Functional genomics: technological challenges and opportunities

ROBERT L. STRAUSBERG1 AND M. J. FINLEY AUSTIN2

1National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892-2590; and
2Brigham and Women's Hospital, Boston, Massachusetts 02115

Strausberg, Robert L., and M. J. Finley Austin. Functional genomics: technological challenges and opportunities. Physiol. Genomics 1: 25–32, 1999.—In April, the Merck Genome Research Institute and the National Cancer Institute's Cancer Genome Anatomy Project, both supporters of functional genomics technology development and research, brought together a group of 27 scientists working at the forefront of this new field. Here we report on the presentations, discussions, and outcomes from this highly interactive and stimulating meeting held at the Banbury Center.

REMARKABLE NEW RESEARCH PLATFORMS have been established during the 1990s that will forever change the way we think about biomedical and biological research. At the start of this decade plans were just being established for the Human Genome Project, aimed toward assembling comprehensive genetic and physical maps of the genome and attainment of the complete DNA sequence of all human chromosomes. To date only a small portion of the human genome has actually been completely sequenced, but it is now expected that over the next few years most of that genome will be sequenced, with a "rough draft" available in the next year. Well before that sequence becomes available, a new vision of comprehensive molecular analysis, in which the functions of individual genes would be studied within the context of all the genes in the cell, is already being established as a key underpinning for biomedical research in the 21st century. This new vision has been captured under the umbrella term of functional genomics, with a variety of related subcategories such as transcriptome, proteome, and physiome.

The vision for this rapid paradigm change has been driven in large part by three scientific advances. First, the concept and then large-scale application of expressed sequence tag (EST) technology has resulted in the identification of tens of thousands of human genes. Currently the human UniGene database includes more than 71,000 clusters (potentially distinct genes) of ESTs and complete cDNAs. Similar work is under way to provide large gene catalogs for mouse and rat genomes. Most of the EST clusters represent previously unknown genes, providing fertile ground for scientific discovery.

Second, over the past few years the genomic sequences of various microbes, including model systems such as E. coli and S. cerevisiae, as well as pathogens including H. pylori and M. tuberculosis, have been determined. Most recently, the essentially complete sequence of a complex multicellular eukaryote, C. elegans, has been obtained. The third advance, computing technology and new bioinformatics methodologies, has allowed for the management of these enormous data sets and their rational mining. In silico analysis of the genomic sequences has revealed large numbers of previously unknown genes. With the complete coding repertoire now known for these (and many future) organisms, we can frame biological questions in a completely different contextual framework. No longer does one need to study a particular gene without knowledge of other potential partners in the networks in which that gene functions. Knowing all the genes immediately drives forward the quest to put together complete pictures of the genetic networks and to understand all of how the genome performs its business.

Thus there is now an enormous opportunity and challenge for the scientific community. Large-scale sequencing for gene discovery has now become routine and is becoming even more aggressive as sequencing technology improves. The challenge is to now apply genomic principles toward building complete catalogs not only of genomes, but of all of their products, and to learn how these macromolecules interface in a dynamic manner to produce complex cells and organisms. Assessing all of the components that contribute to the molecular anatomy and physiology of normal and disease cells and forming predictive hypotheses of changes involved in progressing through different cellular types, be they the formation of normal cells or ones involved in disease, now become the genomics opportunities for the 21st century.

ogy Development and Research Applications” held at the Banbury Center in April 1999. MGRI and CGAP have complementary missions related to developing and applying technology for functional genomics, and this conference represented just one of many joint efforts by the two groups.

The MGRI, a wholly separate not-for-profit corporation, was established with funding from Merck & Co., Inc., to drive the development of functional genomics technologies and to ensure the broad availability of these basic research tools to the entire scientific community. The focus of the MGRI is to enable scientists to develop assays and methodologies that can be applied broadly across genomics research with the objective of improving the accuracy and speed with which functional associations can be made with sequences of genetic information. Examples of research and infrastructure supported by MGRI include 1) targeting gene identification in tissues with disease associations and increasing the utility of the sequence data by identifying complete gene sequences (Merck Gene Index), 2) full mammalian cDNA cloning/sequencing technology, 3) bioinformatics (e.g., development of new algorithms to predict gene function based on sequence content), 4) disease models (e.g., development of methods that create gene-targeted interactions/mutations/expressional control in model organisms for the purposes of studying gene function), and 5) gene expression/function assays (e.g., development of technologies to rapidly assess whole genome gene expression).

The CGAP mission is the development of technology, information, and materials infrastructures toward the goal of comprehensive molecular analysis of cells as they transition from the normal state to the precancerous state and, ultimately, to cancer. For application of functional genomics principles, cancer biology and disease management provide enormous opportunities and challenges. The reason for these opportunities is that cancer is a very dynamic set of diseases and, likewise, the cancer genome is very dynamic, both with respect to actual physical changes in the genome and in alterations in gene expression as the disease progresses. Thus hereditary predispositions represent but one aspect of molecular changes contributing to cancer development. Most often, cancer is associated with somatic changes occurring during the lifetime of an individual. The changes can be of many forms, from point mutations to insertions and deletions, amplifications, and translocations. In addition, other forms of genomic change, such as DNA methylation, are also associated with the transformation to cancer. The genomic alterations are manifested in alterations in gene expression at the RNA and protein level, including differences in posttranslational modification such as phosphorylation.

The spirit of the genomics revolution was fully evident at the Banbury Conference. This was a gathering of a highly interdisciplinary group of scientists, pursuing many different aspects of functional genomics. Perhaps most importantly, there was open sharing of thoughts about not only the enormous progress being made, but also the even greater challenges that lie ahead.

DNA chip technology. Several of the presentations were focused on the development and application of DNA array (chip) technologies. From a technological and philosophical viewpoint the spirit of the genomics revolution is perhaps best captured in the application of technology based on advances in the computer industry to the development of DNA chips. In DNA array technology, molecules are either synthesized on the chip, or presynthesized and then deposited. The DNA molecules range in size from short oligonucleotides (25mers) to cDNAs, to bacterial artificial chromosome (BAC) clones with inserts of ~100 kilobases. These DNA probes can then be used to interrogate unknown target sequences based on specificity of hybridization to the known probes. Although this technology was first envisioned for application in DNA sequencing, the creativity of the community has been harnessed to broaden the utility of these chips for many aspects of functional genomics. This was quite evident at the Banbury Conference, in which approaches to assessing changes in the genome itself or in its expression were the subject of several presentations.

Scientists at Affymetrix Corporation and Stanford University developed two of the pioneering approaches in this field. In the Affymetrix approach high-density oligonucleotide arrays are synthesized on chips by photolithographic technology. For measurements of transcript expression a series of 20 oligonucleotides spanning the known sequence and 20 additional partner oligonucleotides with one-base mismatches to the target sequences are synthesized, hybridization is measured across the oligonucleotide population, and algorithms are employed to measure expression levels for each transcript. One of the great strengths of the Affymetrix approach is that sequence information in databases is used to design on-chip synthesis. Thus there is no need to maintain large collections of cloned DNA molecules. In addition, because the oligonucleotides are relatively short and can be designed for any gene region, the technology can be applied to sequencing, identification of polymorphisms, and potentially for identification of different transcript splice variants. A limiting factor for the technology is that, because photolithography is employed, the average laboratory cannot synthesize its own chips, and the use of photolithographic masks leads to relatively slow turnaround time for development of new chips.

In the Stanford approach, cDNA molecules are spotted robotically on glass slides, and changes in gene expression are measured by labeling a control and experimental transcript population with different fluorescent tags and then measuring the intensity and ratios of the fluorescent signals. This approach has gained much popularity, in part because the technology has been disseminated widely, and the approach is well suited to rapid design and synthesis of new arrays with different sets of cDNA probes.

Gene expression analysis. The power of the cDNA approach for gene expression analysis was convincingly
demonstrated by Paul Spellman (Stanford University) in his very successful effort to catalog yeast genes whose expression is correlated with changes in the cell cycle (4, 19). Using the power of yeast biology/genetics with the DNA arrays, Spellman identified more than 800 genes regulated in a cell cycle-dependent manner. An important principle fully evident from Spellman's presentation is that DNA array experiments need to be addressed in an integrated manner, including biological sample preparation, assay and detection systems, and informatics tools. In keeping with genomic principles of open data dissemination, the complete data sets and analysis tools from these experiments are available at http://cmgm.stanford.edu/pbrown/index.html.

Louis Staudt (NCI) demonstrated the potential power of the cDNA array approach as applied to the study of cancer, in particular gene expression in lymphomas. In this case a cDNA array termed the lymphochip (with cDNAs derived from lymphocytes and precursors to lymphocytes) is being used to profile gene expression in lymphomas and leukemias. In addition to providing basic insights to lymphocyte differentiation, it is hoped that molecular profiling of lymphomas can be achieved and that these profiles can be used to differentiate lymphomas that appear similar but that each have a different molecular basis for disease development. Staudt showed preliminary evidence that such hopes for molecular profiling of these cancers may soon become a reality.

The utility of microarray analysis of gene expression for the study of cancer was further demonstrated by Uri Alon (Princeton University). He presented results obtained from the Levine laboratory when expression was compared from 42 colon tumor tissue samples versus 20 matched normal tissues using Affymetrix technology (data available at http://www.molbio.princeton.edu/colondata). To organize and analyze this massive data set they employed a two-way cluster analysis and developed a novel pattern-visualization method. This allowed one to easily see the striking differences in expression between the two sample sources. They plan to create an open Web-based database with an interactive format linked to the analytic tools.

In his presentation, Steven Gullans (Brigham and Women's Hospital) described his effort to build an expression database based on application of Affymetrix chips. This will become a publicly accessible database (http://www.geneindex.org) profiling gene expression in a wide variety of normal human tissues. Establishment of the normal variation in gene expression will be key to teasing out which changes are meaningful in pathophysiology and which are typical fluctuations, thereby allowing researchers to more rapidly identify genes for more in-depth study.

Jeffrey Green (NCI) is focused toward understanding molecular mechanisms of mammary and prostate tumor development in transgenic mice and applying that knowledge toward identification of potential therapeutic agents. In particular, his laboratory has developed the first transgenic mouse model for prostate cancer (through expression of SV40 Tag in the prostate controlled by the rat c3(1) prostatein regulatory region (18)). Interestingly, female mice with the same genetic construction develop mammary tumors. Green spoke of the application of DNA array technologies to assess the transcriptional changes associated with tumorigenesis in these models. He also described plans for the NCI mouse Cancer Genome Anatomy Project, designed to identify genes and profile transcriptional changes associated with tumor development, and the potential for interfacing those data with gene profiles for corresponding human cancers.

Gary Churchill (The Jackson Laboratory) discussed the critical need for good experimental design in array expression experiments to ensure data are reliable and meaningful. In an analysis of expression data generated from cDNA arrays he found that 7–8% of the variation in the data is noise. He discussed ways to factor out this noise through good design: for example, the need to duplicate clones within a chip to account for variation in hybridization that occurs at different locations on the chip and the need to replicate experiments on multiple chips. Furthermore, he suggested that by using designs with replication, one could derive so-called shrinkage estimators of expression levels. Shrinkage estimators are derived by a Bayesian statistical model and take advantage of all the information available in the many thousands of hybridizations on a typical chip.

Gregory Riggins (Duke University Medical Center) described an alternate approach to gene expression analysis termed serial analysis of gene expression (SAGE) (20). In this approach short sequence tags unique to each transcript are generated and concatenated. Upon sequence analysis, approximately 25 or more independent tags can be identified from each sequencing lane. An important aspect of the SAGE technology is that it is not limited to genes already discovered or inferred from informatics analysis. Indeed, this technology was used to identify hundreds of yeast transcripts not predicted by the currently available informatics tools (21). Riggins described the public CGAP/SAGE database (http://www.ncbi.nlm.nih.gov/SAGE/) as one model for collaborative expression genomics and rapid sharing of gene expression data in the community.

One of the key issues with respect to analysis of gene expression is that for complex tissue systems: the cell populations are very heterogeneous, and therefore the DNA array readouts represent an average of many different cell types. Although this may suffice for many applications, it is not ultimately what will be needed to fully apply genomic principles to normal human development and disease processes. For example, it may be possible (it remains to be seen) to do molecular pathology of cancers based on bulk samples. But to get at the earliest stages of cancer development and understand specificity of gene expression in cells within a tissue (or tumor), assessment of gene expression in homogeneous cell populations will be needed.
In that regard, Michael Emmert-Buck (NCI) discussed his laboratory's goal of developing an integrated three-dimensional molecular model of the prostate gland (3). This vision includes production of transverse cross sections of whole prostate glands, with subsequent microdissection and molecular characterization at the levels of the genome, transcript population, and proteome. Key to the success of such an approach is the development of new sample preparation methods that ensure that the profiles truly reflect the in vivo situation.

In a related approach Gregor Eichele (Baylor College of Medicine) described his efforts in developing and applying RNA high-throughput in situ hybridization within localized regions of tissue sections. He is currently applying this approach toward generation of a database of gene expression in the mouse brain with the goal of assembling a gene activity map of the mouse brain. Clearly, this approach could be closely interfaced with array-based gene expression profiling efforts to gain additional insights into specificity of gene expression in specific tissue regions and different cell types.

Assessing changes in genomic DNA. Additional applications of DNA microarrays were discussed by Jon Pollack (Patrick Brown's Laboratory, Howard Hughes Medical Institute, Stanford University) and Vivian Cheung (Children's Hospital of Philadelphia, http://genomics.med.upenn.edu/vcheung). In their research the applications of DNA arrays are extended to characterization of changes or similarities in genomic DNA. In the Pollack work, human cDNA arrays have been applied to detecting gene amplifications and deletions in tumors and for measuring changes in gene expression from the same samples. In her presentation, Cheung demonstrated the use of DNA arrays for comparative genomic hybridization (CGH) as applied to gene mapping.

Extending the horizons for DNA array technologies even further, Dan Pinkel (University of California at San Francisco) and Vivian Cheung described their efforts to build arrays based on BAC clones. The goal of these projects is to build BAC DNA arrays (covering the genome at 1-megabase intervals) faithfully representing the human genome. In the Pinkel approach the BAC arrays are being developed for CGH. A key performance improvement based on the application of BAC arrays to CGH is the very high resolution that would be attained by using sets of overlapping BACs to give resolution less than the length of the clone. The measurement precision using BACs for CGH permits determination ratios with a standard deviation of <10%, so that low-level copy number changes affecting a single clone in a large array can be detected with high statistical confidence. Cheung is producing a 1-megabase whole genome BAC map that can be applied to 1) genome mismatch scanning for identity by descent studies (localization of disease alleles arising from a common origin) (2), 2) chromosomal mutation detection, and 3) physical mapping.

David Muddiman (Virginia Commonwealth University, http://saturn.vcu.edu/~dcmuddim/) presented a novel and potentially high-throughput approach to the characterization of alterations in DNA based on methylation of cytosine residues of CpG islands through the application of electrospray ionization-Fourier transform ion cyclotron resonance-mass spectrometry (ESI-FTICR-MS) (9). This type of methylation has been implicated in several disease processes, including myotonic dystrophy and fragile X syndrome, genetic diseases related to expansions of trinucleotide repeat sequences, or variable-number tandem repeats (VNTRs). Hypermethylation is identified in affected individuals and not in individuals carrying the wild type or premutation. In Muddiman's approach, methylation-specific PCR incorporates bisulfite conversion of unmethylated cytosines to uracils in the template DNA with subsequent conversion of uracil to thymine during in vitro amplification, and the 15-Da difference between cytosine and thymine is detected by ESI-FTICR-MS. In this approach, locations of methylation can be precisely located and correlated with patterns of repeats, and individuals can be typed for number of trinucleotide repeats at a particular locus. Although obtaining informative spectra from large oligonucleotides using ESI-FTICR-MS is challenging, Muddiman showed a spectrum of a 500-bp PCR product with a mass precision of 0.008%.

Debbie Nickerson (University of Washington) focused her remarks on the challenges in identifying and using single-nucleotide polymorphisms for the study of complex biological traits. She discussed applications of a computer program called Polyphred for automated detection of heterozygous sequences among PCR products analyzed by fluorescence-based DNA sequencing (http://droog.mbt.washington.edu). This program builds on a series of other programs developed by Phil Green, including Phred (base calling), Phrap (sequence assembly), and Consed (editing), to form an integrated and robust system for detection of polymorphisms. Nickerson also focused her presentation on the challenge of assigning a particular phenotypic effect to a single nucleotide polymorphism, even in cases where a single gene has been implicated. In that regard she pointed toward a study of the DCP1 gene encoding angiotensin converting enzyme, in which 17 polymorphisms were in linkage disequilibrium with an Alu element previously correlated with cardiovascular disease (16). Assigning phenotype to any one of these sites (if indeed one of them is responsible) thus becomes quite difficult. The challenges with this single gene example point to the difficulties in pinpointing responsible sites in multigenic diseases with multiple contributing factors.

New approaches to DNA array assembly. Particularly exciting are completely new approaches to synthesis of DNA arrays to allow for onboard DNA synthesis while providing for rapid changes and enhanced flexibility in generating the arrays. Xiaolian Gao (Univ. of Houston, http://www.chem.uh.edu/www/gao/intro.html) described an approach incorporating a programmable photolithographic system, new oligonucleotide deprotection chemistry using photogenerated acids (6), and
newly developed synthesis microreactors. Key to the approach is the replacement of photomasks, production of which requires a time-consuming and expensive process for oligonucleotide chip development. Furthermore, the design and chemistry is not limited to DNA or RNA synthesis but could be used to make peptide arrays, as well.

In a similar technological direction, Skip Garner (University of Texas Southwestern Medical Center) is developing a system for the manufacture of high-density DNA arrays using digital optimal chemistry that employs ultraviolet photochemistry based on the Texas Instruments digital light processor (DLP). The DLP houses millions of mirrors under computer control. This system also alleviates the need for photolithographic masks so that new arrays with different sets of oligonucleotides can be rapidly designed and built. It is interesting and very much in the genome tradition that both the Gao and Garner approaches are in collaboration with scientists at the Display Technology Center at the University of Michigan and Texas Instruments, respectively. These types of interdisciplinary, academic/industrial collaborations among nontraditional partners are likely to rapidly change technological capabilities to meet ever-increasing biological/biomedical needs.

Identifying cellular proteins. Philip Andrews (Univ. of Michigan, http://www.proteome.med.umich.edu) and Ruedi Aebersold (Univ. of Washington, http://weber.u.washington.edu/~ruedilab/aebersold.html) each discussed new approaches for proteome analysis. The vision put forth in their presentations is to develop technologies for ultra-high-throughput analysis of at least the majority (hopefully all) of the proteins present in cells. Important to recognize in this vision is that proteomics is very dynamic and that we are interested in identifying not only the proteins in the cells, but also their functional states, posttranslational modifications, quantity, and rate of synthesis and turnover. Characteristics and components of these systems (in addition to being high throughput) include high-sensitivity analytical methods, high-resolution separation methods including prefractration, quantitation, and new computational approaches.

Particularly important to the Andrews and Aebersold approaches are technology advances in 1) mass spectrometry, in particular matrix-assisted laser desorption/mass spectrometry (MALDI) and electrospray ionization tandem mass spectrometry (ESI-MS/MS), 2) the interface of MALDI-MS and ESI-MS/MS with gel-based separation methods, 3) high-speed, sophisticated data-processing routines, and 4) data search algorithms based on correlating peptide masses and collision-induced dissociation spectra acquired by tandem mass spectrometry with predicted protein/peptide masses derived from genome and EST DNA sequencing databases. Andrews (1, 13–15) discussed the notion of virtual two-dimensional gels. These are generated by mass analysis directly from isoelectric focusing (IEF) gels and by comparing the experimental data with similar theoretical pseudogels calculated from complete DNA sequence information already available for many microbes and soon for several multicellular eukaryotes. He emphasized the need for new software to manage and interpret the truly massive amounts of heterogeneous data generated by these types of functional genomics approaches. Aebersold (7, 8) discussed limitations of current proteome technology with respect to measuring quantitative changes in protein expression and the detection of low-abundance proteins and introduced a new chemistry termed “isotope-coded affinity tags” for quantitative proteome analysis.

George Church (Harvard Medical School) took on a rather herculean task trying to synthesize many of these approaches and spoke on integrating measurements, motifs, and models for comprehensive molecular quantitations of cell populations. His presentation ranged from describing his biomolecule interaction, growth, and expression database (BIGED) found at http://arep.med.harvard.edu, to proteomics, to new technology for spotting high-density DNA microarrays on the fly, to the development of fluorescent in situ sequencing on microarrays. He impressed on the audience the importance of remembering that noncoding regions of the DNA are significant. Further, their study is a challenge because typically one is dealing with interacting proteins of low abundance and small nuclear RNAs.

Structure-based functional genomics. Two of the greatest challenges in functional genomics are to assign potential protein functions and to understand which proteins may perform related activities. Often, homologous proteins are identified based on translations of DNA sequences and comparisons of the primary amino acid sequences. However, often this does not work, in part because similar protein structures can be derived from primary amino acids sequences that are quite dissimilar. Thus there is a great need to be able to perform high-throughput protein structure analysis and to use these data to identify proteins that are functionally related.

Kristin Gunsalus (Rutgers University) described her efforts in Gaetano Montelione's laboratory in building tools for structure-based functional genomics (http://www.nmr.cabm.rutgers.edu/structuralgenomics/). In their approach NMR spectroscopy is applied to understanding relationships between protein structure and function. The laboratory is automating analysis of protein NMR data with the idea that, through structure determination and application of new bioinformatics methods, biochemical functions can be discovered for previously unknown proteins discovered in genome sequencing efforts (12).

In an enlightening presentation, Jeff Skolnick (Scripps Institute) demonstrated just how much can be derived about protein structure and function using purely informatics approaches (5). He summarized work with threading, which uses structural analogy to predict protein folds, as well as his method for ab initio fold prediction. He also showed how such low-resolution structures can be used to predict the biochemical function of proteins. Further, he underscored the challenges and need for broader, more complete data sets of
structural determinations because proteins can have different structures but similar functions.

Michael Yaffe (Harvard Institute of Medicine) described his project in the Cantley laboratory to create an interactive protein signature motif database using peptide library information and to use this database as a tool toward elucidating protein function. Yaffe provided examples of the laboratory's overall strategy based on the use of partially degenerate peptide libraries for determination of optimal motifs in serine kinases and tyrosine kinases, as well as Src-homology domains. In particular, based on primary amino acid sequences, he was able to determine specifically which SH2 domains interface with particular signaling molecules. The results of these studies are to form the starting basis of an interactive World Wide Web protein signature motif database.

Protein networks. The study of protein-protein interactions was discussed actively by several of the participants. In particular, development of technology for and application of the yeast two-hybrid system was discussed extensively. In the two-hybrid system, proteins or protein domains are expressed as fusions with either a DNA-binding domain or an activation domain. If the proteins linked to the DNA-binding and activation domains interact, transcription of a reporter gene encoding a selectable phenotype results. The presentations focused on developing protein interaction maps for the entire sets of proteins produced by yeast (Peter Uetz), C. elegans (Marc Vidal), and Drosophila (Russ Finley) and the expression of human proteins in yeast to identify binding partners (Erica Golemis).

Peter Uetz (Univ. of Washington) described the Yeast Protein Interaction Map Project, an effort to examine all potential protein pairs of yeast with a 6,000 × 6,000 matrix (http://weber.u.washington.edu/~sfIELDS/projects/YPLM/). In this project, all yeast open reading frames (ORFs) are expressed in a GAL4 activation domain vector. The entire set of strains are then mated to a strain expressing a yeast ORF linked to the GAL4 binding domain (product is 6,000 diploids), and the diploids are tested for the selectable marker. The end result would be a database of all of the yeast protein-protein interactions (and noninteractions), as observed through this approach.

Russ Finley (Wayne State Univ. School of Medicine, http://cmmg.biosci.wayne.edu/rfinley/lab.html) described his plans for a similar project to map all protein interactions for a more complex organism, Drosophila. Of course, in Drosophila the number of genes is greater (~15,000), so the matrix for examining all interactions expands significantly. In his discussion, he stressed the need to verify interactions observed through the two-hybrid approach and to interface these data with other approaches aimed at characterization of biological function. In that regard he discussed application of a related technology based on identification of peptides that bind to specific proteins and that are able to disrupt interactions with binding partners (aptamers) (10). The aptamers can be expressed in vivo such that phenotypes derived from blockage of particular protein-protein interactions can be assessed. To demonstrate the utility of this strategy toward assigning function, Finley showed an example of the in vivo expression of an aptamer targeted to a protein regulator of cell divisions required for development of the Drosophila eye, with resultant developmental abnormalities.

A program to map protein interactions in the nematode C. elegans, through application of the two-hybrid system, was introduced by Marc Vidal (Massachusetts General Hospital Cancer Center) (22). He introduced the term interaction sequence tag (IST), defining a pair of two short sequences derived from cDNA encoding interacting proteins, and showed how these tags are being integrated in the C. elegans database (AceDB) in a format very similar to the ESTs. Thus protein-protein interaction data would appear in the same format as other C. elegans genomic resources. In his presentation he also emphasized the need to validate associations discerned from the two-hybrid data. He discussed plans to validate the interactions through the use of interaction-defective alleles (IDAs), generated in a reverse two-hybrid system, defined as amino acid alterations that specifically disrupt protein-protein interactions.

Erica Golemis (at the Fox Chase Cancer Center, http://wwwfcc.edu:80/research/labs/golemis/EG_homepage.html) (11, 17) studies regulation of cell shape control in S. cerevisiae through the identification of signaling proteins. Many of the yeast signaling proteins have homologs in mammalian systems, including human. In several cases these mammalian genes have been identified as oncogenes and antioncogenes. Her laboratory is interested in understanding common principles in yeast and mammalian cell shape control and division. To study the interspecies relationships further, she expressed a library of human cDNAs in yeast and identified human genes that change the morphology of the yeast budding pattern to pseudohyphal. Her application of the two-hybrid system has been to identify potential binding partners for these common yeast-mammalian proteins. For example, she identified a binding partner of the human Krev-1 gene termed Krit1, which also may be a cell shape regulator. The Golemis studies demonstrate the power of first interfacing complex mammalian biology with that of simpler systems such as yeast in order to identify genes contributing to a particular function, and then using the power of the yeast two-hybrid system to extend the network of proteins contributing to that function.

Overall the presentations of the two-hybrid technology demonstrated that this system will likely make very significant contributions to our understanding of protein-protein interactions. At the same time, there was much discussion of false positives and false negatives in the data sets and the need to continue to improve as well as interface with other approaches for defining interactions. The participants discussed the need to very much limit the false positives, as these could needlessly lead researchers in the wrong direction. Indeed, it will be a challenge for functional genomics in general to capture large data sets from
complex biological systems and to provide indications of relative confidence in the observation (quality scores).

In conclusion, Mark Boguski (National Center for Biotechnology Information) provided the wrapup and reflections at the end of the two and half days of presentations. He summarized the current state of genomics information and discussed the need to think about new ways of incorporating this information and approaching biological research. He reminded us that there are currently more than 10 million articles in Medline with 400,000 added each year from 4,000 journals. Of these, greater than one million are indexed as molecular biology and genetics. Further, in GenBank there are more than 3.5 million sequences and 2.6 billion bases from more than 41,000 species, including 301 megabases of finished human sequence. This underlines the enormity of the task at hand—to handle, utilize and learn from all of the information being generated—leading to the conclusion that the small cottage industry style of research will need to restructure to take full advantage of genomics and the resulting multidisciplinary and dual-use technologies.

Probably no scientific field has ever started with greater expectations than functional genomics. The ability to look at cellular, tissue, and maybe organismal molecular biology in a comprehensive manner opens up an entirely new way of approaching biomedical science and in fact is transforming all biological sciences. It truly will change the way we predict, diagnose, and treat disease, as well as how we develop new therapies and create novel therapeutic modalities. With that said, the technical challenges are also enormous, because in order to accomplish the ultimate vision, we need to apply the new vision to biological systems that each bring their own complexities. So we don't only want to look at an organ as the sum of its cellular components; we want to have comprehensive molecular knowledge of all of the component cells and how they interface to produce overall phenotypes. For example, with cancer we need to have complete knowledge not only of the epithelial cells that give rise to cancer but of their molecular changes as they become precancerous and cancerous.

We also need to have comprehensive molecular knowledge about supporting tissues, including endothelial cells and distal sites, such as tissues that support metastasis. Beyond those sites directly related to tumorogenesis, we want to know about the complete genetic makeup of the individual in order to identify other factors contributing to cancer development and affecting outcome.

Moreover, we want this information from many patients in order to understand individual differences, the multitude of ways that a particular type of cancer, or any disease, can arise, and how we can use these data to effectively prevent or treat all diseases at their earliest stages. In addition, at the meeting it became very clear that the building blocks of functional genomics might need to be reduced even further than initially anticipated. For example, it may be necessary to do two-hybrid analysis based on individual protein domains, not just the complete proteins, to enhance our ability to detect interactions.

Perhaps the most challenging scientific aspects of functional genomics will be related to informatics (databases, analysis tools, in silico modeling) and how to capture the biology in those databases. Defining the starting sample and sample preparation methods and capturing the actual biological experiment (or, for example, all of the critical issues related to patients and their treatment in clinical trials) will certainly constitute extremely challenging aspects of the databases. Even in the simplest cases, such as the wonderful model S. cerevisiae, it will be important and challenging to capture all of the biological information, including experimental protocols, in database format. Moreover, we will need seamless interfaces in databases, so that one can ask very simple or complex questions and get back the desired data, which will often come from different studies, captured in different databases.

At the same time, the scientist's greatest challenge will be in defining new ways to interact, collaborate, generate, and share this information. Without sharing of fundamental tools and knowledge, the ability to harness this biological revolution will be stifled. Furthermore, many of the "wish-list" projects, by the very nature of their size, can only be accomplished through large collaborations. Biologists will not only need to learn new ways to study interactions, but also to learn new ways to interact. This will challenge the funders and institutions as well to design programs and structures that support these activities.

Although all of this may seem daunting, in fact there is incredible excitement about the infant field of functional genomics. The Banbury Conference captured that spirit very well and facilitated very open discussions on the technical challenges with various approaches. In addition, the interdisciplinary nature of the conference and diversity of scientific goals all under the umbrella of functional genomics made for intriguing scientific exchange and highlighted the utility of interdisciplinary approaches. Indeed, we will not be surprised to see some new joint efforts arising from this conference's participants. The start of genomic research has provided for an exciting end to 20th-century biological research. The discussions at this conference suggest that there will be no lack of excitement as we begin the new millennium.

Address for reprint requests and other correspondence: R. L. Strausberg, Director, Cancer Genome Anatomy Project, and Assistant to the Director, National Cancer Institute, Bldg. 31, Rm. 11A03, 31 Center Dr., MSC 2590, Bethesda, MD 20892-2590 (E-mail: RLS@nih.gov).

REFERENCES


