Am J Pharmacogenomics 2005; 5 (6): 387-396
1175-2203/05/0006-0387/$34.95/0

Using Bioinformatics for Drug Target Identification from the Genome

Zhenran Jiang and *Yanhong Zhou*

Hubei Bioinformatics and Molecular Imaging Key Laboratory, School of Computer Science, Huazhong University of Science and Technology, Wuhan, China

Contents

Abstract

Genomics and proteomics technologies have created a paradigm shift in the drug discovery process, with bioinformatics having a key role in the exploitation of genomic, transcriptomic, and proteomic data to gain insights into the molecular mechanisms that underlie disease and to identify potential drug targets. We discuss the current state of the art for some of the bioinformatic approaches to identifying drug targets, including identifying new members of successful target classes and their functions, predicting disease relevant genes, and constructing gene networks and protein interaction networks. In addition, we introduce drug target discovery using the strategy of systems biology, and discuss some of the data resources for the identification of drug targets.

Although bioinformatics tools and resources can be used to identify putative drug targets, validating targets is still a process that requires an understanding of the role of the gene or protein in the disease process and is heavily dependent on laboratory-based work.

The classical progression of the pharmaceutical discovery process goes from drug target to lead compound to drug. Effective discovery of disease-associated targets for further validation is the first critical step in this process. The more information we have about potential drug targets, the more opportunities we have to develop successful drugs. Genomics research has deepened the pool of potential drug targets, however, a major challenge for drug development continues to be the rapid and accurate identification of drug targets with true potential. It is reported that just 483 drug targets account for nearly all the drugs currently on the market (45% receptors, 28% enzymes, 5% ion channels, and 2% nuclear receptors).[1] However, it is estimated there might be thousands of

drug targets within the human genome, indicating the huge potential for drug target discovery.[2-4] Currently, most of the new drugs approved by the regulatory authorities are based on protein targets for which marketed drugs already exist.[5] Addressing this 'innovation gap' has resulted in the development of the new paradigm of genomics-based drug discovery, with bioinformatics having a key role in the exploitation of genomic, transcriptomic, and proteomic data to gain insights into the molecular mechanisms that underlie disease and to search for targets that will lead to new drugs.[6-8]

In this review, some of the data resources and computational methods for the identification of potential drug targets are summarized.

1. Defining Drug Targets

Drug targets are membrane or cellular receptors or other molecules that are pivotally involved in a disease process. From a pharmacological viewpoint, a drug target is either inhibited or activated by drug molecules (e.g. small organic molecules, antibodies, therapeutic proteins). Drug molecules can physically attach to a drug target, triggering a cascade of intracellular biochemical reactions, followed by a cellular reaction. Potential drug targets can include:

- genes that are differentially expressed between individuals who are and are not in need of treatment for a particular disease or condition;
- genes that are differentially expressed when that individual is exposed to a drug known to alleviate or exacerbate the symptoms of interest;
- genes that are co-expressed with other genes presumed to be involved in the systems and pathways under study;
- genes that serve as pathway initiators.

Any gene (or its product) falling into any one of those categories may be a gene for which manipulation of its expression might affect disease or symptom progression.[9]

1.1 Characteristics of a Putative Target

Based on the analysis of the molecular targets of current therapies, biologists have revealed that the most successful drug targets share several basic characteristics.[10] First, the most successful targets tend to be amenable to medical intervention using therapeutic drugs that fall into three major classes: small molecules, antibodies, or therapeutic proteins. Secondly, in inhibiting or activating the target should have a clear therapeutic effect. Thirdly, a drug target should have robust assay systems for *in vitro* characterization and high-throughput screening. In addition, an ideal target should be specific and essential disease process, and targeting it should not only address unmet medical needs but also serve major medical markets. The principles described above have been consistent traits associated with targets of proven value and, therefore, can be used as a simple set of rules to guide target discovery, validation, and development.

1.2 Successful Target Classes

Certain classes of proteins are more amenable to drug development than others. Historically, G-protein coupled receptors (GPCRs) have been the major drug target class for the pharmaceutical industry, with ion channels, nuclear hormone receptors, proteases, kinases, phosphodiesterases, phosphatases, and other key enzymes making up the remaining target classes.[2,7,11] Proteins within these target families tend to exert a biological effect that is amplified within cells or organisms by a variety of signaling mechanisms.

1.2.1 G-Protein Coupled Receptors

GPCRs constitute one of the most important families of drug targets in the pharmaceutical industry and are central to the signaling networks that regulates basic cellular processes.[12,13] Over the last decade the number of characterized GPCRs has grown steadily and more than 700 GPCR genes have been identified from human genome.[14] However, given the track record of GPCRs as validated drug targets, the vast number of potentially untapped targets within this superfamily still presents an intriguing challenge for drug target discovery. Despite their importance, the power and utility of microarray technology has not been extended to membrane proteins because of significant technical challenges associated with their fabrication and use. GPCRs, like other membrane-embedded proteins, have characteristics that make their three-dimensional structures extremely difficult to determine experimentally. To date, the three-dimensional structures of GPCRs are unsolved, except for that of the GPCR-bovine rhodopsin.[15] As a result, the structure-based *in silico* methods of drug discovery cannot be used effectively with regards to GPCR targets, and the design of ligands of GPCRs has to rely on ligand-based techniques.

As GPCRs are proven to be important drug targets, the pharmaceutical industry is devoting enormous amounts of money and manpower to identify these targets. The success of this monumental effort depends on the correct identification and delineation of the functions of GPCRs and effectively applying that information toward drug discovery.[13] One of the key areas for innovation is further development of two-hybrid methods suitable for GPCRs and other transmembrane proteins.

1.2.2 Ion Channels

Ion channels are another attractive drug target class. Ion channels have potential as drug targets for several reasons. First, ion channels are required in various normal physiological processes. The dysfunction of ion channels can have a strong impact on cellular function and signaling. Secondly, ion channels belong to one of a few protein classes that are highly amenable to regulation

by small molecule drugs. Thirdly, ion channels are expressed in numerous cell types and occur as large families of related genes with cell-specific expression patterns. Despite their remarkable physiological value, ion channels remain a relatively unexploited therapeutic target class, especially in comparison with target areas such as GPCRs or kinases. Major challenges have been the lack of high-throughput screening assays and available targeted libraries of candidate ion channel modulators.[16-18]

However, ion channels are currently experiencing renewed interest from pharmaceutical and drug discovery companies due to the progress of new high-throughput technological approaches. The large number of diseases that are attributable to ion channel dysfunction are the primary drivers for the development of ion channel targets. According to a report by the US Food and Drug Administration (FDA) in 2003, the number of new approved drugs targeting ion channels is equal to or even higher than those targeting proteases, polymerases, and reverse transcriptases. In the post-genomics era, progress in function genomics will reveal the tissue-specific distributions of ion channels and a greater understanding of these proteins, meaning that ion channels will play an increasingly important role as therapeutic drug targets in a number of areas, including asthma, inflammation, arrhythmia, and CNS disorders.[19]

1.2.3 Nuclear Hormone Receptors

Nuclear hormone receptors are outstanding targets for drug discovery, not only because of their profound roles in human physiology and diseases but also because their structures allow them to interact with small chemical molecules.[20]

The current members of the nuclear receptor gene family can be divided into two main classes: the 'validated' nuclear receptors, whose ligands and endocrine pathways are established and as a result serve as bona fide drug targets for human disease; and the 'orphan' nuclear receptors, whose ligands, target genes, and physiological function are not completely understood, and offer new first-in-class targets for large therapeutic areas, in particular, cardiovascular and metabolic disorders. In addition, several members of the nuclear hormone receptor superfamily are directly involved in tumor progression, or conversely, have shown tumor-suppressive potential through modulation of cell proliferation, differentiation, and apoptosis (e.g. the anticancer drugs tamoxifen and flutamide act by targeting nuclear receptors). Using advanced structure-based bioinformatics tools, Inpharmatica Ltd has identified 16 proteins with previously unrecognized structural similarity to the ligand-binding domain of the nuclear receptors, all clearly outside of the known family members.[21] The detailed knowledge of the structural mechanism underlying activation and inhibition of nuclear receptors by small molecule modulators begets important therapeutic opportunities.

1.2.4 Proteases and Kinases

Given the importance of altered protease expression/function in many diseases, proteases and their substrates are increasingly viewed as important drug targets.[22,23] Proteases exert high-order post-translational control over a diverse range of cellular functions. Elucidating the substrate repertoire of a protease is critical to understanding its biological role. Serine proteases, the largest human protease gene family, have been implicated in the growth and progression of solid tumor cancers, including breast and prostate cancer.

Protein kinases also present distinctive advantages as potential drug targets.[24] The human genome contains over 500 different protein kinases, which are the key regulatory enzymes that catalyze the phosphorylation of proteins at about 100 000 different sites to reversibly control their functional activities. Defects in specific protein kinases have been linked to over 400 diseases, including cancer, diabetes mellitus, and Alzheimer disease, and about 25% of all pharmaceutical industry research and development is now focused on the discovery and evaluation of protein kinase inhibitors for therapeutic applications.[25] Numerous specific kinases have been identified as attractive drug targets for inflammation, cancer, and other diseases. It has been reported that cyclin-dependent kinase-5 (CDK5) may play a role in microtubule-associated protein tau (MAPT) phosphorylation and contribute to the pathogenesis of Alzheimer disease.[26]

2. Strategies for Drug Target Identification

2.1 Tools and Resources for Drug Target Identification

Drug target identification involves acquiring a molecular level understanding of a specific disease state and includes analysis of gene sequences, protein structures, protein-protein interactions, and metabolic pathways.[27,28] The ultimate goal of the process is to discover macromolecules that can become binding targets for lead compounds, each one a potential drug. In the age of genomics, the process of drug target identification needs to incorporate and integrate different sources of data including genetic, transcriptomic, proteomic, and metabolomic data. Relational databases are increasingly effective in facilitating pharmaceutical research and development as they broaden the range of analytical functions and expand the class of data models supported.

Table I lists some important databases for drug target identification. One of the most important resources is the human genome itself and associated annotations. The public data infrastructure is also as important as the data and includes algorithms for sequence analysis, gene expression analysis, proteomics analysis and that for protein structure prediction – one of the most computationally intensive exercises in the drug discovery process.[29] Although these resources represent a good general reference, they also possess significant limitations; most importantly, in many cases

Table I. Databases and web sites of interest for drug target identification

Database	Web address	Description	Reference
GPCRDB	http://www.gpcr.org/7tm	GPCR database	30
LGICdb	http://www.ebi.ac.uk/compneur-srv/LGICdb	Database of ligand-gated ion channels	31
IUPHAR	http://www.iuphar-db.org/iuphar-ic	The IUPHAR voltage-gated ion channels	32
MEROPS	http://www.merops.ac.uk	Peptidase database	33
KinG	http://hodgkin.mbu.iisc.ernet.in/~king/	Protein kinase database	34
NuclearDB	http://www.receptors.org/NR	Nuclear receptor database	35
NUREBASE	http://www.ens-lyon.fr/LBMC/laudet/nurebase	Database of nuclear hormone receptors	36
TTD	http://xin.cz3.nus.edu.sg/group/cjttd/ttd.asp	Therapeutic target database	37
AurSCOPE	http://www.aureus-pharma.com	Drug target databases	38
LifeSpan Drug Target™	http://www.lsbio.com'products/	Database of localization data for comprehensive families of drug targets	39
KEGG	http://www.genome.jp/kegg/	Kyoto Encyclopedia of Genes and Genomes	40
TRMP	http://xin.cz3.nus.edu.sg/group/trmp/trmp.asp	Therapeutically Relevant Multiple Pathways database	41
MetaCyc	http://metacyc.org/	Metabolic pathways and enzymes from various organisms	42
LIGAND	http://www.genome.ad.jp/ligand	Database of chemical compounds in biological pathways	43
OMIM	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM	Online Mendelian Inheritance in Man – catalog of human genes and genetic disorders	44
HPID	http://www.hpid.org	The Human Protein Interaction Database	45
PhRMA	http://www.phrma.org	Pharmaceutical Research and Manufacturers of America (Washington, DC)	46
BIND	http://www.blueprint.org/bind/bind.php	Biomolecular Interaction Network Database	47
HP-2DPAGE	http://www.mdc-berlin.de/~emu/heart/	2-dimensional electrophoresis database of human myocardial proteins	48
PharmGKB	http://www.pharmgkb.org	National Institutes of Health pharmacogenetics research database	49

the amount of data contained in them are insufficient to be used for effectively for identifying drug targets.

2.2 Bioinformatics Strategies for Drug Target Identification

With the development of bioinformatics, many computational techniques have been proposed for searching novel drug targets from genomic information. In this paper, the bioinformatics approaches for drug target identification are summarized as four classes: (i) the gene-to-target approach; (ii) the disease-to-target approach; (iii) the gene network approach; and (iv) the protein interaction network approach.

2.2.1 The Gene-to-Target Approach

Selection of a Certain Target Class

For the gene-to-target strategy, the first step is to select a common class of drug targets, then to design computational methods to find new members of this class and to predict their function based on available knowledge and information of the target class. Suitable target classes are those protein families whose members have been proven to be successful targets historically, such as GPCRs, ion channels, kinases, and nuclear hormone receptors. Here we select GPCRs as a case study based on the enormous amount of current pharmaceutical research aimed at understanding their structure and function.

Predicting New Target Genes from a Certain Class

Once we have chosen a particular class of targets, the next step is to screen sequence databases and identify all the possible candidates of that class. Recent studies demonstrated that discovering new members of a target class is important not only for finding useful drug targets but also for understanding the molecular basis of diseases. Early efforts to predict new targets of a particular class relied on two strategies.

- *Data mining the genome:* mining the human genome sequence can detect new protein coding genes and find new members of particular target classes.[27,50,51] GPCRs represent the most important target class for drug discovery. Many strategies have been used to identify novel GPCRs for various sequenced genomes. The common strategies attempt to find similar sequences of known GPCRs from sequence databases using primary database search tools (e.g. BLAST) or more sophisti-

cated ones that are coupled with the search of pattern databases such as PRINTS.[52,53] However, in many cases, these have not been sufficiently successful for the identification of GPCRs, since GPCRs make up a highly divergent family, with strikingly little sequence similarity shared between members. In order to overcome these limitations, other *in silico* approaches that incorporate such features as amino acid compositions, physiochemical properties,[54,55] and transmembrane topology patterns of GPCRs have been proposed.[56,57] In addition, the incorporation of *ab-initio* gene prediction techniques should also be useful in the discovery of new GPCR targets.[58]

- *Data mining the expressed sequence tags (ESTs):* genome-wide sequencing projects aim to identify all genes contained in genomes. The huge number of ESTs provides a valuable resource for gene identification, characterization, and tissue-specific gen expression analysis.[7,59,60] One of the most important applications of EST databases (e.g. dbEST) in target discovery is to indentify new genes of a target class and infer relative gene expression levels.[61] Wittenberger et al.[62] demonstrated a comprehensive EST database search method to identify new members of the GPCR superfamily. They found at least 14 ESTs that are promising candidates for new putative GPCRs, and five of them, namely GPR84, GPR86, GPR87, GPR90, and GPR91 sequences, were experimentally validated. Furthermore, it was found that GPR86 is central to the pathophysiology of hematopoiesis and immune system disease. Marvanová et al.[63] also investigated the use of ESTs as a starting point to map brain expression patterns and to identify potential novel drug targets. There are many factors that prevent ESTs from being widely exploited, including alternative splicing and the "error prone" characteristic of ESTs. Further studies are needed to tackle these problems in order for ESTs to be more fully utilized.[64]

Predicting the Function of New Genes

One essential requirement for a drug target to be useful is to understand its function.[27] The elucidation of gene function *in silico* is an important field for bioinformatics in target discovery. Nevertheless, determining protein function is one of the most challenging problems in the post-genomic era.

Functional annotation of completely sequenced genomes has proved to be a formidable task, and large segments of genes are as yet uncharacterized. Even in well studied genomes, such as *Escherichia coli*, ~30% of the genes are annotated as being of unknown function. In the malarial parasite *Plasmodium falciparum*, ~60% of genes lack functional assignments.[65] A significant limitation in understanding gene function is the lack of assays evaluating signal-specific cellular metabolic events downstream of the anticipated changes in gene expression and protein phosphorylation. The availability of entire genome sequences and high-throughput capabilities to determine gene function has shifted the research focus from the study of single proteins or small complexes to that of the entire proteome. However, the technology for discovering gene function is lagging behind the advances made in genomic sequencing.

Accurate computational function prediction, which is helpful for speeding up the functional annotation of gene products, has become an increasingly important problem in the field of bioinformatics.[66-69] By searching similar protein sequences with known function annotations, one can draw some inferences about the function of the uncharacterized gene. More sophisticated methods of incorporating sequence information such as sorting signals, post-translational modifications and domains to predict protein function have also been developed.

The function of a protein is highly correlated with its three-dimensional structure and the structural information is also very important to drug discovery and design. However, for many known protein sequences, their three-dimensional structure information is lacking. Therefore, further studies are needed to develop more accurate structure prediction methods and strategies of linking structure to function. Heterogeneous data should be integrated to take this problem. Recently, some researchers have initiated systems biology approach to predict target gene function.[70]

2.2.2 The Disease-to-Target Approach

Focusing on a Specific Disease

The identification of therapeutic targets requires knowledge of the etiology of a disease and the associated biological systems. The disease-to-target approach first focuses on a specific disease, or at least diseases in specific therapeutic categories. Then, various techniques such as gene expression analysis and linkage analysis are adopted to identify disease relevant genes and drug targets. Many pharmaceutical companies have focused on specific diseases for drug target identification.

Identifying Disease Relevant Genes and Drug Targets

Microarray technology, which can be used for measuring the expression levels of thousands of genes simultaneously in a single experiment and generating gene expression profiles can be utilized to discover disease relevant genes and drug targets.[71] Microarray experiments can not only identify novel candidate molecular targets and biochemical pathways that may be therapeutically exploited, but also increase our understanding of the biology of a disease process, and define how a specific compound affects the regulatory networks involved in cellular metabolism, or affects a specific cellular pathways, which may ultimately lead to the identification of other potential drug targets.[72] The first step is to compare the gene expression patterns in various disease stages of healthy tissue, and to identify those genes with differential expression in different conditions. The subsequent process then focuses on whittling down the candidate genes to those that seem central to the disease process, and whose products are likely to be amenable

to therapeutic intervention. For example, Cellzome Ltd (Heidelberg, Germany) has tried to identify and validate potential drug targets associated with Alzheimer disease by this strategy, and has developed a series of small-molecule gamma secretase modulators for the treatment of patients with this disease.

Separating genes causally involved in a disease from innocent bystander genes is a crucial problem in the analysis of disease expression profiles.[73] Many statistical methods have been proposed to detect the expression difference of single gene.[74] These methods generally produce long list of differentially expressed genes, but they provide few clues to which of these changes are important.[75] One promising method is to analyze the alterations of expression at functional level, such as biological pathways, which holds the tremendous potential to detect subtle but coordinate alterations in the expression of groups of functionally related genes and unveil the most relevant genes and functions that contribute to diseases. Currently, some tools are available to provide such analysis, e.g. OntoExpress,[76] GOAL,[77] and MageKey.[78]

For inherited diseases, analyzing chromosome regions that are linked to disease phenotypes can also identify the relevant genes and potential drug targets. The linkage analysis method has been widely used to locate disease loci.[79,80] Within the chromosome region of a disease locus mapped by this strategy, however, there are often hundreds of candidate genes. In order to find the disease-relevant gene, further experiments are needed to check the candidate genes for disease-causative mutations. Obviously, it will be very time-consuming and expensive if the candidate genes are randomly selected in the experimental search for disease-causative mutations. Therefore, the prediction of disease-relevant genes and prioritization of candidate genes for mutation analysis is one of the crucial steps in the identification of disease relevant genes.[68,81-83] Currently, the major information used to choose the candidate disease genes for mutation analysis include the gene function, gene expression patterns and features of gene sequences, based on which several computational methods and tools to predict disease relevant genes have also been developed in the recent years.[68,81-84] In addition, Pettipher et al.[85] used genetic association approach for the identification of GPCRs involved in inflammatory disease, leading to the identification of genetically associated targets, including TSHR, EDG6, and CRTH2. In this regard, the discovery of disease relevant genes may provide an essential starting point for drug discovery.

Building Predictive Disease Models

Moving all targets forward through development is prohibitive in terms of cost and time. From the perspective of drug target identification for human diseases, predictive disease models that are suitable for rigorous experimentation can support the case for discovery or validation of a target in humans. We cannot realistically hope to characterize all the relevant molecular interactions one-by-one as a requirement for building a predictive disease model. The ultimate goal of the disease model is to be able to model a disease process at the molecular level, to predict which specific chemical compounds are best suited to treating the disease for a genetically defined patient population, and to perform all binding experiments *in silico*. Intradigm Corporation (London and Cambridge, UK) has developed a unique and proprietary method that employs efficacy in animal disease models as a starting point for target discovery and validation. Intradigm's target discovery method, which combines gene perturbation of animal disease with pathway analysis, selectively and rapidly identifies those novel targets, operating in complex biological processes, which are activated as disease pathology expands or contracts. Achieving these goals *in silico* will dramatically improve drug discovery process and pave the way for personalized medicine based on a molecular level understanding of both the patient and the illness.

Phenotypes are generally difficult to recognize and validate, especially at the cellular level. Providing an association between phenotype and genotype is critical to being able to understand and create models of disease. This association is also the key to targeting critical pathways in disease and identifying the genes and proteins that regulate biological processes, thus identifying better drug targets. Predictive disease models that are suitable for rigorous experimentation can give well informed linkages between genotypes and individual phenotypes.[86]

Using Pharmacogenomics in Drug Target Identification

Inherited differences in drug targets, as well as polymorphisms in genes with a role in drug metabolism and disposition, have an influence on the efficacy and safety of therapeutics. The field of pharmacogenomics has the potential to lead to the identification of new drug targets, an improved understanding of the causes for variable drug response, and greater knowledge of the mechanistic basis for drug action and disease pathophysiology.[87,88] The critical strategy for a pharmaceutical company going forward is one that uses pharmacogenomics and biomedical informatics to better define disease targets. Until these are clear, and until some form of biomedical informatics is put into place, therapeutic design is going to be flawed by poorly defined targets.

The study of single nucleotide polymorphisms (SNPs) is crucial for characterizing molecular targets and can also validate the role of these targets in diseases.[89] SNP technology is expected to contribute substantially to the fields of pharmacogenomics and personalized medicine, disease mechanisms, and drug target discovery. An important prerequisite before these next generation achievements can be reached is the ability to analyze complex biological associations, and to identify their relevance for clinical problems. There are great expectations to the potential value in exploiting the accumulating amount of genetic/biological data.[90]

2.2.3 The Gene Network Approach

The aim of the network-based strategy is the reconstruction of endogenous metabolic, regulatory, and signaling networks with which potential drug targets interact. The reason is that if a drug target participates in many biological pathways, the inhibition of this target may interfere with many activities associated with those pathways and, therefore, it may not be a good candidate for drug target.

Genetic interactions are central to the understanding of the molecular structure and function, cellular metabolism, and response of organisms to their environments. If such interaction patterns can be measured for various kinds of tissues and the corresponding data interpreted, potential clinical benefits are obvious for diagnostics, identification of candidate drug targets, and predictions of drug effectiveness. It has already been shown that it is possible to infer a predictive model of a genetic network by overexpressing each gene of the network and measuring the resulting expression at steady state of all the genes in the network.[91] Using the inferred model, we can endeavor to make useful predictions by mathematical analysis and computer simulations. Model-based and computational analysis can open up a window on the physiology of an organism and disease progression. Recently, several computational methods have been proposed along with gene network models such as Boolean networks,[92] differential equation models,[93] and Bayesian networks,[94,95] to infer gene regulatory networks. These quantitative approaches can be applied to natural gene networks and used to generate a more comprehensive understanding of cellular regulation and elucidation of the underlying gene regulatory mechanisms.

2.2.4 The Protein Interaction Network Approach

Proteins are the principal targets of drug discovery. Protein expression in normal and diseased human tissue holds the key to developing more effective drugs to treat a wide variety of diseases. High-throughput proteomics, potentially identifying hundreds to thousands of protein expression changes in model systems following perturbation by drug treatment or disease, lends itself particularly well to target identification in drug discovery, and complements the genomics approach.

Protein-protein interaction data can be utilized in drug target identification.[96] Protein interaction maps can reveal novel pathways and functional complexes, allowing 'guilt by association' annotation of uncharacterized proteins and ascribing the role of these proteins into biochemical pathways and networks. Generation of a comprehensive human protein interaction map would facilitate identification of proteins that could be targeted for therapeutic and diagnostic applications. Once the pathways are mapped, these need to be analyzed and validated functionally in a biological model.[97] There are numerous studies aimed at mapping pathways that are involved in disease processes. The goal is to identify the key nodes in a complex network of genes and proteins (and small metabolites, i.e. the metabolome) that can serve as drug targets.

Using an *in situ* proteomics technology involving whole cell imaging, MelTec GmbH & Co. KG (Magdeburg, Germany) have tried to predict key proteins involved in disease pathways. The ability to monitor pathways with subcellular resolution in the proper tissue context further increases the significance of target predictions, and the ability to associate specific proteins with disease and to localize those proteins within tissues at the cellular level is of critical importance for identifying and validating drug targets.

Although a diagrammatic representation of the information on a pathway facilitates the understanding of the network topology and identification of drug targets, its capacity to predict cell behavior in response to an environmental or genetic change is very limited.[98] The identification of proteins is only the beginning of the process; data analysis and validation of potential protein targets that follows is a time-consuming and labor-intensive process as well.

With our ever-increasing understanding of the complexity of human protein interactions that impact directly on the safety and efficacy of therapeutic interventions, new technology in systems biology allows synergistic interpretation of both types of data in the context of functional networks. Technically, cellular processes are presented as 'interactome', an interconnected network of signaling, regulatory, and biochemical modules and pathways. Using a representative set of human protein-protein interaction data, the network analysis enables a comprehensive view of disease-implicated pathways which enables the discovery and validation of modules and pathways involving disease-specific protein drug targets as specific network modules.

Powerful bioinformatics software enables rapid interpretation of protein-protein interactions, comparative pathway analysis, accelerated functional assignment, and drug target discovery.[99,100] Current challenges to fully exploit the available experimental proteomics data include the integration of information available across several databases and the in-depth characterization of the data using new and advanced algorithms.

2.3 Systems Biology and Drug Target Discovery

The fact that the total number of genes in the human genome is surprisingly small suggests that much of the complexity of human biology resides outside the DNA sequence itself. The recent availability of large-scale heterogeneous (genomic, proteomic, and metabolomic) data is responsible for the major growth spurt of systems biology. Systems biology – that is, the computational integration of data generated by the suite of genetic, transcriptomic, proteomic, and metabonomic platforms to understand function through different levels of biomolecular organization –

offers exciting new prospects for determining the causes of human disease and finding possible cures.[101,102]

Systems biology is currently one of the hottest areas of biotechnology research today and is becoming central to the strategy of many biopharmaceutical and genomic companies. There is little doubt that biomedicine and the pharmaceutical industry stand to be significant beneficiaries of the promise of systems biology. The Cambridge-Massachussets Institute of Technology (MIT) Institute (CMI) brings together two of the world's leading universities in a dynamic and unique academic partnership to further the application of systems biology and stem cell research to the study and identification of drug targets in complex diseases such as cancer and inflammation. Bioseek, Inc. (Burlingame, CA, USA) has also developed quantitative, automated primary human 'cell systems biology' models of inflammation, autoimmunity, and cardiovascular disease that embody disease-relevant complexity for drug discovery.

Systems biology aims at understanding complex biological networks through a combination of (comprehensive) experimental analysis and (quantitative) mathematical modeling.[103,104] At present, however, it is largely unclear which knowledge and data will be required for establishing realistic mathematical models. Related to this, it is equally important to ask to what extent the already available data allow for meaningful model development. Therefore, systems biology will also encompass the development of tools and experimental approaches to produce quantitative data. This type of information will help us to better understand diseases and, hence, systems biology will become an integral part of drug target identification.

3. Discussion

Whether the number of actual drug targets is correct or not, the currently available data strongly suggest that the present number of known and well validated drug targets is still relatively small. Bioinformatics is making practical contributions in identifying large numbers of potential drug targets; however, target validation efforts are required to link them to the etiology of known diseases and/or to demonstrate that the novel targets have relevant therapeutic potential. A number of problems are still present in the current approaches.

- An increasing number of bioinformatics tools coupled with the lack of an integrated and systematized interface for their selection and utilization is becoming widely acknowledged.
- The processing and exploitation of useful information from genomics data pose a challenging problem. Sophisticated bioinformatics platforms should be constructed for integrating genetic and gene expression data and their use in the selection of genes as novel targets.
- The identification and validation of drug targets depends critically on knowledge of the biochemical pathways in which potential target molecules operate within cells.

Although database mining and transcriptional profiling clearly have increased the number of putative targets, the current focus is to assign function to new gene targets in a high-throughput manner. This requires a restructuring of the classical linear progression from gene identification, functional elucidation, target validation, and screen development. For this reason, the complexity of the drug discovery process in the post-genome era requires the application of integrated approaches for the rapid advancement of target-to-drug. The study of biochemical pathways is the focus of numerous drug discovery research efforts, and is central to the strategy of drug target identification.

4. Concluding Remarks

Up until about 20 years ago, drug discovery was chemistry-driven, conducted by trial-and-error, and had a paucity of defined targets. Genomics and proteomics technologies have created a paradigm shift in the drug discovery process. With the complete sequencing of the human genome, it is now possible to think of the whole pharmaceutical process as a computational approach, with confirmatory experiments at each decision-point. Genomics-based drug discovery and development is reliant on sophisticated bioinformatics and data management tools. As the molecular dynamics data become more copious and complex, we may need to develop new *in silico* methods to provide the reliable, guiding hypotheses for experimental design.

It should be emphasized that although bioinformatics tools and resources can be used to identify putative drug targets, validating targets is still a process that requires understanding the role of the gene or protein in the disease process and is heavily dependent on laboratory-based work. Target validation needs high-throughput screening and selectivity analysis for therapeutic compounds that inhibit a mutant protein involved in the disease process.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (grants no. 90203011 and no. 30370354), and the Ministry of Education of China (grant no. 505010).

References

1. Drews J. Genomic sciences and the medicine of tomorrow. Nat Biotechnol 1996; 14: 1516-8
2. Drews J. Drug discovery: a historical perspective. Science 2000; 287: 1960-4
3. Terstappen GC, Reggiani A. *In silico* research in drug discovery. Trends Pharmacol Sci 2001; 22: 23-6
4. International Human Genome Sequence Consortium. Finishing the euchromatic sequence of the human genome. Nature 2004; 431: 931-45
5. Zambrowicz BP, Sands AT. Knockouts model the 100 best-selling drugs: will they model the next 100? Nat Rev Drug Discov 2003; 2: 38-51
6. Searls DB. Using bioinformatics in gene and drug discovery. Drug Discov Today 2000; 4: 135-43

7. Whittaker PA. What is the relevance of bioinformatics to pharmacology? Trends Pharmacol Sci 2003; 24: 434-9
8. Dahl SG, Kristiansen K, Sylte I. Bioinformatics: from genome to drug targets. Ann Med 2002; 34: 306-12
9. Allison DB. Statistical methods for microarray research for drug target identification 2002. Proceedings of the American Statistical Association, Biopharmaceutical Section [CD-ROM]. Alexandria (VA): American Statistical Association, 2002
10. Sands AT. The rule of three: selecting the best new drug targets. Drug Discovery & Development 2003 Mar [online]. Available from URL: http://www.dddmag.com/ [Accessed 2005 Oct 10]
11. Drews J, Ryser S. Classic drug targets. Nat Biotechnol 1997; 15: 1318-9
12. Bouvier M. Structural and functional aspects of G protein-coupled receptor oligomerization. Biochem Cell Biol 1998; 76: 1-11
13. Nambi P, Aiyar N. G protein-coupled receptors in drug discovery. Assay Drug Dev Technol 2003; 1: 305-11
14. Fredriksson R, Schioth HB. The repertoire of G-protein-coupled receptors in fully sequenced genomes. Mol Pharmacol 2005; 67: 1414-25
15. Palczewski K, Kumasaka T, Hori T, et al. Crystal structure of rhodopsin: a G protein-coupled receptor. Science 2000; 289: 739-45
16. Clapham DE. TRP channels as cellular sensors. Nature 2003; 426: 517-24
17. Terstappen GC. Ion channel screening technologies today. Drug Discov Today 2005; 2: 133-40
18. Cahalan MD, Chandy KG. Ion channels in the immune system as targets for immunosuppression. Curr Opin Biotechnol 1997; 8: 749-56
19. Bennett PB, Guthrie HRE. Trends in ion channel drug discovery: advances in screening technologies. Trends Biotechnol 2003: 21, 563-9
20. Duarte J, Perrière G, Laudet V, et al. NUREBASE: database of nuclear hormone receptors. Nucl Acids Res 2002; 30: 364-8
21. Inpharmatica successfully closes GBP13.9m (USD25m) third round financing. 2004 Nov 8 [press release; online]. Available from URL: http://www.inpharmatica.com/news/2004/081104.htm [Accessed 2005 Nov15]
22. Docherty AJ, Crabbe T, O'Connell JP, et al. Proteases as drug targets. Biochem Soc Symp 2003; 70: 147-61
23. Leung D, Abbenante G, Fairlie DP. Protease inhibitors: current status and future prospects. J Med Chem 2000; 43: 305-41
24. Cohen P. Protein kinases: the major drug targets of the twenty-first century. Nat Rev Drug Discov 2002; 1: 309-15
25. Shears SB. Rounding up the usual suspects: protein kinases as therapeutic targets. Drug Discovery Today 2005; 10: 240-2
26. Lau LF, Seymour PA, Sanner MA, et al. Cdk5 as a drug target for the treatment of Alzheimer's disease. J Mol Neurosci 2002; 19 (3): 267-73
27. Smith C. Drug target identification: a question of biology. Nature 2004; 428: 225-31
28. Augen J. The evolving role of information technology in the drug discovery process. Drug Discov Today 2002; 7: 315-23
29. Head-Gordon T, Wooley JC. Computational challenges in structural and functional genomics. IBM Systems J 2001; 40 (2): 265-96
30. Horn F, Bettler E, Oliveira L, et al. GPCRDB information system for G protein-coupled receptors. Nucleic Acids Res 2003; 31: 294-7
31. Le Novère N, Changeux J-P. LGICdb: the ligand-gated ion channel database. Nucleic Acids Res 2001; 29: 294-5
32. IUPHAR receptor database [online]. Available from URL: http://www.iuphar-db.org/iuphar-rd [Accessed 2005 Oct 10]
33. Rawlings ND, O'Brien E, Barrett AJ. MEROPS: the protease database. Nucleic Acids Res 2002; 30: 343-6
34. Krupa A, Abhinandan KR, Srinivasan N. KinG: a database of protein kinases in genome. Nucleic Acids Res 2004; 32: 153-5
35. Horn F, Vriend G, Cohen FE. Collecting and harvesting biological data: the GPCRDB and NucleaRDB information systems. Nucleic Acids Res 2001; 29: 346-9
36. Duarte J, Perrière G, Laudet V, Robinson-Rechavi M. NUREBASE: database of nuclear hormone receptors Nucleic Acids Res 2002; 30: 364-368
37. Chen X, Ji ZL, Chen YZ. TTD: therapeutic target database. Nucleic Acids Res 2002; 30: 412-5
38. Aureus Pharma website [online]. Available from URL: http://www.aureus-pharma.com [Accessed 2005 Oct 10]
39. LifeSpan Biosciences website [online]. Available from URL: http://www.lsbio.com [Accessed 2005 Oct 10]
40. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 2000; 28: 27-30
41. Zheng CJ, Zhou H, Xie B, et al. TRMP: a database of therapeutically relevant multiple-pathways. Bioinformatics 2004; 20 (14): 2236-41
42. Krieger CJ, Zhang P, Mueller LA, et al. MetaCyc: a multiorganism database of metabolic pathways and enzymes. Nucleic Acids Res 2004; 32: 438-42
43. Goto S, Okuno Y, Hattori M, et al. LIGAND: database of chemical compounds and reactions in biological pathways. Nucleic Acids Res 2002; 30: 402-4
44. Hamosh A, Scott AF, Amberger J, et al. Online mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 2002; 30: 52-5
45. Human Protein Interaction Database [online]. Available from URL: http://www.hpid.org [Accessed 2005 Oct 10]
46. PhRMA website [online]. Available from URL: http://www.phrma.org/ [Accessed 2005 Oct 10]
47. Bader GD, Betel D, Hogue CWV. BIND: the biomolecular interaction network database. Nucleic Acids Res 2003; 31: 248-50
48. Heart High-Performance 2-DE Database [online]. Available from URL: http://www.mdc-berlin.de/~emu/heart/ [Accessed Nov 15]
49. Hewett M, Oliver DE, Rubin DL, et al. PharmGKB: the pharmacogenetics knowledge base. Nucleic Acids Res 2002; 30: 163-5
50. Foord SM. Receptor classification: post genome. Curr Opin Pharmacol 2002; 2: 561-6
51. Manning G, Whyte DB, Martinez R, et al. The protein kinase complement of the human genome. Science 2002; 298: 1912-34
52. Lapinsh M, Gutcaits A, Prusis P, et al. Classification of G-protein coupled receptors by alignment-independent extraction of principal chemical properties of primary amino acid sequences. Protein Sci 2002; 11: 795-805
53. Takeda S, Kadowaki S, Haga T, et al. Identification of G protein-coupled receptor genes from the human genome sequence. FEBS Lett 2002; 520: 97-101
54. Karchin R, Karplus K, Haussler D. Classifying G-protein coupled receptors with support vector machines. Bioinformatics 2002; 18: 147-59
55. Bhasin M, Raghava GP. GPCRpred: an SVM-based method for prediction of families and subfamilies of G-protein coupled receptors. Nucleic Acids Res 2004; 32: W383-9
56. Sugiyama Y, Poluliakh N, Shimizu T. Identification of transmembrane protein functions by binary topology patterns. Protein Eng 2003; 16: 479-88
57. Inoue Y, Ikeda M, Shimizu T. Proteome-wide classification and identification of mammalian-type GPCRs by binary topology pattern. Comput Biol Chem 2004; 28: 39-49
58. Zhou YH, Yang L, Wang H, et al. Prediction of eukaryotic gene structures based on multilevel optimization. Chin Sci Bull 2004; 49 (4): 321-8
59. Adams MD, Kerlavage AR, Fields C, et al. 3,400 new expressed sequence tags identify diversity of transcripts in human brain. Nat Genet 1993; 4: 256-67
60. Allikmets R, Gerrard B, Glavac D, et al. Characterization and mapping of three new mammalian ATP-binding transporter genes from an EST database. Mamm Genome 1995; 6: 114-7
61. Boguski MS, Lowe TM, Tolstoshev CM. dbEST: database for "expressed sequence tags". Nat Genet 1993; 4 (4): 332-3
62. Wittenberger T, Schaller HC, Hellebrand S. An expressed sequence tag (EST) data mining strategy succeeding in the discovery of new G-protein coupled receptors. J Mol Biol 2001; 307: 799-813
63. Marvanová M, Törönen P, Storvik M, et al. Synexpression analysis of ESTs in the rat brain reveals distinct patterns and potential drug targets. Mol Brain Res 2002; 104: 176-83
64. Zhou YH, Jing H, Li YE, et al. Identification of true EST alignments and exon regions of gene sequences. Chin Sci Bull 2004; 49 (23): 2463-9
65. Aggarwal K, Lee KH. Functional genomics and proteomics as a foundation for systems biology. Brief Funct Genomic Proteomic 2003; 2: 175-84
66. Gabaldon T, Huynen MA. Prediction of protein function and pathways in the genome era. Cell Mol Life Sci 2004; 61: 930-44
67. Hennig S, Groth D, Lehrach H. Automated gene ontology annotation for anonymous sequence data. Nucleic Acids Res 2003; 31: 3712-5
68. Jensen LJ, Gupta R, Staerfeldt HH, et al. Prediction of human protein function according to gene ontology categories. Bioinformatics 2003; 19: 635-42
69. Schug J, Diskin S, Mazzarelli J, et al. Predicting gene ontology functions from ProDom and CDD protein domains. Genome Res 2002; 12: 648-55
70. Guffanti A. Modeling molecular networks: a systems biology approach to gene function. Genome Biol 2002; 3 (10): 4031.1-4031.3

71. Meltzer PS. Spotting the target: microarrays for disease gene discovery. Curr Opin Genet Dev 2001; 11: 258-63

72. Marton MJ, De Risi JL, Bennet HA, et al. Drug target validation and identification of secondary drug target effects using DNA microarrays. Nat Med 1998; 4: 1293-301

73. Miklos GL, Maleszka R. Microarray reality checks in the context of a complex disease. Nat Biotechnol 2004; 22: 615-21

74. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A 2001; 98: 5116-21

75. Schulze A, Downward J. Navigating gene expression using microarrays: a technology review. Nat Cell Biol 2001; 3: E190-5

76. Draghici S, Khatri P, Martins RP, et al. Global functional profiling of gene expression. Genomics 2003; 81: 98-104

77. Volinia S, Evangelisti R, Francioso F, et al. GOAL: automated Gene Ontology analysis of expression profiles. Nucleic Acids Res 2004; 32: W492-9

78. Wang WQ, Zhou YH, Bi R. Correlating genes and functions to diseases by systematic differential analysis of expression profiles. Lecture Notes Comput Sci 2005; 3645: 11-20

79. Pericak-Vance MA, Yamaoka LH, Haynes CS, et al. Genetic linkage studies in Alzheimer's disease families. Exp Neurol 1988; 102: 271-9

80. Corder EH, Saunders AM, Strittmatter WJ, et al. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. Science 1993; 261: 921-3

81. Freudenberg J, Propping P. A similarity-based method for genome-wide prediction of disease-relevant human genes. Bioinformatics 2002; 18 Suppl. 2: S110-5

82. Van Driel MA, Cuelenaere K, Kemmeren PP, et al. A new web-based data mining tool for the identification of candidate genes for human genetic disorders. Eur J Hum Genet 2003; 11: 57-63

83. Perez-Iratxeta C, Bork P, Andrade MA. Association of genes to genetically inherited diseases using data mining. Nat Genet 2002; 31: 316-9

84. Zhou YH, Zhou QX, Liu HL, et al. Predicting disease genes for familial dilated cardiomyopathy based on the codon usage bias. Chin Sci Bull 2005; 50 (18): 2028-32

85. Pettipher R, Mangion J, Hunter MG. et al. Identification of G-protein-coupled receptors involved in inflammatory disease by genetic association studies. Curr Opin Pharmacol 2005: 5, 412-7

86. Richard M, Bruskiewich AB, Cosico WE, et al. Linking genotype to phenotype: the International Rice Information System (IRIS). Bioinformatics 2003; 19: 63-5

87. Lindpaintner K. Science and society: the impact of pharmacogenetics and pharmacogenomics on drug discovery. Nature Rev Drug Discovery 2002; 1: 463-9

88. Lindpaintner K. Pharmacogenetics and pharmacogenomics in drug discovery and development: an overview. Clin Chem Lab Med 2003; 41 (4): 398-410

89. Campbell DA, Valdes AM, Spurr N. Making drug discovery a SN(i)P. Drug Discov Today 2000; 5: 388-96

90. Evans WE, McLeod HL. Pharmacogenomics: drug disposition, drug targets, and side effects. N Engl J Med 2003; 348: 538-49

91. Cabusora L, Sutton E, Fulmer A, et al. Differential network expression during drug and stress response. Bioinformatics 2005; 21: 2898-905

92. Shmulevich I, Dougherty ER, Kim S, et al. Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks. Bioinformatics 2002; 18: 261-74

93. Akutsu T, Miyano S, Kuhara S. Inferring qualitative relations in genetic networks and metabolic pathways. Bioinformatics 2000; 16: 727-34

94. Imoto S, Savoie CJ, Aburatani S, et al. Use of gene networks for identifying and validating drug targets. J Bioinform Comput Biol 2003; 1: 459-74

95. Gardner TS, di Bernardo D, Lorenz D, et al. Inferring gene networks and identifying compound mode of action via expression profiling. Science 2003; 301: 102-5

96. Stagljar I, Hottiger MO, Auerbach D, et al. Protein-protein interactions as a basis for drug target identification. Innov Pharm Technol 2002; 1: 66-9

97. Rain JC, Selig L, De Reuse H, et al. The protein-protein interaction map of *Helicobacter pylori*. Nature 2001; 409: 211-5

98. Luan Y, Li H. Model-based methods for identifying periodically expressed genes based on time course microarray gene expression data. Bioinformatics 2004; 20: 332-9

99. Nariai N, Kim S, Imoto S, et al. Using protein-protein interactions for refining gene networks estimated from microarray data by Bayesian networks. Pac Symp Sympos Biocomput World Sci 2004; 9: 336-47

100. Bader JS, Chaudhuri A, Rothberg JM, et al. Gaining confidence in high-throughput protein interaction networks. Nat Biotechnol 2004; 22: 78-85

101. Butcher EC, Berg EL, Kunkel EJ. Systems biology in drug discovery. Nat Biotechnol 2003; 22: 1253-9

102. Hood L, Galas D. The digital code of DNA. Nature 2003; 421: 444-8

103. Nicholson JK, Wilson ID. Understanding 'global' systems biology: metabonomics and the continuum of metabolism. Nat Rev Drug Discov 2004; 2: 668-76

104. Parsons AB, Brost RL, Ding H, et al. Integration of chemical-genetic and genetic interaction data links bioactive compounds to cellular target pathways. Nat Biotechnol 2003; 22: 62-9

Correspondence and offprints: Dr *Zhenran Jiang*, School of Computer Science, Huazhong University of Science and Technology, Wuhan 430074, China.
E-mail: jiangzhenran@163.com