DRUG DISCOVERY

AND GENOMIC TECHNOLOGIES

Deciphering Therapeutic Targets

Tomi K. Sawyer

ARIAD Pharmaceuticals,
Cambridge, MA, USA

The recent publication of human genome maps by the International Human Genome Sequencing Consortium (10) and Celera Genomics (20) provides a milestone toward the continuing emergence of a “genes-to-drugs” paradigm for drug discovery. These studies estimate that approximately 30 000 human genes exist—equivalent to about 1% of the human genome. Genes provide the molecular template for the human proteome—the complete set of proteins encoded by the genome. It is also important not to exclude noncoding regions of the genome because these sequences contain regulatory elements for gene transcription and related functions that are manifest, for example, in alternative start and stop sites for transcription. The mapping of the human genome represents a major milestone for the drug discovery effort as it moves toward the implementation of sophisticated genomic technologies, which include functional genomics, proteomics, and informatics. These technologies complement the armamentarium of existing core competencies, such as high-throughput screening and mechanism-based assays, medicinal and combinatorial chemistry, structural biology and drug design, cell biology and pharmacology, in vivo disease models, pharmacokinetics and drug metabolism, and preclinical toxicology, as tools for the development of new medicines.

Drug Discovery in Transition

Contemporary drug discovery is built on a foundation of early pioneering work, which has culminated in the successful production of many drugs for cancer, cardiovascular disease, central nervous system (CNS) disease,

endocrine and metabolic disorders, immune and inflammatory diseases, and infectious diseases. In retrospect, past achievements have focused on about 500 therapeutic targets (5). Interestingly, based on a simple biochemical classification, therapeutic targets have included receptors and ion channels (50%), enzymes (28%), hormones and factors (11%), DNA and nuclear receptors (4%), and other targets (2%). On the other hand, a prospective view of therapeutic targets needs to incorporate the spectrum of possibilities afforded by knowledge of the human genome, and it is likely that more than 5000 therapeutic targets may become available for drug discovery in the new millennium. Hence, the quest to identify, assign function to, and validate the thousands of potential new therapeutic targets presents a formidable challenge. Accordingly, the modus operandi of drug discovery is to integrate key genomic technologies to fulfill a genes-to-drugs paradigm (Figure 1).

Impact of the Human Genome Map

The human genome has not yet been entirely mapped (about 95% has been sequenced and assembled), and detailed annotation, comparative genomic analysis, and molecular dissection of complex phenotypes will be essential to comprehensively exploit the blueprint of life. The complex nature of the human genome and its estimated 3.2 gigabases is realized with respect to the challenges of mapping genes to contribute protein-coding sequences versus relatively expansive noncoding sequences. Nonetheless, the functions of predicted proteins encoded by the human genome (20) suggest that the major categories include nucleic acid enzymes (7.5%), transcription factors (6%), receptors (5%), hydrolases (4%; proteases dominate this category), and kinases (2.8%). Further, nearly 50% of predicted human proteins do not yet readily fit into these functional categories. Another analysis of the human genome map identifies predicted proteins in terms of structure (10), as shown in Table 1, which summarizes the top 25 most frequently occurring protein families or structural domains. Finally, it is noteworthy that the human genome codes for greater than 10 000 secreted proteins. Collectively, these data describing predicted human proteins, extrapolated from the se-

Figure 1. A genes-to-drugs paradigm for drug discovery. Not all core competencies are shown.

DRUG DISCOVERY

AND GENOMIC TECHNOLOGIES

Table 1. Predicted Human Proteins Families or Domains^a

Rank (No.)	Structural Category
1 (765)	Immunoglobulin domain
2 (706)	C2H2 zinc finger
3 (575)	Protein kinase
4 (569)	GPCR superfamily
5 (433)	P loop motif
6 (350)	Reverse transcriptase
7 (300)	rrm domain
8 (277)	G protein BWD-40 repeat
9 (276)	Ankyrin repeat
10 (267)	Homeobox domain
11 (252)	PH domain
12 (242)	EF hand family
13 (222)	EGF-like domain
14 (215)	SH3 domain
15 (210)	RING finger
16 (188)	Leucine-rich repeat
17 (171)	KRAB box
18 (165)	Fibronectin type III domain
19 (160)	PDZ domain
20 (153)	Helicase C-terminal domain
21 (150)	TPR repeat
22 (133)	Ion transport protein
23 (131)	Helix-loop-helix DNA-binding domain
24 (123)	C2 domain
25 (119)	SH2 domain

^aNot comprehensive; adapted from Reference 10.

quence of almost the entire genome, impact drug discovery significantly by providing a vast spectrum of potential therapeutic targets.

Functional Genomics, Proteomics, and Informatics

The genes-to-drugs paradigm for drug discovery begins with disease gene identification and expression using genomic technologies, including a platform of core competencies, to ultimately yield validated therapeutic targets. Some major technologies for disease gene investigation include expressed sequence tag sequencing, secreted protein analysis, differential display, expression profiling, and positional cloning. A number of disease genes have already been directly identified as a result of positional cloning methods using sequence data (10); some examples include genes associated with breast cancer susceptibility (*BRCA2*), X-linked lymphoproliferative disease (*XLP*), familial hyperlysinemia (*AASS*), total color blindness (*CNGB3*), optic atrophy (*OPA1*), X-linked congenital stationary night blindness (*XLCSNB*), hypophosphotemic

rickets (*FGF23*), and progressive pseudorheumatoid dysplasia (*WISP3*). High-throughput microarray-based methods will further enable the annotation of the human genome by validating predicted exons, grouping exons into genes on the basis of co-regulated expression and defining full-length mRNA transcripts (19). A map of the human transcriptome—the complete set of mRNAs transcribed by the genome—has been developed, and this map reveals a clustering of highly expressed genes in specific chromosomal regions and may provide a tool for the identification of genes that are overexpressed or suppressed in cancers (3).

Knowledge of the complete sequence of the human genome will facilitate the ultimate goal of identifying all genes that contribute to disease; efforts to correlate functional categorization of 923 known disease genes with disease characteristics (e.g., frequency, mode of inheritance, age of onset, and reduction of life expectancy) have been launched (13). It is worth mentioning that several new therapeutic targets have recently emerged from the genome sequencing project, including the serotonin receptor subtype 5-HT_{3B} (4), the cysteinyl leukotriene receptor subtype CysLT₂ (9), and β site amyloid precursor protein-cleaving enzyme homolog BACE2 (6). Furthermore, the analysis of potential therapeutic targets, predicted from the recent human genome map and having a minimum of 70% structural identity to at least 50 amino acids relative to established drug targets, has resulted in the identification of a number of apparently new receptors, ion channels, and enzymes (10). A map of approximately 1.42 million single nucleotide polymorphisms (SNPs) distributed throughout the human genome has been assembled (12) and is expected to aid in the identification of disease genes for both drug discovery and diagnostic applications.

Separate from the mapping of the human genome, but critical to human functional genomics studies and the development of in vivo disease models, is mapping of the mouse genome and phenotypic analysis (15). Functional annotation of mouse genome maps (predicted to include about 30 000 genes, similar to the human genome) remains incomplete. However, high-throughput DNA-based mutation detection systems and functional characterization of “trapped” and engineered mutations provide possible routes to further map and exploit the mouse as a powerful tool in the genes-to-drugs paradigm for drug discovery (11).

Beyond the mapping of the human genome, future studies will employ functional genomics (7,8,17), proteomics (1,14,21), and informatics (2,16,18) to provide global analysis (experimental and computational) of RNA and protein expression, protein localization, protein-protein interactions, and chemical modulation of protein pathways. Databases for genome mapping, nucleic acid and protein sequence, and nucleic acid and protein structure exist. Internet resources provide invaluable information and make available databases of relevance to drug discovery and genomic technologies (Table 2).

DRUG DISCOVERY

AND GENOMIC TECHNOLOGIES

Table 2. Some Internet Resources for Drug Discovery and Genomic Technologies

Name of Resource	Web Site
Human Genome Project	http://www.nhgri.nih.gov/HGP
International Human Genome Sequencing Consortium	http://www.ensembl.org/genome/central
Nature & Nature (Genetics) Genomics Gateway	http://www.nature.com/genomics/links
Jackson Laboratory Transgenic/Targeted Mutation Database	http://tbase.jax.org
BioMedNet Mouse Knockout and Mutation	http://www.biomednet.com/db/mkmd
The Rat Genome Database	http://ratmap.gen/gu.se
TIGR Microbial Genome Database	http://www.tigr.org/tdb/mdb/mdbcomplete.html
The Nucleic Acid Database Project	http://ndbserver.rutgers.edu
Protein Data Bank	http://www.rcsb.org/pdb
BioMagResBank Repository of NMR Data	http://www.bmr.b.wisc.edu/Welcome.html
ExPASy Molecular Biology and Proteomics	http://www.nhgri.nih.gov
European Molecular Biology Laboratory	http://www.embl-heidelberg.de
National Center of Biotechnology Information	http://www.ncbi.nlm.nih.gov/Structure
National Human Genome Research Institute	http://www.nhgri.nih.gov/80/Data/#mapping
The Sanger Center	http://www.sanger.ac.uk
Whitehead Institute/MIT Center for Genome Research	http://www-genome.wi.mit.edu
Celera Genomics	http://www.celera.com
The Institute for Genomic Research	http://www.tigr.org

REFERENCES

1. Aebersold, R., and D.R. Goodlet. 2001. Mass spectrometry in proteomics. *Chem. Rev.* 101:269-295.
2. Birney, W., A. Bateman, M.E. Clamp, and T.J. Hubbard. 2001. Mining the draft human genome. *Nature* 409:827-828.
3. Caron, H., B. van Schaik, M. van der Mee, F. Baas, G. Riggins, P. van Sluis, M.-C. Hermus et al. 2001. The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science* 291:1289-1292.
4. Davies, P.A., M. Pistis, M.C. Hanna, J.A. Peters, J.J. Lambert, T.G. Hales, and E.F. Kirkness. 1999. The 5-HT3B subunit is a major determinant of serotonin-receptor function. *Nature* 397:359-363.
5. Drews, J. 2000. Drug discovery: a historical perspective. *Science* 287:1960-1964.
6. Fan, W., B.D. Bennett, S. Babu-Kahn, Y. Luo, J.-C. Louis, M. McCaleb, M. Citron, R. Vassar, and W.G. Richards. 1999. BACE maps to chromosome 11 and BACE homolog, BACE2, reside in the obligate Down Syndrome region of chromosome 21. *Science* 286:1255.
7. Futreal, P.A., A. Kasprzyk, E. Birney, J.C. Mullikin, R. Wooster, and M.R. Stratton. 2001. Cancer and genomics. *Nature* 409:850-852.
8. Harris, S. and S.M. Foord. 2000. Transgenic gene knock-outs: functional genomics and therapeutic target selection. *Pharmacogenomics*. 1:433-443.
9. Heise, C.E., B.F. O'Dowd, D.J. Figueroa, N. Sawyer, T. Nguyen, D.S. Im, R. Stocco, J.N. Bellfeuille et al. 2000. Characterization of the human cysteinyl leukotriene 2 receptor. *J. Biol. Chem.* 275:30531-30536.
10. International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* 409:860-921.
11. International Mouse Mutagenesis Consortium. 2001. Functional annotation of mouse genome sequences. *Science* 291:1251-1255.
12. International SNP Map Working Group. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409:928-933.
13. Jimenez-Sanchez, G., B. Childs, B., and D. Valle. 2001. Human disease genes. *Nature* 409:853-855.
14. Legrain, P. and L. Selig. 2000. Genome-wide protein interaction maps using two-hybrid systems. *FEBS Lett.* 480:32-36.
15. Nolan, P.M. 2000. Generation of mouse mutants as a tool for functional genomics. *Pharmacogenomics* 1:243-255.
16. Roos, D.S. 2001. Bioinformatics—trying to swim in a sea of data. *Science* 291:1260-1261.
17. Rosamond, J. and A. Allsop. 2001. Harnessing the power of the genome in the search for new antibiotics. *Science* 287:1973-1976.
18. Sander, C. 2000. Genomic medicine and the future of health care. *Science* 287:1977-1978.
19. Shoemaker, D.D., E.E. Schadt, C.D. Armour, Y.D. He, P. Garrett-Engle, P.D. McDonough, P.M. Loerch et al. 2001. Experimental annotation of the human genome using microarray technology. *Nature* 409:922-927.
20. Venter, J.C., M.D. Adams, E.W. Myers, P.W. Li, R.J. Mural, G.G. Sutton, H.O. Smith et al. 2001. The sequence of the human genome. *Science* 291:1304-1351.
21. Yoshida, M., J.A. Loo, and R.A. Lepley. 2001. Proteomics as a tool in the pharmaceutical drug design process. *Curr. Pharm. Des.* 7:291-310.

Suggestions for contributions to the "Drug Discovery and Genomic Technologies" section are welcomed by its editor, Dr. Tomi K. Sawyer (tomi.sawyer@ariad.com)