserved and somehow linked to pertinent proteomics data. To not create and support comprehensive, standardized databases that are universally usable and freely accessible, is to severely compromise the ability to link protein deficiencies to genetic defects and to disease. Essentially, a lack of high-quality databases will significantly delay efforts toward achieving predictive and preventative medicine. Fortunately, coordinated efforts are currently underway to avoid such unnecessary delays. For a more thorough description of the issues facing the Proteomics Standards Initiatives as well as its mission and planned strategies, the reader is referred to the PSI's published meeting reviews.[75−77] Parenthetically, we are attempting to develop a database at the Institute for Systems Biology, (Systems Biology Expression and Management system, or SBEAMs), that will be able to acquire all relevant types of global data sets (DNA, RNA, proteins, interactions, phenotypic data, etc.) and begin to do the integrations that are an essential part of systems biology.

Computational Integration

The goal of cataloguing all of the cellular elements under various conditions and in various organisms is well underway, and becoming increasingly possible as global technologies mature. The next phase is to understand how these elements are coordinated to form functional biological systems. Systems-level integration of data is still in its infancy, but a number of new concepts have emerged. Assimilating information from disparate data sets serves at least two important purposes. First, data integration minimizes the noise that is inherent in data generated through large-scale, high-throughput biology. An excellent example of this is demonstrated in the transcription factor analyses carried out by Lee et al., 2003, in which genome-wide location data was filtered with microarray expression data to attach confidence to their protein-DNA interactions.[19] Similarly, as described earlier, we used the combined information from four distinct studies to generate a list of Gal4p-binding promoters. The second benefit of data integration is that it serves to reveal new biological phenomena, which would not be readily apparent from any single analysis. For example, the study of the galactose utilization system in yeast allowed us to integrate mRNA and protein concentration data to suggest that approximately half of the protein concentrations are controlled by posttranscriptional mechanisms. Without the integration of these two different data types, this conclusion could not be reached. The ultimate goal is to characterize the information flow through protein networks that interconnect the extracellular microenvironment with the control specified by gene regulatory networks which, in turn, active the peripherial batteries of genes to execute the effector functions of development and physiological responses. To successfully understand the interfacing of these protein and gene regulatory networks will require, ultimately, the integrations of many of the different data types arising from DNA, RNA, protein, metabolites, small molecules, and many different aspects of phenotype.

Proteomics and the Future of Medicine

The field of proteomics is rapidly evolving. A major task is to determine how best to use the currencies of this field to effect change in health care. It is important to recognize that this challenge is as much political, social, ethical, and legal as it is technological. A number of factors will dictate the success of proteomics, not the least of which is an organized effort to define the overall goals of this discipline, establish the immediate priorities, and outline coordinated strategies. A collaborative, organized approach was necessary to initiate the Human Genome Project, and will be even more essential for characterizing the far more challenging task of whole proteomes. The field of proteomics may benefit by the creation of the Human Proteome Organization (HUPO). As its mission, HUPO endeav-

ors to consolidate national and regional proteome organizations, to disseminate and promote proteomics research, and to coordinate proteomic initiatives. If researchers, worldwide, coordinate their efforts through HUPO, the establishment and maintenance of this organization may represent the most essential milestone for realizing the potential of proteomics in predictive and preventative medicine. It is worth pointing out that there was a similar organization created at the beginning of the human genome program, HUGO—the human genome organization. HUGO, however, did not contribute significantly to the genome project because of a lack of leadership and relevance to the tasks at hand. It will be interesting to see HUPO's impact on the emerging field of proteomics.

In addition to worldwide collaborations, proteomics and systems biology require collaborations between different scientific disciplines. Proteomics alone requires a critical mass of scientists who are adept in the areas of cell biology, protein biochemistry, mass spectrometry, computation, mathematics, statistics, informatics, and engineering—all of whom need to work freely work together to solve technical and biological problems. In addition, to truly harness the power of proteomics requires that findings be integrated with data from other studies including genomic analyses, transcriptome profiling, metabolite measurements, etc. Currently, typical faculty environments in academia do not cultivate cross-disciplinary science. In addition, there are high costs associated with the use of global, high-throughput technologies. If these issues are not dealt with, the dissemination of systems biology throughout the scientific community will be delayed, and young scientists training in these environments will be ill-equipped for practicing modern biology. To avoid this, large-scale technologies need to be accessible to more researchers in academic institutions. This will require the establishment of core facilities run by personnel who can maintain them at the cutting-edge, both within the academic setting, and as core facilities servicing the greater academic community. Methods of promoting interdepartmental and interfaculty collaborations are also crucial. While new grant initiatives, aimed at promoting systems biology, provide good incentive for cross-disciplinary science, this is only a start. Clearly, integrative systems-biology research centers or institutes will emerge, although the challenge will be not to burden them with academic bureaucratic structures that have been fashioned for small science. Although no clear consensus on this issue exists, it is anticipated that academic departments will increasingly need to interface with integrative research centers, and educational programs will be restructured so that science programs can be built around systems biology and its cross-disciplinary approach to science.

It is important to ensure that advances in proteomics do not merely remain at an academic, or research level. Many of the goals of systems biology and proteomics are long-term, and the ability to improve diagnosis and treatment of certain diseases will present itself far before we acquire a full knowledge of the human proteome. As scientists develop new technologies and arrive at new findings, they have an obligation to foresee their applications. Bridging the gap between basic research and clinical applications, however, is a major challenge that will only be circumvented by effective partnerships between academia, and the pharmaceutical and biotechnology industries. Each of these entities has its individual strengths, but on their own, neither can produce the technologies and drugs necessary to predict and prevent disease.

Predictive, preventative, and personalized medicine is a concept with a whole host of ethical issues. Although a full discussion of ethics is beyond the scope of this review, the ethical challenges posed by introducing genomics, proteomics, and emerging technologies into medicine certainly need to be underlined. For instance, there is the concern that an individuals genomic or proteomic profile could end up in the hands of employers or insurance companies; that allocating resources toward personalized medicine is not the most effective use of available funds; that proteomic data could be made confidential and/or proprietary; and that academic departments could be exploited or even become commercialized as a result of the necessary formation of corporate-academic alliances. These are just a few of many ethical concerns associated with the permeation of proteomics and other emerging capabilities into the health care field. Similar to the partnerships required for solving various technical challenges, dealing with these ethical issues will require alliances among researchers, health care workers, insurance companies, educators, legislators, policy makers, industry, media, and the public.

Summary

In conclusion, the emerging fields of systems biology and proteomics offer exciting and promising advances toward predictive, preventative, and personalized medicine. To fully realize the potential of these technologies and new insights, however, a number of issues and challenges remain. First and foremost, researchers need to continue to learn how to do systems biology. This will require developing new global technologies for genomics, proteomics, metabolomics, and phenotyping. It will require developing software that can capture, store, analyze, graphically display, integrate, model and disperse the global data sets of systems biology. We must learn how to determine the natures of protein and gene regulatory networks and their integrations. We must learn how to integrate many types of data and to analyze and integrate global data sets across the dynamic transitions of development or physiological responses. We also must deal with the challenge of providing access for the laboratories practicing small science to these global technologies and powerful computational tools. Finally, we must have access to biological samples from a large number of normal and diseased patients to begin the global correlative studies that will establish the foundational framework of predictive medicine and pave the way for moving forward into preventive medicine.

To summarize, we highlight some of the key considerations, outlined in this review, for integrating systems biology and proteomics into medicine:

- Understanding protein and gene regulatory networks of biological systems will improve drug development efforts and eventually will lead to preventive drugs. This approach will serve as the foundation for preventive medicine. These networks have key nodal points, the targeting of which will allow one to circumvent the disease potentials emerging from defective genes (somatic or inherited) or pathological environmental stimuli. These nodes may therefore be more effective targets for therapeutic interventions.
- The boundaries are fading between basic research and the clinical applications of systems biology and proteomics. Proteomics will play a major role both in developing better multiparameter diagnostics and in the search for new therapeutic targets. Proteomics is an immature technology and will require enormous resources to promote the development of

appropriate technologies, software and strategies. It will also play a major role in designing preventive drugs.

• Integrating different types of biological information will be critical both for understanding biological systems and for accurately diagnosing and monitoring disease. Computers are essential to this integration.

• Nanotechnology and microfluidics platforms are emerging, which promise to revolutionize research and medicine. With these technologies, multivariate measurements can be obtained efficiently and with small samples, and studying systems at a nanoscale will be feasible.

• Despite a number of technical concerns to address, replacing single-molecule biomarker analysis with serum proteome multiparameter diagnostics may represent the most promising advance toward early detection of diseases such as cancer.

• There is a large and growing list of applications for studying proteomes. Devising innovative ways to combine platforms, integrate their information, and exploit their unique advantages, will expedite their application in clinical practice and make maximal use of their capabilities.

• Diverse alliances need to be formed between academia and industry to expedite the development of new systems and to integrate them into the clinic.

• A majority of academic researchers need better access to high-throughput facilities for global technologies, including DNA arrays, sequencing, genotyping, and various proteomics platforms.

• Outlining objectives, defining strategies, and coordinating efforts are all essential for efficiently dealing with the enormous challenges in proteomics. If the expectations of HUPO are met, the establishment of this and other organizations will be instrumental to future successes in this field.

Acknowledgment. The authors wish to acknowledge Reudi Aebersold (Institute for Systems Biology) for helpful insights and Steve Quake and James Heath (both from California Institute of Technology) for interesting discussions.

References

(1) Hood, L.; Galas, D. *Nature* **2003**, *421*, 444−448.
(2) Ideker, T.; Galitski, T.; Hood, L. *Annu. Rev. Genomics Hum. Genet.* **2001**, *2*, 343−372.
(3) Heath, J. R.; Phelps, M. E.; Hood, L. *Mol. Imaging Biol.* **2003**, *5*, 312−325.
(4) Lohr, D.; Venkov, P.; Zlatanova, J. *Faseb J.* **1995**, *9*, 777−787.
(5) Reece, R. J. *Cell. Mol. Life Sci.* **2000**, *57*, 1161−1171.
(6) Peng, G.; Hopper, J. E. *Mol. Cell. Biol.* **2000**, *20*, 5140−5148.
(7) Peng, G.; Hopper, J. E. *Proc. Natl. Acad. Sci U S A* **2002**, *99*, 8548−8553.
(8) Mylin, L. M.; Bhat, J. P.; Hopper, J. E. *Genes Dev.* **1989**, *3*, 1157−1165.
(9) Mylin, L. M.; Johnston, M.; Hopper, J. E. *Mol. Cell. Biol.* **1990**, *10*, 4623−4629.
(10) Bhaumik, S. R.; Green, M. R. *Genes Dev.* **2001**, *15*, 1935−1945.
(11) Larschan, E.; Winston, F. *Genes Dev.* **2001**, *15*, 1946−1956.
(12) Ideker, T.; Thorsson, V.; Ranish, J. A.; Christmas, R.; Buhler, J.; Eng, J. K.; Bumgarner, R.; Goodlett, D. R.; Aebersold, R.; Hood, L. *Science* **2001**, *292*, 929−934.
(13) Ren, B.; Robert, F.; Wyrick, J. J.; Aparicio, O.; Jennings, E. G.; Simon, I.; Zeitlinger, J.; Schreiber, J.; Hannett, N.; Kanin, E.; Volkert, T. L.; Wilson, C. J.; Bell, S. P.; Young, R. A. *Science* **2000**, *290*, 2306−2309.
(14) Kellis, M.; Patterson, N.; Endrizzi, M.; Birren, B.; Lander, E. S. *Nature* **2003**, *423*, 241−254.
(15) Davidson, E. H.; Rast, J. P.; Oliveri, P.; Ransick, A.; Calestani, C.; Yuh, C. H.; Minokawa, T.; Amore, G.; Hinman, V.; Arenas-Mena, C.; Otim, O.; Brown, C. T.; Livi, C. B.; Lee, P. Y.; Revilla, R.; Rust, A. G.; Pan, Z.; Schilstra, M. J.; Clarke, P. J.; Arnone, M. I.; Rowen, L.; Cameron, R. A.; McClay, D. R.; Hood, L.; Bolouri, H. *Science* **2002**, *295*, 1669−1678.
(16) Yuh, C. H.; Bolouri, H.; Davidson, E. H. *Science* **1998**, *279*, 1896−1902.
(17) Shen-Orr, S. S.; Milo, R.; Mangan, S.; Alon, U. *Nat. Genet.* **2002**, *31*, 64−68.
(18) Milo, R.; Shen-Orr, S.; Itzkovitz, S.; Kashtan, N.; C*hkl*ovskii, D.; Alon, U. *Science* **2002**, *298*, 824−827.
(19) Lee, T. I.; Rinaldi, N. J.; Robert, F.; Odom, D. T.; Bar-Joseph, Z.; Gerber, G. K.; Hannett, N. M.; Harbison, C. T.; Thompson, C. M.; Simon, I.; Zeitlinger, J.; Jennings, E. G.; Murray, H. L.; Gordon, D. B.; Ren, B.; Wyrick, J. J.; Tagne, J. B.; Volkert, T. L.; Fraenkel, E.; Gifford, D. K.; Young, R. A. *Science* **2002**, *298*, 799−804.
(20) Gygi, S. P.; Rist, B.; Gerber, S. A.; Turecek, F.; Gelb, M. H.; Aebersold, R. *Nat. Biotechnol.* **1999**, *17*, 994−999.
(21) Griffin, T. J.; Gygi, S. P.; Ideker, T.; Rist, B.; Eng, J.; Hood, L.; Aebersold, R. *Mol. Cell. Proteomics* **2002**, *1*, 323−333.
(22) Baliga, N. S.; Pan, M.; Goo, Y. A.; Yi, E. C.; Goodlett, D. R.; Dimitrov, K.; Shannon, P.; Aebersold, R.; Ng, W. V.; Hood, L. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 14 913−14 918.
(23) Gygi, S. P.; Corthals, G. L.; Zhang, Y.; Rochon, Y.; Aebersold, R. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 9390−9395.
(24) Fey, S. J.; Larsen, P. M. *Curr. Opin. Chem. Biol.* **2001**, *5*, 26−33.
(25) Goodlett, D. R.; Keller, A.; Watts, J. D.; Newitt, R.; Yi, E. C.; Purvine, S.; Eng, J. K.; von Haller, P.; Aebersold, R.; Kolker, E. *Rapid Commun. Mass Spectrom.* **2001**, *15*, 1214−1221.
(26) Peters, E. C.; Horn, D. M.; Tully, D. C.; Brock, A. *Rapid Commun. Mass Spectrom.* **2001**, *15*, 2387−2392.
(27) Munchbach, M.; Quadroni, M.; Miotto, G.; James, P. *Anal. Chem.* **2000**, *72*, 4047−4057.
(28) Flory, M. R.; Griffin, T. J.; Martin, D.; Aebersold, R. *Trends Biotechnol.* **2002**, *20*, S23−29.
(29) Oda, Y.; Huang, K.; Cross, F. R.; Cowburn, D.; Chait, B. T. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 6591−6596.
(30) Zhou, H.; Watts, J. D.; Aebersold, R. *Nat. Biotechnol.* **2001**, *19*, 375−378.
(31) Goshe, M. B.; Conrads, T. P.; Panisko, E. A.; Angell, N. H.; Veenstra, T. D.; Smith, R. D. *Anal. Chem.* **2001**, *73*, 2578−2586.
(32) Ficarro, S. B.; McCleland, M. L.; Stukenberg, P. T.; Burke, D. J.; Ross, M. M.; Shabanowitz, J.; Hunt, D. F.; White, F. M. *Nat. Biotechnol.* **2002**, *20*, 301−305.
(33) Goshe, M. B.; Veenstra, T. D.; Panisko, E. A.; Conrads, T. P.; Angell, N. H.; Smith, R. D. *Anal. Chem.* **2002**, *74*, 607−616.
(34) Ranish, J. A.; Yi, E. C.; Leslie, D. M.; Purvine, S. O.; Goodlett, D. R.; Eng, J.; Aebersold, R. *Nat. Genet.* **2003**, *33*, 349−355.
(35) Shiio, Y.; Eisenman, R. N.; Yi, E. C.; Donohoe, S.; Goodlett, D. R.; Aebersold, R. *J. Am. Soc. Mass Spectrom.* **2003**, *14*, 696−703.
(36) Han, D. K.; Eng, J.; Zhou, H.; Aebersold, R. *Nat. Biotechnol.* **2001**, *19*, 946−951.
(37) Petricoin, E. F.; Liotta, L. A. *J. Nutr.* **2003**, *133*, 2476S−2484S.
(38) Petricoin, E. F.; Ardekani, A. M.; Hitt, B. A.; Levine, P. J.; Fusaro, V. A.; Steinberg, S. M.; Mills, G. B.; Simone, C.; Fishman, D. A.; Kohn, E. C.; Liotta, L. A. *Lancet* **2002**, *359*, 572−577.
(39) Petricoin, E. F., 3rd; Ornstein, D. K.; Paweletz, C. P.; Ardekani, A.; Hackett, P. S.; Hitt, B. A.; Velassco, A.; Trucco, C.; Wiegand, L.; Wood, K.; Simone, C. B.; Levine, P. J.; Linehan, W. M.; Emmert-Buck, M. R.; Steinberg, S. M.; Kohn, E. C.; Liotta, L. A. *J. Natl. Cancer Inst.* **2002**, *94*, 1576−1578.
(40) Cottingham, K. *Anal. Chem.* **2003**, *75*, 472A−476A.
(41) Diamandis, E. P. *J. Natl. Cancer Inst.* **2003**, *95*, 489−490.
(42) Diamandis, E. P. *Clin. Chem* .**2003**, *49*, 1272−1275.
(43) Petricoin, E., 3rd; Liotta, L. A. *Clin. Chem.* **2003**, *49*, 1276−1278.
(44) Cahill, D. J.; Nordhoff, E. *Adv. Biochem. Eng. Biotechnol.* **2003**, *83*, 177−187.
(45) Cutler, P. *Proteomics* **2003**, *3*, 3−18.
(46) Jona, G.; Snyder, M. *Curr. Opin. Mol. Ther.* **2003**, *5*, 271−277.
(47) Liotta, L. A.; Espina, V.; Mehta, A. I.; Calvert, V.; Rosenblatt, K.; Geho, D.; Munson, P. J.; Young, L.; Wulfkuhle, J.; Petricoin, E. F., 3rd *Cancer Cell.* **2003**, *3*, 317−325.
(48) Lopez, M. F.; Pluskal, M. G. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* **2003**, *787*, 19−27.
(49) Phizicky, E.; Bastiaens, P. I.; Zhu, H.; Snyder, M.; Fields, S. *Nature* **2003**, *422*, 208−215.
(50) Zhu, H.; Bilgin, M.; Bangham, R.; Hall, D.; Casamayor, A.; Bertone, P.; Lan, N.; Jansen, R.; Bidlingmaier, S.; Houfek, T.; Mitchell, T.; Miller, P.; Dean, R. A.; Gerstein, M.; Snyder, M. *Science* **2001**, *293*, 2101−2105.

(51) Paweletz, C. P.; Charboneau, L.; Bichsel, V. E.; Simone, N. L.; Chen, T.; Gillespie, J. W.; Emmert-Buck, M. R.; Roth, M. J.; Petricoin, I. E.; Liotta, L. A. *Oncogene* **2001**, *20*, 1981−1989.
(52) Grubb, R. L.; Calvert, V. S.; Wulkuhle, J. D.; Paweletz, C. P.; Linehan, W. M.; Phillips, J. L.; Chuaqui, R.; Valasco, A.; Gillespie, J.; Emmert-Buck, M.; Liotta, L. A.; Petricoin, E. F. *Proteomics* **2003**, *3*, 2142−2146.
(53) Craven, R. A.; Banks, R. E. *Proteomics* **2001**, *1*, 1200−1204.
(54) Li, C.; Hong, Y.; Tan, Y. X.; Zhou, H.; Ai, J. H.; Li, S. J.; Zhang, L.; Xia, Q. C.; Wu, J. R.; Wang, H. Y.; Zeng, R. *Mol. Cell Proteomics* **2004**.
(55) Aebersold, R. *Nature* **2003**, *422*, 115−116.
(56) Zhang, H.; Li, X. J.; Martin, D. B.; Aebersold, R. *Nat. Biotechnol.* **2003**, *21*, 660−666.
(57) Anderson, N. L.; Anderson, N. G. *Mol. Cell. Proteomics* **2002**, *1*, 845−867.
(58) Liotta, L. A.; Ferrari, M.; Petricoin, E. *Nature* **2003**, *425*, 905.
(59) West, J. L.; Halas, N. J. *Annu. Rev. Biomed. Eng.* **2003**, *5*, 285−292.
(60) Dubertret, B.; Skourides, P.; Norris, D. J.; Noireaux, V.; Brivanlou, A. H.; Libchaber, A. *Science* **2002**, *298*, 1759−1762.
(61) Akerman, M. E.; Chan, W. C.; Laakkonen, P.; Bhatia, S. N.; Ruoslahti, E. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 12 617−12 621.
(62) Dahan, M.; Levi, S.; Luccardini, C.; Rostaing, P.; Riveau, B.; Triller, A. *Science* **2003**, *302*, 442−445.
(63) Fritz, J.; Baller, M. K.; Lang, H. P.; Rothuizen, H.; Vettiger, P.; Meyer, E.; Guntherodt, H.; Gerber, C.; Gimzewski, J. K. *Science* **2000**, *288*, 316−318.
(64) Raiteri, R.; Nelles, G.; Butt, H.-J.; Knoll, W.; Skladal, P. *Sens. Actuat. B* **1999**, *61*, 231−217.
(65) Wu, G.; Ji, H.; Hansen, K.; Thundat, T.; Datar, R.; Cote, R.; Hagan, M. F.; Chakraborty, A. K.; Majumdar, A. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 1560−1564.
(66) Hansen, K. M.; Ji, H. F.; Wu, G.; Datar, R.; Cote, R.; Majumdar, A.; Thundat, T. *Anal. Chem.* **2001**, *73*, 1567−1571.
(67) Wu, G.; Datar, R. H.; Hansen, K. M.; Thundat, T.; Cote, R. J.; Majumdar, A. *Nat. Biotechnol.* **2001**, *19*, 856−860.
(68) Eng, J. K.; McCormack, A. L.; Yates, J. R., III *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 796−789.
(69) Perkins, D. N.; Pappin, D. J. C.; Creasy, D. M.; Cottrell, J. S. *Electrophoresis* **1999**, *20*, 3551−3567.
(70) Clauser, K. R.; Baker, P.; Burlingame, A. L. *Anal. Chem.* **1999**, *71*, 2871−2882.
(71) Field, H. I.; Fenyo, D.; Beavis, R. C. *Proteomics* **2002**, *2*, 36−47.
(72) Hernandez, P.; Gras, R.; Frey, J.; Appel, R. D. *Proteomics* **2003**, *3*, 870−878.
(73) Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R. *Anal. Chem.* **2003**, *75*, 4646−4658.
(74) Green, P. *Science* **1998**, *279*, 1115−1116.
(75) Orchard, S.; Kersey, P.; Hermjakob, H.; Apweiler, R. *Comp Func Genom* **2003**, *4*, 16−19.
(76) Orchard, S.; Kersey, P.; Zhu, W.; Montecchi-Palazzi, L.; Hermjakob, H. *Comp. Func. Genom.* **2003**, *4*, 203−206.
(77) Orchard, S.; Hermjakob, H.; Apweiler, R. *Proteomics* **2003**, *3*, 1374−1376.
(78) Davidson, E. H. *Genomic Regulatory Systems: Development and Evolution*; Academic Press: San Diego, CA, 2001.
(79) Davidson, E. H.; Rast, J. P.; Oliveri, P.; Ransick, A.; Calestani, C.; Yuh, C. H.; Minokawa, T.; Amore, G.; Hinman, V.; Arenas-Mena, C.; Otim, O.; Brown, C. T.; Livi, C. B.; Lee, P. Y.; Revilla, R.; Schilstra, M. J.; Clarke, P. J.; Rust, A. G.; Pan, Z.; Arnone, M. I.; Rowen, L.; Cameron, R. A.; McClay, D. R.; Hood, L.; Bolouri, H. *Dev. Biol.* **2002**, *246*, 162−190.
(80) Hong, J. W.; Quake, S. R. *Nat. Biotechnol.* **2003**, *21*, 1179−1183.

PR0499693