Annu. Rev. Pharmacol. Toxicol. 2000. 40:193–208

THE IMPACT OF GENOMICS ON DRUG DISCOVERY

C. Debouck and B. Metcalf
Discovery Chemistry & Platform Technologies, SmithKline Beecham Pharmaceuticals, Research & Development, King of Prussia, Pennsylvania 19406; e-mail: Christine_M_Debouck@sbphrd.com, Brian_Metcalf@sbphrd.com

Key Words EST, cathepsin K, G-protein coupled receptors, microarrays, proteomics

■ **Abstract** High-throughput gene sequencing has revolutionized the process used to identify novel molecular targets for drug discovery. Thousands of new gene sequences have been generated but only a limited number of these can be converted into validated targets likely to be involved in disease. We describe here some of the approaches used at SmithKline Beecham to select and validate novel targets. These include the identification of selective tissue gene product expression, such as for cathepsin K, a novel osteoclast-specific cysteine protease. We also describe the discovery and functional characterization of novel members of the G-protein coupled receptor superfamily and their pairing with natural ligands. Lastly, we discuss the promises of gene microarrays and proteomics, developing technologies that allow the parallel analyses of tissue expression patterns of thousands of genes or proteins, respectively.

THE PARADIGM SHIFT—FROM GENE TO SCREEN

Traditionally, the identification of molecular targets in drug discovery has been biologically driven. An interesting enzyme or receptor activity implicated in normal physiology or disease was characterized and isolated, usually from animal tissues. Micropurification, monitored by functional assays such as radioligand binding or other techniques, led to the cloning of the gene encoding the target of interest. Following the expression of the gene in a recombinant host, the desirable activity was confirmed and used to run a high-throughput compound screen or to support rational drug design. This ‘‘from-function-to-gene’’ process was time consuming, but it delivered defined targets whose function was understood. The advent of high-throughput gene sequencing resulted in the rapid identification of thousands of novel genes, most without known function. As a result, the drug discovery scientists have had the challenging task of leveraging genes of unknown function into attractive therapeutic targets, a paradigm shift often referred to as the ‘‘from-gene-to-screen’’ process. We describe here a number of approaches

0362–1642/00/0415–0193$14.00

applied in SmithKline Beecham (SB) laboratories to identify and validate novel therapeutic targets from the wealth of gene sequence information currently available.

EST Sequencing

High-throughput sequencing was first applied to complementary DNA (cDNA) libraries by Adams et al (1) and has since rapidly led to a buildup of expressed sequence tags (ESTs) in corporate and public (2) databases. By definition, this approach gives partial sequences of expressed genes from the various cells and tissues used for the cDNA library preparation, but it does not include the sequence of the intervening noncoding DNA. This ''junk'' DNA constitutes about 97% of the human genome. Although these ESTs are only partial sequences of 200–600 nucleotides, bioinformatic approaches are able to assemble ESTs derived from the same gene through the identification of overlapping fragments. Assemblies of EST (contigs) are then created, giving an estimate of the number of expressed genes (as distinct from gene fragments) in the library under consideration. Taken to their logical conclusion, such approaches could eventually identify the complete repertoire of expressed human genes and, thereby, an extremely large number of potential human drug targets. The full complement of the human genome is estimated to be 80,000–100,000 genes. Analogously, application of high-throughput sequencing techniques to microbial genomes has made possible the sequence determination of several entire bacterial genomes. This wealth of sequence information will allow researchers to characterize shared pathways and metabolic networks, and to identify all possible targets for intervention in infection control (3).

"Smart" Libraries

The approach initially taken by scientists at SB is illustrated in Figure 1. Here, ''smart'' cDNA libraries are prepared from various sources likely to either play a role in disease etiology or be enriched in cell surface or secreted proteins. For example, libraries were prepared from tissues of particular therapeutic interest, such as prostate, kidney, left heart ventricle, bone marrow, or subsections of the brain, or from cell lines where an activated form can be compared with a resting form. In collaboration with Human Genome Sciences, we constructed more than 500 cDNA libraries, subjected them to random sequencing from the 5′ end, searched for sequence homology, and assembled contigs. After first-pass sequencing of 500 ESTs, the libraries that were deemed high quality were further sequenced, to a depth of 2,000–10,000 ESTs. Because many genes are expressed at low level, this type of ''shallow'' sequencing is expected to identify genes that are moderately to abundantly expressed in the tissue under study. The relative abundance of ESTs for a given gene in a given library can be a selection criterion for choosing a novel gene as a molecular target, as illustrated below for cathepsin K. On the other hand, random EST sequencing, much like the lottery, can draw

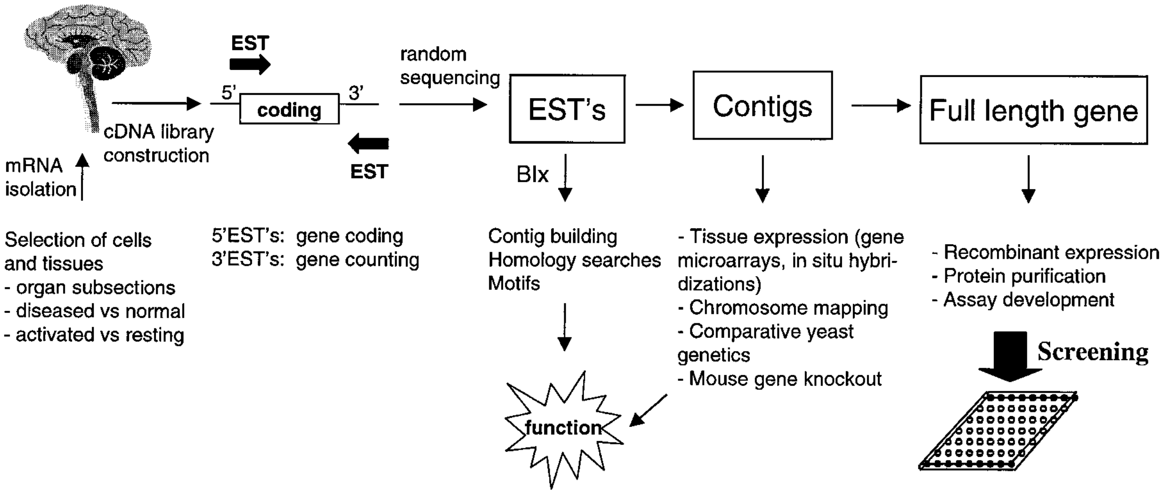

Figure 1 Outline of the gene-to-screen process. The main experimental steps for the identification and validation of novel drug targets are depicted and discussed further in the text. EST, Expressed sequence tags.

some low-abundance, important novel genes, such as the orphan G-protein coupled receptors also discussed in this chapter. In order to reduce the continuing identification of abundantly expressed clones and thereby increase the probability of identifying distinct novel genes, including those expressed at low levels, our group and others made serious attempts to construct normalized cDNA libraries. However, sequence analysis of such libraries revealed that normalization only resulted in limited removal of redundant genes and did not justify the investment of additional time and effort (4).

Identifying Novel Protein Sequences

ESTs are obtained by single-pass sequencing of the ends of cDNA inserts using library vector-specific primers, thereby affording sequence information at higher speed and lower cost compared with gene-specific primers. Thus, sequencing can be conducted from the 5′ or the 3′ end of the inserts or from both ends, and this choice impacts the potential for novel gene discovery. Because reverse transcriptase synthesizes cDNA starting from the 3′ end of the mRNA, all cDNA inserts derived from the same gene have the same 3′ EST sequence. This feature is useful when the primary interest is gene counting to determine the gene expression profiles in a given tissue. However, because 3′ untranslated regions can be long, 3′-derived EST sequences often lack protein coding information. On the other hand, reverse transcriptase tends to stall and fall off the mRNA during cDNA synthesis and will often not reach the 5′ end of the mRNA, producing incomplete and ragged 5′ ends from a given gene. As a result, 5′-derived EST sequences typically contain protein coding information, and those derived from the same gene tend to overlap with one another, leading to the generation of in silico full-length gene sequences. For this reason, we focused our efforts on 5′ EST sequencing to favor the novel gene discovery process. We also felt that electronic gene counting would not be accurate because of the libraries not being sequenced in depth and that it would be supplanted by high-throughput ‘‘wet’’ gene expression profiling technologies (see below).

The protein sequences translated from the EST contigs often represent enough coding sequence for the molecular function of the gene to be recognized by homology to known human, animal, or other available sequences, for example a cysteine protease, a zinc finger, a G-protein coupled receptor, etc. In addition, appropriate probes can be derived from these partial sequences to carry out tissue distribution studies, e.g. by in situ hybridization. When coupled with the source of the cDNA library and the putative function of the gene product, enough information may be obtained to suggest further interest. The full-length gene is then cloned and expressed in an appropriate recombinant host. It should be noted that although the molecular function of the gene may be surmised at this stage, the physiological function is by no means certain. Drug discovery programs that are commenced based on this information have some degree of risk, owing to lack of knowledge of the physiological role of the selected target. Further work, such as construction of transgenic animals, can be undertaken to further underpin the importance of the target.

Millions of ESTs have been generated (e.g. 1,048,756 at http://www.tigr.org/ and 1,505,046 at http://www.ncbi.nlm.nih.gov/dbEST/) and subjected to extensive bioinformatic analyses, including assembly into contigs, and nucleic acid and protein homology searches. The challenge for drug discovery scientists is to identify those genes that play critical roles in normal physiology or in the etiology of diseases, and to elucidate their function both biochemically and biologically. This is then followed by the development of assays for high-throughput screening and the establishment of the relevant animal models for preclinical testing of compounds. Sequence homology and tissue distribution of a novel gene are two critical pieces of the large jigsaw puzzle of its function. Other high-throughput as well as more specialized, low-throughput techniques need to be deployed to add more pieces to the puzzle and to start discerning the picture of their role in normal physiology or disease. We describe below examples of how scientists at SB have selected novel therapeutic targets starting from hundreds of thousands of ESTs: the protease cathepsin K and the orphan G-protein coupled receptor superfamily.

CATHEPSIN K—A PROTOTYPIC GENOMICS-DERIVED DRUG DISCOVERY TARGET

If one were interested in discovering a treatment for osteoporosis, for example, a logical starting point in the context of genomics would be the osteoclast, a highly specialized cell that is responsible for bone resorption during the bone remodeling process. An imbalance between bone resorption and formation results in pathological states such as osteoporosis. Clearly one would make a cDNA library from the human osteoclast (see Figure 1). Because there is no cell line representative of the osteoclast, the SB approach was to use surgical specimens of osteoclastoma tumor, from which osteoclasts could be extracted using osteoclast-specific antibodies attached to magnetic beads (5). Upon sequencing a cDNA library prepared from human osteoclasts, it was found that approximately 4% of the ESTs generated encoded a novel cysteine protease, which was named cathepsin K (6). ESTs encoding other cysteine proteases were rare in the osteoclast library. Subsequent studies at both the message and the protein level demonstrated that cathepsin K was expressed exclusively in osteoclasts. Indeed, in situ hybridizations showed high levels of expression in osteoclasts but not other bone cells, whereas immunocytochemistry revealed a polar distribution of the enzyme right at the site of contact between the osteoclasts and the bone resorption pit, which suggests proteolytic digestion of the bone ‘‘in the making.’’

A research program was started involving the cloning and recombinant expression of full-length cathepsin K and the detailed characterization of its proteolytic properties (7). At the time, the conceptual underpinnings of the decision to commit to a program that eventually could cost as much as $400 million (the cost of development of a new drug) and to initiate a drug discovery program were limited to the following: The target is highly expressed in a human cell of interest, the osteoclast; the target is selectively expressed in the osteoclast; and the target is a new cysteine protease. It was assumed that inhibition of cathepsin K would lead to treatments for osteoporosis. Starting extensive drug discovery efforts with such a paucity of information seemed a high-risk strategy. Hence, firmer demonstration of the involvement of cathepsin K in bone resorption was important.

Pycnodysostosis—The Human Cathepsin K Knockout

When efforts on cathepsin K were under way, the mutations causative for the rare inherited osteochrondrodysplasia, pycnodysostosis, were localized to the cathepsin K gene. Affected individuals present with osteopretrosis and bone fragility. In humans, three separate mutations that could lead to loss of function of cathepsin K have been identified. Osteoclasts from such individuals demineralize normally but do not degrade the bone protein matrix. Cathepsin K was thereby confirmed to be the major protease responsible for degradation of the bone matrix (8). It should be noted that cathepsin K–deficient mice have since been constructed. These also exhibit an osteopetrotic phenotype and produce osteoclasts impaired in bone resorptive activity (9, 10).

The results of this human genetic study (and the mouse knockout) put the drug discovery program aimed at inhibiting cathepsin K on a firmer conceptual footing. Cathepsin K was confirmed as the major protease in human osteoclasts, and loss of function was shown to lead to an impairment in bone resorption. Our laboratories determined the crystal structure of cathepsin K (11) and undertook inhibitor design and synthesis (12). We further demonstrated that inhibitors of cathepsin K inhibit bone resorption both in vitro and in vivo (13). The gene-to-function-to-potential-drug paradigm was thus demonstrated.

Are Genes that Are Expressed in High Abundance Ideal Drug Targets?

The traditionalist might well question the EST paradigm with the choice of cathepsin K as a drug target. If high message abundance were to translate to high protein abundance, then the efficacy of a drug directed at that target might be limited by the sheer amount of protein it must bind to. A compelling example is the development of resistance to the antitumor drug, PALA, and hence perhaps its clinical failure as a result of induced expression of its target enzyme, aspartic acid transcarbamylase (14). Conversely, the success of the statins as cholesterol-lowering agents reflects the status of their molecular target, HMGCo-A reductase, as the rate-limiting enzyme in cholesterol biosynthesis (15). Cathepsin K was chosen as a drug target in part because of the high abundance of its expression in the target cell. Although the abundance of cathepsin K has not resulted in lack of efficacy of its inhibitors in animal models of osteoporosis (M Lark & G Stroup, unpublished data), high levels of expression could prove to be a fatal flaw in the EST strategy for other abundantly expressed targets.

AN ALTERNATIVE APPROACH—YOUR FAVORITE GENE SUPERFAMILY

The choice of cathepsin K as a molecular target for osteoporosis resulted primarily from identification of its high and selective expression in human osteoclasts. As it was a cysteine protease, its enzymatic mechanism of action could be extrapolated from studies on prototypic cysteine proteases. Therefore, approaches to its inhibition could be envisaged by experienced medicinal chemists. The drug discovery program thus appeared from the outset to be chemically tractable. Alternatively, the most abundantly expressed gene product could have been a partner in a spurious protein-protein interaction with no enzymatic activity and involved in unknown functions. Such a scenario would not have presented an attractive target for drug discovery.

To bias toward chemical tractability, one could select all members of gene superfamilies that have proven records in drug discovery and then attempt to assign biological function to them. Such superfamilies, when translated into protein families, could be the G-protein coupled receptors [also known as 7 transmembrane receptors (7TMRs)], ion channels, nuclear hormone receptors, and proteases, or more tenuously kinases. We chose to focus initially on the identification and functional characterization of novel 7TMRs.

The G-Protein Coupled Receptor Superfamily

Recent internal analysis pointed out that 37 marketed drugs are targeted at 21 distinct 7TMRs, representing $21 billion dollars in annual worldwide sales in 1997 (M Birkeland & P Agarwal, personal communication). There are currently over 250 known human 7TMRs, not including sensory olfactory receptors. It is estimated that the human genome will be shown to contain over 5000 members of this family, with 500–1000 being unrelated to odor-detecting receptors and, hence, of likely interest for drug discovery. Given the proven chemical tractability of agonism and antagonism within this superfamily, this represents a formidable opportunity.

At SB, we have identified 170 novel 7TMRs. Full-length cDNA clones have already been isolated for 117 of these, and 35 have been linked with ligands either in-house or by other groups. Our approach to functionally characterize these orphan receptors has been to first identify native or surrogate ligands for these receptors (16). To this end, we use stably or transiently transfected mammalian cell lines to guide biofractionation from various tissues following cytosolic calcium mobilization to detect agonist activity. In parallel, we test a bank of known bioactive substances. Once the agonist is detected, the receptor/agonist pair is configured into a high-throughput screening assay. Biological function is then sought using a combination of technologies to identify tissue distribution and responses in pharmacological assays chosen to reflect the tissue distribution. This approach also offers the tangible possibility that low-molecular-weight antagonists might be discovered early in the program by high-throughput screening and hence could serve as tool compounds to allow the function of the molecular target to be more readily assigned.

The discovery of the orexins 1 and 2 and their receptors and the pairing of melanin-concentrating hormone (MCH) to its receptor serve to illustrate these approaches.

Orexins and Orexin Receptors

Fifty stable transfectant HEK293 cell lines, each harboring a novel orphan 7TMR, were challenged with high-performance liquid chromatography fractions derived from various tissue extracts, with monitoring of cytoplasmic Ca^{2+} mobilization as a measure of agonism of the transfected receptor (17). In order to identify responses from endogenous receptors, three different transfectants, each expressing a distinct orphan 7TMR, were monitored and only those signals that were unique to a single cell line were pursued.

With this approach, several fractions from rat brain extracts elicited responses in a transfectant that expressed HFGAN72, a receptor originally identified as an EST from human brain. Purification of the active fractions exposed two peptides, which were called orexin-A and orexin-B. Orexin-A is a 33–amino acid peptide with an N-terminal pyroglutamate, a C-terminal amidation, and four cysteine residues that form two intramolecular disulfide bonds. Orexin-B, a 28–amino acid peptide, was 46% identical to orexin-A and was C-terminally amidated. Subsequent cloning of the cDNA for orexin-A revealed that the open reading frame encoded both orexin-A and orexin-B. Orexin-A and orexin-B are therefore expressed as a prepropeptide and processed proteolytically at dibasic amino acid residues like many known bioactive peptides. As orexin-B was considerably less active on HGFAN72 than was orexin-A, the presence of an orexin-B receptor was postulated. Such a receptor was subsequently identified in a BLAST search of the GenBank database using the sequence of HFGAN72 to identify paralogs. Subsequently OX_1R (HFGAN72) was found to be 64% identical to OX_2R at the protein sequence level.

Tissue Distribution of Orexin and Orexin Receptors and Potential Therapeutic Indication of Orexin Receptor Antagonists

In situ hybridization and immunohistochemical analyses in rat brains showed that orexin-containing neurons were present in the lateral and posterior hypothalamic areas. Because the lateral hypothalamic area has been implicated in the regulation of feeding behavior, orexins-A and -B may be involved in feeding behavior. To investigate this hypothesis, orexin-A was administered acutely into the lateral ventricle of male rats. A dramatic stimulation of feeding behavior was observed. Intuitively, an antagonist of orexin-A would appear to offer approaches to the treatment of obesity, diabesity, and diabetes mellitus.

The Receptor for Melanin-Concentrating Hormone

As part of the battery of orphan 7TMRs under study, a 353–amino acid human orphan known as SLC-1 was cloned from a human fetal brain cDNA library and expressed in HEK 293 cells (18). These cells were then challenged with a collection of known bioactive subtances, including >500 neuropeptides. Of these, only the cyclic neuropeptide MCH elicited a robust, dose-dependent elevation of intracellular calcium. MCH has long been implicated in the regulation of food intake and energy balance, but its receptor has been unknown. SLC-1 was shown for the first time by in situ hybridization and immunohistochemical techniques to be expressed in two nuclei of the hypothalamus, the ventromedial and dorsomedial nuclei, areas known to be involved in feeding. The pairing of MCH with SLC-1 allows configuration of a high-throughput screen and opens the route for the discovery of antagonists, which are likely to be useful in the treatment of obesity.

The Promiscuous Paradigm

The success in drug discovery engendered by members of the 7TMR superfamily as drug targets is not unrelated to a degree of promiscuity of ligand structures that cross-react among family members. The interplay of dopamine, epinephrine, norepinephrine, and serotonin (5-HT) is illustrative. This cross-reactivity of ligands probably reflects the derivation of discrete family member receptors from common ancestors. The promiscuity of ligands transfers to promiscuous structural templates found in antagonists of disparate receptors. An example of a promiscuous template is the biphenyl moiety found in the structure of the angiotensin antagonist losartan (Figure 2*A*) (19) and in various 5-HT1B receptor antagonists (Figure 2*B*) (20). Angiotensin II is an acidic ligand whereas 5-HT is a basic one, yet antagonists for their respective receptors can be presented on the same biphenyl scaffold.

Another example comes from SB programs with the commonality of the imidazoleacrylate scaffold of the angiotensin antagonist (Figure 3*A*) and the endothelin antagonist (Figure 3*B*) (21). In this case, cross-screening of angiotensin antagonists in an enthothelin receptor screen led to leads that could be converted into selective endothelin antagonists.

Combinatorial Libraries Based on Promiscuous Templates

Linking the two concepts of pairing novel orphan receptors within the 7TMR family with their ligands and antagonist structural promiscuity suggests the interfacing of combinatorial libraries based on promiscuous templates with the stable

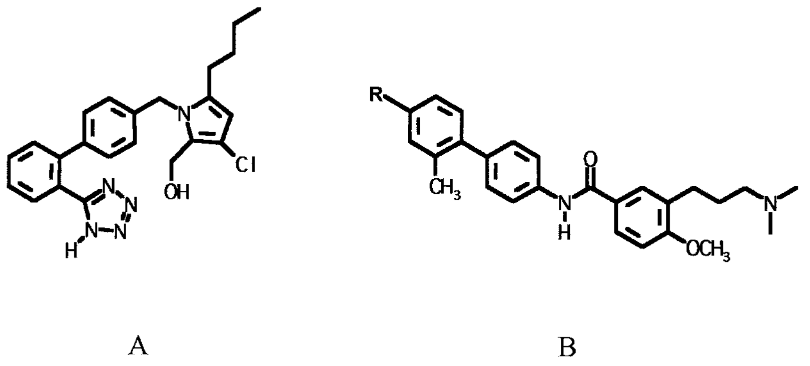

Figure 2 The structure of the angiotensin antagonist losartan (*A*) and the 5-HT1B receptor antagonist (*B*).

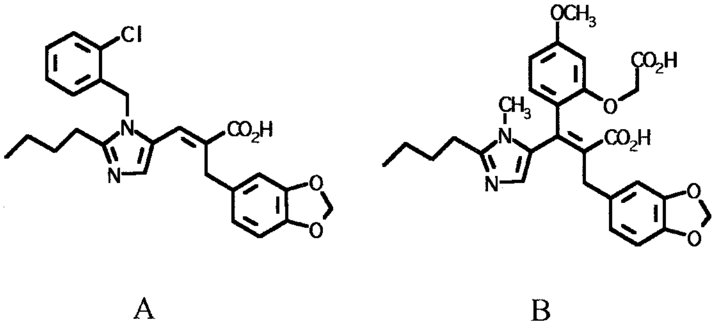

Figure 3 The imidazoleacrylate scaffold of the angiotensin antagonist (*A*) and the endothelin antagonist (*B*).

transfectants of orphan receptors in a high-throughput screening mode. For example, a combinatorial library was prepared based on the biphenyl scaffold and is proving to be a useful resource in the search for antagonists for novel 7TMRs (22). Similarly, the imidazole acrylate substructure common in Figure 3*A* and *B* would also offer a likely template for combinatorial expansion.

THE MOLECULAR TECHNOLOGIES REQUIRED

The prototypical examples cited above of the discovery of cathepsin K and its tantalizing validation as a drug discovery target, and the generic approach to the discovery and characterization of novel 7TMRs, their cognate ligands, and their antagonists, rely on a number of platform technologies. These include rapid full-length cDNA cloning, tissue localization, combinatorial chemistry, high-throughput screening, and the underlying automation and data management. Given the multiplicity of the target opportunities presented by genomics, a major investment in these downstream platform technologies is required. These different technologies each have their own evolution, although striking convergence is seen between genomics, combinatorial chemistry, and high-throughput screening with respect to handling thousands of objects, dependence on miniaturization and laboratory automation, creation of information and data tidal waves, and need for effective data management and mining.

Many approaches are needed to appropriately validate novel drug discovery targets. These include tissue distribution in normal and diseased tissues, chromosome localization, and analysis of orthologous genes in model systems such as yeast, the worm *Caenorhabditis elegans*, fruit flies, or mice. Yeast and *C. elegans* are particularly powerful systems because their entire genome sequence has been determined (23, 24), and ingenious genetic techniques are available allowing facile gene disruption, gene complementation, etc. Because the estimated number of human genes is high (80,000–100,000), scientists are striving to develop high-throughput technologies in which all or many genes are analyzed in parallel. For example, a critical technology in target validation is the use of high-density gene microarrays to monitor spatial and temporal gene expression. In addition, the correlation or lack of correlation of mRNA levels with translated protein and subsequent posttranslational processing and modifications can be accessed by proteomics.

The Gene Microarrays

Gene microarrays allow the rapid parallel analysis of the expression of thousands of genes against hundreds of tissues, cell types, and conditions whether using oligonucleotide arrays (25) or arrays of gene fragments (26). Because of the compatibility of glass with fluorescently labeled probes, the most effective gene microarrays are those constructed on glass surfaces, as opposed to arrays on nylon membranes (27). Because fluorescent probes exhibit little-to-no signal dispersion, very dense array spacing is possible, and deposition of DNA samples at densities of thousands of discrete DNA spots per 1.0 cm^2 have been attained. Another advantage of dense arrays on small glass surfaces is that the volumes required for hybridization are very small, thereby allowing the use of probes from small amounts of tissues.

The number and variety of microarray applications is virtually unlimited (28). Gene microarrays are being used extensively to generate broad and in-depth data on gene expression patterns in normal and diseased tissues, both in human and in animal systems. Although many effective drugs have been developed against targets that are widely expressed in the body (e.g. the angiotensin converting enzyme), highly selective tissue expression of a drug target, such as that seen for cathepsin K, is attractive, as the potential for unwanted side effects may be more restricted. Perhaps the most promising application of microarrays is the study of differential expression in disease. The up- or down-regulation of a gene can be the cause of the disease or its result. In addition, microarrays are a powerful tool to help dissect the mechanism of action of drugs and drug candidates. They will also increasingly contribute to the analysis of metabolic pathways for drugs, the understanding and prediction of toxic or adverse events in vitro and in vivo, as well as the potential identification of surrogate markers to follow the dose and even efficacy of a drug in the clinical setting.

Microarray-based assays yield several hundred data points for thousands of genes, and this in turn demands the development of fully automated and standardized software systems for the collection, quality scoring, and tracking of all data points. The rush to apply this new technology should not ignore quality assurance because low-quality data will only generate poor biological conclusions. It will therefore be critical that strict and broadly accepted quality standards be developed, so that data from different laboratories can be compared and even combined.

Microarrays are not the panacea for gene expression analysis. They have a number of limitations that must be addressed by complementary technologies. First, the sensitivity of detection is about 1 in 100,000. More sensitive but lower throughput methods such as quantitative polymerase chain reaction, e.g. the recently automated Taqman technology (29), must be utilized for the analysis of genes that are expressed in low abundance. Second, the labor-intensive in situ hybridization and immunocytochemistry methods will continue to make important contributions because they provide a critical link to histology and cytology that the best microdissection of cells from tissues is unlikely to provide. Third, mRNA levels may not be paralleled at the protein level. To address this, high-throughput methods for the analysis of differential expression at the protein level are being developed to characterize translational and posttranslation regulation. This emerging technology is referred to as proteomics (30).

Proteomics

Recent technological advances in protein analysis have paved the way for proteomics, which characterizes the protein complement, proteome, of a cell or tissue as opposed to the mRNA transcripts, transcriptome, studied by microarrays. These advances were made possible by significant progress in two-dimensional gel electrophoresis and ultrasensitive mass spectrometry. The improvements to two-dimensional gel electrophoresis included larger format to allow separation of thousands of protein spots and reproducibility and image scanning to position spots and to quantitate their intensity between gels. Mass spectrometry has provided a quantum leap in sensitivity, as even the most sensitive Edman sequencers could not sequence the majority of the spots. Currently, the extensive deployment of proteomic approaches to support target validation is limited by the inherent low throughput of the method. Early applications are likely to be restricted to selected subsets of the proteome, such as phosphorylated proteins where the proteins under study are defined by those revealed by anti-phosphotyrosine or phosphoserine antibodies (31), or to the bacterial proteome. Other defined subsets of the proteome that are yielding to these approaches include the ribosomal protein complexes where interacting partners can be defined (32).

FUTURE OPPORTUNITIES

By the spring of 2000, 90% of the human genome will be sequenced, inaugurating the twenty-first century with a historical achievement for mankind. Bioinformaticians and molecular biologists will be challenged to select the protein coding information out of 2.9 billion noncoding nucleotides. Gene-calling algorithms are being developed to allow the rapid identification of open reading frames, which will need to be confirmed as truly expressed genes as opposed to pseudogenes. This exercise will be greatly facilitated by the availability of the millions of ESTs generated to date (33, 34). Regulatory sequences, including promoters and enhancers, will become available for study and will facilitate therapeutic intervention at the level of transcription. The study of regulatory elements will benefit tremendously from being interfaced with tissue expression data generated by microarray analysis, prompting identification of common regulatory features in genes with coregulated expression profiles (35).

The accuracy of genome sequence data will be very high (99.99%), and hence, nucleotide sequence differences, known as single-nucleotide polymorphisms, in the same gene derived from different individuals will be reliable. These single-nucleotide polymorphisms will constitute the basis for pharmacogenetic studies of differences between individuals with respect to drug efficacy and to manifestation of adverse events (36). The completion of animal genome sequences will be expected not too long after that of the human genome, allowing for unprecedented comparisons of biochemical pathways and physiological phenomena between humans, mice, and other animal models used in research for decades.

High-throughput screening will migrate from the current 384 format to a high-density 1536 format, aided by fluorescent detection methodologies (37). Combinatorial chemistry will evolve in high-speed medicinal chemistry where thousands of compounds can be prepared as singles, rather than as mixtures. These advances are discussed elsewhere in this edition (38).

LITERATURE CITED

1. Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, et al. 1991. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* 252:1651–56
2. Pandey A, Lewitter F. 1999. Nucleotide

sequence databases: a gold mine for biologists. *Trends Biochem. Sci.* 24:276–80

3. Karp PD, Krummenacker M, Paley S, Wagg J. 1999. Integrated pathway-genome databases and their role in drug discovery. *Trends Biotechnol.* 17:275–81
4. Hillier LD, Lennon G, Becker M, Bonaldo MF, Chiapelli B, et al. 1996. Generation and analysis of 280,000 human expressed sequence tags. *Genome Res.* 6:807–28
5. James IE, Walsh S, Dodds RA, Gowen M. 1991. Production and characterization of osteoclast-selective monoclonal antibodies that distinguish between multinucleated cells derived from different human tissues. *J Histochem. Cytochem.* 39:905–14
6. Drake FH, Dodds R, James I, Connor J, Debouck C, et al. 1996. Cathepsin K but not cathepsins B, L or S is abundantly expressed in human osteoclasts. *J. Biol. Chem.* 271:12511–16
7. Bossard MJ, Tomaszek TA, Thompson SK, Amegadzie BY, Hanning CR, et al. 1996. Proteolytic activity of human osteoclast cathepsin K: expression, purification, activation and substrate identification. *J. Biol. Chem.* 271:12517–24
8. Gelb BD, Shi G-P, Chapman HA, Desnick RJ. 1996. Pycnodysostosis, a lysosomal disease caused by cathepsin K deficiency. *Science* 273:1236–38
9. Saftig P, Hunziker E, Wehmeyer O, Jones S, Boyde A, et al. 1998. Impaired osteoclastic bone resorption leads to osteopetrosis in cathepsin-K-deficient mice. *Proc. Natl. Acad. Sci. USA* 95:13453–58
10. Gowen M, Lazner F, Dodds R, Field J, Tavaria M, et al. 1999. Cathepsin K knockout mice develop osteopetrosis due to a deficit in matrix degradation but not demineralization. *J. Bone Miner. Res.* In press
11. Zhao B, Janson C, Amegadzie B, D'Alessio K, Griffin C, et al. 1997. Crystal structure of human osteoclast cathepsin K complex with E64. *Nat. Struct. Biol.* 4:109–11
12. Yamashita DS, Smith WW, Zhao B, Janson CA, Tomaszek TA, et al. 1997. Structure and design of potent and selective cathepsin K inhibitors. *J. Am. Chem. Soc.* 119:11351–52
13. Votta BJ, Levy MA, Badger A, Bradbeer J, Dodds RA, et al. 1997. Peptide aldehyde inhibitors of cathepsin K inhibit bone resorption both in vitro and in vivo. *J. Bone Miner. Res.* 12:1396–406
14. Kensler TW, Mutter G, Hankerson JG, Reck LJ, Harley C, et al. 1981. Mechanism of resistance of variants of the Lewis lung carcinoma to N-(phosphonacetyl)-L-aspartic acid. *Cancer Res.* 41:894–904
15. Shapiro DJ, Rodwell VW. 1971. Regulation of hepatic 3-hydroxy-3-methylglutaryl coenzyme A reductase and cholesterol synthesis. *J. Biol. Chem.* 246: 3210–16
16. Stadel JM, Wilson S, Bergsma DJ. 1997. Orphan G protein-coupled receptors: a neglected opportunity for pioneer drug discovery. *Trends Pharmacol. Sci.* 18: 430–37
17. Sakurai T, Amemiya A, Ishii M, Matsuzaki I, Chemelli RM, et al. 1998. Orexins and orexin receptors: a family of hypothalamic neuropeptides and G protein-coupled receptors that regulate feeding behavior. *Cell* 92:573–85
18. Chambers J, Ames RS, Bergsma DJ, Muir A, Fitzgerald LR, et al. 1999. Melanin-concentrating hormone is the cognate ligand for the orphan G-protein-coupled receptor SLC-1. *Nature* 400: 261–65
19. Chiu AT, McCall DE, Price WA, Wong PC, Carini DJ, et al. 1990. Nonpeptide angiotensin II receptor antagonists. VII. Cellular and biochemical pharmacology of DuP 753, an orally active antihypertensive agent. *J. Pharm. Exp. Ther.* 252:711–18

20. Clitherow JW, Scopes DIC, Skingle M, Jordan CC, Feniuk W, et al. 1994. Evolution of a novel series of [(N,N-dimethylamino)propyl]- and piperazinylbenzanilides as the first selective 5-HT$_{1D}$ antagonists. *J. Med. Chem.* 37:2253–57
21. Elliott JD, Bryan DL, Nambi P, Ohlstein EH. 1996. A novel series of non-peptide endothelin receptor antagonists. In *Peptides: Chemistry, Structure and Biology*, ed. PTP Kaumaya, RS Hodges, pp. 673–75. Kingswinford, UK: Mayflower Sci.
22. Chenera B, Finkelstein JA, Veber DF. 1995. Photodetachable arylsilane polymer linkages for use in solid phase organic synthesis. *J. Am. Chem. Soc.* 117:11999–2000
23. Mewes HW, Albermann K, Bahr M, Frishman D, Gleissner A, et al. 1997. Overview of the yeast genome. *Nature* 387(Suppl):7–65
24. *C. elegans* Sequencing Consort. 1998. Genome sequence of the nematode *Caenorhabditis elegans.* A platform for investigating biology. *Science* 282:2012–18
25. Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, et al. 1996. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.* 14:1675–80
26. Schena M, Shalon D, Davis RW, Brown PO. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270:467–70
27. Mooney J, Kayne P, O'Brien S, Debouck C. 2000. Construction and applications of gene microarrays on nylon membranes. In *PCR5: Differential Display: A Practical Approach*, ed. R Leslie, H Robertson. New York: Oxford Univ. Press. In press
28. Debouck C, Goodfellow P. 1999. DNA microarrays in drug discovery and development. *Nat. Genet.* 21:48–50
29. Wang T, Brown MJ. 1999. mRNA quantification by real time TaqMan polymerase chain reaction: validation and comparison with RNase protection. *Anal. Biochem.* 269:198–201
30. Blackstock WP, Weir MP. 1999. Proteomics: quantitative and physical mapping of cellular proteins. *Trends Biotechnol.* 17:121–27
31. Soskic V, Görlach M, Poznanovic S, Boehmer FD, Godovac-Zimmermann J. 1999. Functional proteomics analysis of signal transduction pathways of the platelet-derived growth factor β receptor. *Biochemistry* 38:1757–64
32. Link AJ, Eng J, Schieltz DM, Carmack E, Mize GJ, et al. 1999. Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* 17:676–82
33. Marra MA, Hillier L, Waterston RH. 1998. Expressed sequence tags—ESTablishing bridges between genomes. *Trends Genet.* 14:4–7
34. Bailey LC Jr, Searls DB, Overton GC. 1998. Analysis of EST-driven gene annotation in human genomic sequence. *Genome Res.* 8:362–76
35. Bucher P. 1999. Regulatory elements and expression profiles. *Curr. Opin. Struct. Biol.* 9:400–7
36. Kleyn PW, Vesell ES. 1998. Genetic variation as a guide to drug development. *Science* 281:1820–21
37. Pope AJ, Haupts UM, Moore KJ. 1998. Homogeneous fluorescence readouts for miniaturized high-throughput screening: theory and practice. *Drug Disc. Today* 4:350–62
38. Ohlstein EH, Ruffolo RR Jr, Elliott JD. 2000. Drug discovery in the next millennium. *Annu. Rev. Pharmacol. Toxicol.* 40:177–90