

CHALLENGES FOR BIOMEDICAL INFORMATICS AND PHARMACOGENOMICS

Russ B. Altman and Teri E. Klein

Stanford Medical Informatics, Stanford, California 94305-5479;
e-mail: russ.altman@stanford.edu, teri.klein@stanford.edu

Key Words bioinformatics, pharmacogenetics, computation, databases

■ **Abstract** Pharmacogenomics requires the integration and analysis of genomic, molecular, cellular, and clinical data, and it thus offers a remarkable set of challenges to biomedical informatics. These include infrastructural challenges such as the creation of data models and databases for storing these data, the integration of these data with external databases, the extraction of information from natural language text, and the protection of databases with sensitive information. There are also scientific challenges in creating tools to support gene expression analysis, three-dimensional structural analysis, and comparative genomic analysis. In this review, we summarize the current uses of informatics within pharmacogenomics and show how the technical challenges that remain for biomedical informatics are typical of those that will be confronted in the postgenomic era.

WHAT IS BIOMEDICAL INFORMATICS?

Biomedical informatics is the study of information flow within biology and medicine. The use of computational techniques in biomedical research dates back to the first general purpose computers but interest in the techniques has exploded in the last decade (1). The increased interest stems from the availability of experimental techniques that create data that simply cannot be manually analyzed and require computational intervention. Many areas of biology and medicine are being revolutionized by the introduction of new experimental techniques, accompanied by informatics methodologies that fundamentally change the way that investigators do their work.

The two flows of information that are studied by informatics are the flow of information from the DNA code to biological function and the flow of information in the design and analysis of experiments. In the first flow, we are interested in the transfer of information *within* biology, while in the other, we are interested in the transfer of information *about* biology. Thus, the first information flow deals with

the central dogma of biology: DNA is transcribed into RNA, RNA is translated into protein, and protein molecules have functions that carry out biological processes. Interacting proteins produce signaling and metabolic pathways that coalesce to form networks at the cellular level, and cells interact at an organismal level to produce physiology. Informatics approaches to studying different aspects of this flow, therefore, include methods for gene finding (2–5), 3D structure prediction (6, 7), modeling of genetic networks (8–14), and statistical population biology (15, 16).

In the second flow, we are interested in the ways in which biological and medical information is gathered. This flow begins with a scientific hypothesis, followed by a plan to collect data, execution of an experiment, analysis of the results, and subsequent refinement of the hypothesis. Informatics applications within this flow are usually created to support investigators in the practice of science. Informatics approaches to studying this flow, therefore, include methods for organizing and searching databases of literature, sequence, and function, as well as methods for helping to create and evaluate scientific models (17). If both of these information flows are included in a definition of biomedical informatics, then virtually all biomedical informatics research can be placed in one or both of these areas.

Biomedical informatics has gained prominence recently because biologists can now collect more data. The success of the genome sequencing projects has catalyzed a new way of thinking in biology, whereby data are collected on a large scale and without a particular hypothesis in mind. The data are then placed in a database, and scientists with hypotheses can extract information from the database in order to evaluate the merits of the hypotheses. This leads to a fundamental change in how some investigators do their work: Instead of first moving to the laboratory, they first move to the database, and only after assessment of the available data are experiments planned. There has been much debate about the merits of such an approach, but there is no doubt that the emergence of these large-scale and high-throughput methods for data collection makes such an approach feasible (18). The data explosion is not limited to DNA sequencing, and we are seeing increased capacity to assess the levels of mRNA expression (19, 20), to detect protein-protein interactions (21), to locate gene products within the cell (22), to detect and identify compounds using mass spectroscopy (23), and even to understand the detailed atomic three-dimensional structure of macromolecules and their small molecule ligands (24).

As long as clever experimentalists continue to create these high-throughput experimental methods, informatics professionals will have a surfeit of data and data analytic challenges. Success in informatics usually means the acceleration in understanding the processes of interest, and increased access to the information required to generate and test scientific hypotheses. One of the areas that has recently attracted the attention of biomedical informaticians is pharmacology, and particularly pharmacogenetics and pharmacogenomics.

WHAT ARE PHARMACOGENETICS AND PHARMACOGENOMICS?

Pharmacogenetics is the study of how variation in genes affects the response to drugs. It has existed as a field for more than four decades, and forms the basic intellectual framework for understanding phenomena such as the idiosyncratic responses to anesthesia, to opiates, and to anticancer agents (25). Pharmacogenetics has tended to study single genes with a focused, hypothesis-directed set of experiments. Most pharmacogenetic studies begin with the recognition of high variability in response to a medication and then a search for the genetic basis of the variation. Thus, for example, the blood levels of a metabolite may be measured and noted to vary widely, investigations of the pathway of metabolism may suggest that enzymes in this pathway are behaving differently, and the genetic analysis of these enzymes might detect variations in the protein sequence (or regulatory sequence) that explain different catalytic rates or binding constants.

Pharmacogenomics emerged recently, and scientists do not entirely agree on its relationship to pharmacogenetics (26–29). “Genomics” is generally used to indicate the study of the entire complement of genes within an organism, and the -omics suffix has been used generally to indicate the comprehensive analysis of the capabilities of an organism. Thus, proteomics studies the full set of proteins and how they interact within a cell. Based on this understanding, pharmacogenomics can be construed as the study of the entire complement of pharmacologically relevant genes, how they manifest their variations, how these variations interact to produce phenotypes, and how these phenotypes affect drug response. A key element of pharmacogenomics is, not surprisingly, the large-scale and high-throughput collection of data, including DNA sequence variations, mRNA expression analysis, enzyme kinetic assays, and cellular localization experiments. The move toward these types of experiments of course creates a magnet for biomedical informatics investigators, who see an opportunity to apply their methodologies to an exciting area with promise to revolutionize medical care.

The development of pharmacogenomics is a natural sequela to the success of the initial human genome sequencing project. The promise of that project was that an understanding of all human genes would create the opportunity for new diagnostic, prognostic, and therapeutic technologies. The variation in response to medications across patients can be large, and the occurrence of side effects and adverse events limits the success of many therapeutic strategies. A systematic understanding of the gene systems that modulate response to medications may therefore change the way medications are prescribed. With the success of pharmacogenomics, it may become possible routinely to check the genetic background of a patient in order to ensure that the prescribed medications are effective and free from adverse side effects (30–33). Although it is not entirely clear how many of the 35,000 genes assigned in the rough draft of the human genome are relevant to drug response (or even how to define “relevance”), a systematic analysis of pharmacology

textbooks indicates a core set of 500 to 1000 genes, as shown in the PharmGKB Web site.¹

WHAT ARE THE APPROACHES TO PHARMACOGENOMICS?

There are generally two approaches to pharmacogenomic research, which are summarized as the “genotype-to-phenotype” and the “phenotype-to-genotype” approaches. In the genotype-to-phenotype approach, the investigators start with a set of genes that are known (or strongly suspected) to be important in modulating the response to drugs, and then they search for variation in their sequences (that is, their genotype). Given an understanding of genetic variation, they can search for the phenotypic consequences. Examples of approaches amenable to genotype-to-phenotype analysis might include gene families known to be important for pharmacokinetics (the study of how medications are absorbed, distributed, and cleared from the body), such as phase I metabolism enzymes (the mixed function oxygenases of the cytochrome p450 system) (34), phase II metabolism enzymes (the conjugation system) (35), and membrane transporter molecules (36). Other systems amenable to the genotype-to-phenotype approach are those that are involved in pharmacodynamics (the study of how medications have their therapeutic effect) and those whose mechanisms are well understood at the receptor and pathway level. Examples might include the well described pathways of inflammation in asthma (37, 38), the purine/pyrimidine biosynthetic pathways that are targeted by some anticancer agents (39), or the enzyme cascade that controls blood clotting (40). The steps of a genotype-to-phenotype approach can be summarized in this simplified way:

1. Identify the genes that belong to the system that is involved in modulating drug response.
2. Catalog the variation in the DNA sequences across the population.
3. Search for phenotypes associated with the sequence variation.
4. Confirm clinical relevance of the genotype-phenotype associations.

Each of these steps is nontrivial and complicated. The first involves searching DNA databases and perhaps using comparative genomic techniques to identify target genes (41, 42). The second includes high-throughput experimental methods for detecting DNA variations, including single nucleotide polymorphisms (SNPs), the most common type of DNA variation (43–45), and also for associating individual SNPs into haplotypes (46). The third step involves the collection of molecular, cellular, or clinical data (reviewed below), and the final step requires clinical trials to prove the associations of interest and to demonstrate clinical relevance.

¹<http://www.pharmgkb.org/>

Phenotype-to-Genotype Approaches

Phenotype-to-genotype approaches toward pharmacogenomic discovery are different. Instead of identifying a family of genes in which to characterize genetic variation, investigators search for a phenotypic measure that shows significant variation. This measure can be a clinical measure (such as the rate of clearance of a drug or the peak level of the drug for a given dose), a cellular measure (the rate of cellular uptake of a drug or the profile of gene expression), or a molecular measure (the enzymatic turnover rate of an enzyme or a substrate binding constant). In any case, it is the phenotypic variation that first draws attention and then follows a search for the genes that are responsible for this variation. The steps of a phenotype-to-genotype approach, therefore, can thus be summarized:

1. Identify a phenotype that shows significant variation.
2. Search for genes that may explain this variation.
3. Characterize genetic variations and check for association with the phenotype.
4. Confirm proposed genetic basis for the variation and its clinical relevance.

The challenges in the first step are to identify phenotypes that are both clinically relevant and also measurable. The second step is the most difficult and requires the investigator to use any means available to identify genes that could be involved with the phenotypes. It may involve using animal models and comparative genomics, DNA microarray analysis to measure changes in expression in response to drugs, database (literature and sequence) searches for associations between genes and related phenotypes, or analytic chemistry methods to identify gene products contributing to variation (47). The third step is similar to the second step of the genotype-to-phenotype process. A major challenge in this step is the large amount of variability in human genes that is not functionally significant, so investigators must focus efforts on variations that can be shown to have functional consequence. The final step is focused particularly on this problem of ensuring that the discovered genetic component really explains the phenotypic variation of interest.

Both approaches to pharmacogenomics have strengths and weaknesses. Investigators must assess the current knowledge base for a given drug class of interest in order to determine whether there is enough genetic information to justify a genotype-to-phenotype approach, or whether there are more striking phenotypic data suggesting a phenotype-to-genotype approach.

CHALLENGES FOR BIOMEDICAL INFORMATICS IN PHARMACOGENOMICS

The challenges for biomedical informatics within the study of pharmacogenomics all follow directly from the preceding discussion. Pharmacogenomics is relatively new, so the current excitement derives, in part, from the great range of opportunity

for contributions that now exists. One of the key themes in pharmacogenomics is that the relevant informatics expertise includes information from molecular biology (sequences, structures, pathways) as well as from clinical medicine (medications, diseases, side effects), and of course from pharmacology (pharmacokinetics and pharmacodynamics). Thus it represents a new wave of informatics problems where both basic biological and clinical information must be combined and analyzed. Whereas previously bioinformatics focused solely on issues of relevance to molecular biology (sequence and structure analysis), applications are now moving closer to the parts of clinical informatics that focus on the organization of clinical information, particularly for research purposes. The main challenges for biomedical informatics within pharmacogenomics fall into nine areas:

1. Representing the diversity of pharmacogenomic data
2. Developing standards for data exchange
3. Integrating data from multiple data resources
4. Mining literature for knowledge
5. Using expression data to understand regulation
6. Understanding the structural basis for variability
7. Using comparative genomics
8. Managing laboratory information
9. Protecting sensitive patient information

Representing the Diversity of Pharmacogenomic Data

One of the principal challenges for pharmacogenomics is the creation of data structures that store relevant information in a form that is easy for computer programs to manipulate. There is a difference between data formats that are useful for human readers (journals, tables, figures) and those that are useful for computers (data structures in computer programs that label all data for easy retrieval and analysis). The classes of data that must be represented are diverse, as are the connections between the data that must be maintained.

GENOMIC DATA The representation of DNA sequence information for pharmacogenomics is similar to that required for many other applications. The main requirement is that the gene structure for a protein product be understood and labeled so that observed DNA sequence variations can be interpreted as belonging to coding or noncoding regions, and their likely significance can be evaluated. The key concepts that must be modeled include genomic sequence, unprocessed mRNA transcript, processed transcript, and protein sequence. Within each of these models are details such as the 3'- and 5'-untranslated regions of genes, genetic regulators (enhancers, silencers), exons that are coding or noncoding (or partially coding), and alternative splicing strategies. There have been a number of proposed standards for tracking genomic data, including the data

structures behind Genbank (48), Human Genome Database (49), BIOML², and others. The human genome browsers offered by UC Santa Cruz³, Ensembl⁴, National Center for Biotechnology Information (NCBI)⁵, and Celera⁶ offer a basic look at gene structure, but these are still evolving because the genome is in draft. In addition to the representation of basic gene structure, it is critical to also understand the locations and types of genetic variation. The dbSNP resource at NCBI provides an excellent source of reported SNPs (50), including those submitted by The SNP Consortium (51), an industrial group that is performing large-scale SNP detection and submitting many of these to the public domain.

Genome data are also made more useful by their connections to databases of biological function, including the Online Mendelian Inheritance in Man (OMIM) database of inherited human disorders (52, 53), and a number of specialty databases that provide valuable in-depth information about individual gene families, such as the Cell Signaling Network Database (54), the transcription factor database TRANSFAC (55), and the protein kinase database (56).

MOLECULAR AND CELLULAR DATA The characterization of phenotype is important for both the genotype-to-phenotype methods as well as the phenotype-to-genotype methods. Phenotype is difficult to precisely define, but it can be thought of as functional features of gene products, ranging in detail from the molecular to the individual and population levels. Unfortunately, phenotype data are not as “digital” as sequence data, so they are much more difficult to represent. Nevertheless, the success of pharmacogenomics depends on the establishment of standards for describing these data.

In pharmacogenomics, a few types of molecular and cellular data are clearly critical to represent. These include enzyme kinetic data (such as the binding and catalytic constants of enzymes, and their associated kinetic parameters), three-dimensional structural data (when available) for enzymes and their substrates/ligands, and protein localization data (often images) that show where different gene products are found within the cell. Standard representations of these data types are not generally available, but the creation of databases to store such information will require that they be developed. Fortunately, there is fairly good agreement on the basic vocabulary of enzyme kinetics and the definition of these parameters in basic pharmacology texts (and associated programs for computing the parameters), as well as the representation of three-dimensional structural data in the Cambridge Crystallographic Database⁷ and the Protein Data Bank (PDB)⁸.

²<http://www.bioml.com/BIOML/>

³<http://genome.ucsc.edu/>

⁴<http://www.ensembl.org/genome/central/>

⁵<http://www.ncbi.nlm.nih.gov/genome/guide/central.html>

⁶<http://.public.celera.com/index.cfm>

⁷<http://www.ccdc.cam.ac.uk/>

⁸<http://www.rcsb.org/pdb/>

Microarray expression data are clearly going to become an important phenotypic data source for pharmacogenomics (57–59), and standards are in active development at this time, although they have not settled yet. There is a proposal for a MicroArray Markup Language (MAML) that would specify a minimal set of information for exchange. MAML is part of a larger effort to develop standards for microarray data and databases in the Microarray Gene Expression Database⁹ effort. The challenges here involve developing standards for representing the experimental conditions, the quality control parameters, the list of genes being assayed, and the actual expression measurements (and background measurements) recorded.

CLINICAL DATA Pharmacogenomics requires a connection to clinical medicine in order to establish the relevance and importance of the systems that are studied. As such, clinical medicine can be considered simply another phenotype, as molecular or cellular data. However, the techniques used to collect and describe clinical data are sufficiently different from basic biological data that the approaches should be distinguished. At the most basic level, pharmacogenomics information resources need to store clinical information pertaining to the most commonly measured phenotypes: one, pharmacokinetic profiles of drug levels in response to dosing and two, measures of pharmacodynamic efficacy based on the target effects. In addition, the incidence of side effects in response to medications must be represented.

Although there is a large literature on the representation of clinical data, much of this is for the purposes of supporting the delivery of clinical care and not for clinical research. Clinical research requires precision in ways different from clinical care, so the portability of most clinical data standards is not clear. The standards that exist for coding diagnoses (International Classification of Diseases standard¹⁰), pathology (Systematized Nomenclature of Medicine¹¹), procedures (Current Procedural Terminology¹²), and others offer a good starting point for pharmacogenomics research but in general do not provide the precision required for high-quality data storage.

As is the case for enzyme kinetics, there is a fair amount of uniformity within the pharmacology community on how to represent pharmacokinetic profiles. Programs such as ADAPT II¹³, NONMEM¹⁴, SAAM 30, and CONSAM¹⁵ exist that all allow first- and second-order kinetics and associated parameters to be computed. A standard set of parameters, including K_i , K_m , and V_{max} , have fairly consistent definitions and thus provide a good initial opportunity for modeling of the data.

One issue that arises in modeling phenotypic data is the relationship between raw data and the more processed parameters and intermediate representations.

⁹<http://www.mged.org/>

¹⁰<http://www.cdc.gov/nchs/about/otheract/icd9/abtcd10.htm>

¹¹<http://www.snomed.org/>

¹²<http://www.ama-assn.org/ama/pub/category/3113.html>

¹³<http://www.usc.edu/dept/biomed/BMSR/Software/adptmenu.html>

¹⁴<http://c255.ucsf.edu/nonmem0.html>

¹⁵<http://www-saam.nci.nih.gov/index.html>

Although the raw data are used to compute these intermediate representations (for example, the raw time points of blood levels are used to compute pharmacokinetic parameters), it can be difficult to determine the appropriate level of data to make routinely available in databases. The raw data can be cumbersome, and the computed parameters may be of real interest to most. However, there are times when the raw data must be retrieved in order to check conclusions or alternative interpretations. Thus, a major challenge for pharmacogenomic information resources is to provide easy access both to the computed/derived parameters as well as to the basic information upon which they are based.

Developing Communication Standards in Pharmacogenomics

One way in which informatics technologies can help accelerate progress in a field is the development of standards for representing and exchanging data. It is clear that shared understanding of the basic data elements within pharmacogenomics is a critical building block with which to build an information infrastructure. Methods for communicating these data are therefore equally important. The two main areas that require progress are the definition of shared syntax (how information is structured in a data file) and semantics (how the information should be interpreted by others). Two contributing technologies within informatics that address these problems are technologies for defining shared vocabularies and technologies for exchanging them.

A standard vocabulary is a controlled set of terms that can be used instead of free text to communicate information. For example, whereas an abstract in Medline is free text, the list of Medical Subject Heading keywords from the MEDLINE database¹⁶ represents a controlled vocabulary. The advantage of a controlled vocabulary is that computer programs can be written to expect certain phrases and can be instructed how to process data based on the occurrence of these phrases. Whereas free text has much more power for expressiveness, it is difficult for computer programs to understand because of the inherent ambiguities in human-to-human communication. Some of these difficulties are addressed by natural language processing techniques (reviewed below). However, a principal means to address these problems is to develop and adopt controlled vocabularies.

Some vocabularies have been developed to facilitate the delivery of clinical care, and these may be helpful in pharmacogenomics. Many contributing vocabularies have been related to one another as part of the Unified Medical Language System project at the National Library of Medicine (60). To support these endeavors fully, however, these vocabularies need to be supplemented by establishing the following standards.

1. Human gene names and links to other organisms. The Human Genome Nomenclature Committee has created a reference set of symbols for human

¹⁶<http://www.ncbi.nlm.nih.gov/PubMed/>

genes, and this should stabilize over time and provide a useful set of indices for the human genome browsers. Included in this activity is the identification of the function of new genes of pharmacogenomic interest, particularly transporters [which are classified in a taxonomy by Saeir¹⁷ (61)] and the cytochrome p450 system (classified based on isoform similarity) (62, 63).

2. Drug and compound names. There are efforts proposed to build a controlled list of drug categories, their structural and biological features, and the associated specific compounds. The Unified Medical Language System (UMLS) contains the 1997 Food and Drug Administration Standard Product Nomenclature.¹⁸
3. Side effects. Standards are required for coding drug side effects, at a clinical level and perhaps at a lower biological level. A vocabulary that is used for clinical trials is a good initial start. The UMLS contains the World Health Organization Adverse Drug Reaction Terminology and the coding symbols for a thesaurus of adverse reaction terms (COSTART) from the Food and Drug Administration.¹⁹

Data Exchange Standards

eXensible Markup Language (XML) has emerged as a common standard for the exchange of data (64)²⁰. XML is a syntax for specifying how text data can be labeled so that computer programs can load the data items into their memory structures. Without a standard such as XML, competing file formats would abound, and programs would work only with a subset of these formats. It then becomes very difficult to exchange data and run programs. Similarly, databases often are constructed independently and organize their data quite differently (including decisions as simple as whether a drug dose should be specified as a single string “20 mg oral” or as separate fields). XML provides a partial solution to this problem by providing a standard syntax for specifying the elements within the file and how they are presented. It becomes relatively easy, then, to read files in one format and then to translate them into a newer format. Even better, however, is for the community to adopt a single format for uniform representation of data. The PharmGKB database²¹ is attempting to define some XML standards for data that will allow basic pharmacogenetic data to be formatted in a standard manner.

A more difficult issue is the semantics of the data representation. Syntax only enforces the order of the elements and their basic types (integers, real numbers, strings, and the like), but specifications of semantics require more constraints on the

¹⁷<http://www-biology.ucsd.edu/~msaier/transport/titlepage.html>

¹⁸<http://www.fda.gov/cder/ndc/database/default.htm> and <http://www.fda.gov/cder/ob/default.htm>

¹⁹<http://www.fda.gov/cder/aers/index.htm>

²⁰<http://www.w3.org/XML/>

²¹<http://pharmgkb.stanford.edu/xml-schemas.html>

logical relationships between data items (for example, a “nonsynonymous SNP” must be at a genome sequence position that is a SNP, must be within a coding region, and must change the amino acid for which it is coded). The specification of semantics is an active area of research, but the Resource Description Framework (written in XML itself²²) is an attempt to enable the standardization of semantics. In addition, knowledge base management systems support the logic and constraint checking that is required for computationally enforcing semantics (65, 66).

Integrating Data From Diverse and Heterogeneous Databases

Pharmacogenomics research is marked by the diversity of databases that must be used in order to answer important questions. For example, in order to find all three-dimensional protein structures with SNPs that change the amino acid in the coding region of proteins that are involved in diabetes, we must combine gene sequence data [such as are found in GENBANK (48)], SNP data [dbSNP (50)], three-dimensional structural databases [Protein Data Bank (67)], databases of genetic diseases and their gene defects [OMIM (52)], and the medical literature. Other queries might require databases of drugs or drug-drug interactions, which are not publicly available at this time.

The problem of integrating data is a difficult one within computer science. One approach is to create a single large data model and to dump the contents of all the contributing databases into the new “mega” database. This approach, called consolidation, suffers because as the contributing databases evolve, the consolidated database becomes out of date. In addition, it is very difficult to build a large data model. Another class of approaches to database integration is federation, which can take three forms. In the first, databases are linked together loosely with hyperlinks on the web, offering little help to automatic programs but useful for human users. In the second, programs are written to extract certain types of data from each database and combine them to answer queries that require data from more than one database (68). In the third, programs are written to extract the data on a regular basis from the contributing databases and dump them into a common database that is then updated regularly [but functions as a consolidated database between updates (69)].

Mining the Published Literature for Pharmacogenomic Data

The Medline/PubMED resource contains references to more than ten million biomedical publications, and many of these offer online abstracts. In addition, many journals are now making their articles available online in full text. Although this mode of publication is effective for human users, it is difficult for computers to extract information from natural language text. At the same time, there are decades of biological knowledge stored in written natural language text. In order

²²<http://www.w3.org/RDF/>

to avoid the loss of this information and to assist in the automatic population of databases, informatics researchers are building systems for extracting information from text. The general problem of understanding the full details of a natural language text has been studied for more than four decades and remains unsolved (70); however, the more tractable goal of reliably identifying relationships within text is within reach. For example, texts can be analyzed to extract protein names (71), and protein-protein (11, 72, 73) or protein-drug interactions (74), based on the occurrence of protein names and verbs such as “inhibits,” “activates,” “represses,” “enhances,” etc.).

Within pharmacogenomics there are good opportunities for natural language processing (NLP) techniques to assist in the organization of data. First, there is no definitive list of drug-gene interactions, and the literature (both published medical literature and the U.S. patent application literature²³) is filled with associations that are pharmacokinetic (e.g., “X is metabolized by CYP2D6”) and pharmacodynamic (e.g., “Y is active at the beta-adrenergic receptor”). In general, NLP techniques work best in well-defined domains that use standardized vocabulary. A second area of opportunity within pharmacogenomics is the extraction of cellular localization information (“X is localized to the Golgi”) from text (75). A third area that holds much promise is the processing of mRNA expression data with microarrays (8, 20, 57, 59, 76–79). Because of the great volume of information generated by these experiments, it is often useful to cluster the genes based on expression pattern, and it is very challenging subsequently to summarize the key features of each cluster. NLP-based techniques may be able to combine the published information about genes with information about how they cluster to create automatic cluster labels that provide biological insight. A fourth area for NLP applications is in the identification of genes of pharmacogenomic interest from text and in the classification of these genes as primarily of pharmacokinetic or pharmacodynamic importance. The language within pharmacokinetic papers is quite idiosyncratic (including discussions of “area under curve” and “bioavailability”), so it may be relatively straightforward to classify abstracts that are discussing this topic and then to extract key data elements from them.

Using Expression Data to Assess the Phenotypes of Drug Response

The emergence of DNA microarrays to measure mRNA expression has created excitement in many areas of biomedical research, including pharmacogenomics (8, 20, 57, 59, 76–79). These microarrays use hybridization of amplified RNA from samples of interest to DNA of known sequence (that have been affixed to small spots that are arranged into a square array) in order to measure the level of gene expression in the samples. The use of microarrays for pharmacogenomics has only begun, but it has the potential of bringing great gains because they can be used to

²³<http://www.uspto.gov/>

address the most difficult steps of both genotype-to-phenotype and phenotype-to-genotype approaches. In genotype-to-phenotype investigations, microarrays can assist in the third step to find phenotypes at the cellular and molecular level that are associated with variations in genotype (and in the context of administering certain drugs). In phenotype-to-genotype studies, microarray measurements can be used in the second step to find genes whose expression alters in the context of an important new phenotype.

The most common informatics analyses of microarray data are currently clustering of genes and classification of genes based on shared expression patterns (80). These groupings can be used as evidence for “guilt-by-association” assignments of function, whereby the function of a gene is assumed to be similar to the function of genes with which it is grouped. The most exciting work in microarrays, however, involves combining information from microarrays with other data sets. One important study measured the expression levels of genes within sixty cancer cell lines and compared the sensitivity of each of these cell lines to over 70,000 different potential anticancer drugs (58). The key expression features that identified potential sensitivity to the anticancer drugs were defined, and cell lines were clustered based on common potential sensitivities. Microarray analysis has also been used to study the pharmacogenomics of cystic fibrosis (81), schizophrenia (82), and others. We expect that in the future microarray analysis of cells before and after drug exposure will provide an important set of pharmacogenomic data for determining the full set of changes that occur at a cellular level, as well as for determining the cells’ kinetics (59).

Understanding the Structural Consequences of Genetic Variations

Pharmacology has always had a strong structural component because the three-dimensional structure of drugs can be critical for understanding mechanisms of action and for building pharmacophores for drug design (83). The increasing number of structures in the 3D structural database also allows us to model the interactions between proteins and their ligands in order to gain a high-resolution understanding of drug action. The PDB now has over 15,000 individual structures (67), and the emergence of high-throughput structure determination efforts promises to maintain a rapid rate of data acquisition (84). The availability of more 3D structures makes it increasingly likely that a homologous protein of known structure will be available for most proteins of pharmacogenomic interest. The types of structural analyses that are becoming important include the following methods

1. Methods for docking small molecules into binding pockets of proteins in order to predict affinity. There have been a number of successful reports in this area, including algorithms based on energetics and based on statistical analysis of pockets (85–87).
2. Methods for homology modeling in order to build models of protein variations. A variety of programs have been developed and tested and offer good

options for building 3D models of proteins that are globular and share 30% or more sequence identity with a known structure (88). In these applications, the location and possible functional significance of nonsynonymous SNPs in the coding portion of proteins can be evaluated (89). A recent paper estimated that 30% of all nonsynonymous SNPs may be associated with significant changes in function (90) based on an analysis of a large set of mutations in DNA binding proteins. There are also some early indications that even synonymous SNPs may change RNA stability and affect the level of activity for some proteins.

3. Methods for predicting protein-protein interactions. It is clear that many proteins have multiple partners with which they interact as activators, inhibitors, or otherwise as modifiers. There has been progress in molecular docking algorithms (91–96) that allows investigators to combine geometric and energetic properties for the purpose of understanding how two protein surfaces may interact.

Comparing Genomes to Develop Pharmacogenomic Models

Comparative genomics is the study of multiple genomes from different organisms in order to define the shared characteristics between organisms as well as the distinguishing characteristics of each organism (41, 42). For pharmacogenomics, comparative genomics can identify analogs to human drug response phenotypes in organisms that are more readily manipulated experimentally. In general, the coding regions of rat, mouse, and pig are more strongly conserved relative to human than the noncoding regions. Recent work shows that noncoding regions can be conserved across species and can be important for regulating gene expression (97). Thus, the availability of complete genomes for related species is likely to assist pharmacogenomics researchers in identifying areas outside coding regions that may be worth studying for polymorphisms. Because the background rate of polymorphisms in humans is so high, it is critical to have these clues from comparative genomics to guide the selection of regions to emphasize in the search for critical variations. These analyses depend on accurate alignments of large segments of genomic DNA, and special purpose algorithms have been developed (97, 98).

Comparative genomics techniques can also be used to understand metabolic and genetic regulatory pathways. Metabolic databases such as EcoCYC (99) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) (100) have been constructed for a number of complete bacterial genomes and are beginning to emerge for eucaryotes as well. These resources will allow a computational analysis of pathways in determining genes that should be studied and perhaps in determining their role in metabolism or drug action (101). Genetic regulatory pathways are summarized in the Cell Signaling Network Database (54) and the signal transduction knowledge environment (STKE²⁴), and can also be studied with databases of transcription factors and regulatory sequences (55). The use of comparative

²⁴<http://stke.sciencemag.org/>

microarray expression analysis is also being evaluated as a strategy for filtering important signals from microarray expression experiments (102, 103).

Managing Laboratory Information Data

Although it is not peculiar to pharmacogenomics, the development of reliable laboratory information management systems (LIMS) is as critical to this field as any. Tracking the large number of samples that must be tracked of patients, tissues, cell lines, and individual genes and gene products is a nontrivial bookkeeping challenge. Excellent systems have been developed in the context of disease research networks for web-based tracking and linking of samples to core data sets (104). Pharmacogenomics offers a particularly challenging application area for LIMS because of the diverse array of data and samples that are relevant and because genomic data are being linked to clinical data for the purposes of finding genotype-phenotype associations. In addition, the emergence of publicly available tissue samples for sampling genomic diversity (105) creates a tracking problem for samples worldwide.

Protecting the Confidentiality and Privacy of Clinical Phenotype Data

The study of pharmacogenomics and the desire to disseminate data pertaining to pharmacogenomics raise a number of important issues in ethics and patient confidentiality. The need to integrate molecular data with clinical data implies that clinical phenotype information may be distributed on the internet. It is crucial, however, that the privacy and security of patient identity be maintained while disseminating these data. Simple methods of “de-identification” (in which basic identifying information such as name, address, and other demographic information is removed) for patient protection is not adequate. As more information is provided about a patient, even though it is not directly identifying, it can often be combined with other data sources (such as hospital discharge records and voter and driver registration information) to reconstruct, either exactly or probabilistically, the identity of patients (106). There are precedents for the publication of patient-related data in journals as well as in databases (such as Genbank). Nonetheless, it is critical to ensure that the availability of clinical phenotype data sets does not lead to the loss of study-subject confidentiality or privacy.

There are generally two approaches to protecting patient privacy. The first, called mediation, inserts a computer program in between a user and a database and monitors the queries that are asked and the answers that are provided by the database. The mediator has rules about the kinds of queries that can be written and the kinds of responses that can be supplied from the database. It stops queries that are inappropriate (based on the rules, which often include information about the privileges of the user), and it filters results that are inappropriate. These mediators have been shown to provide reasonable semi-automated protections (107). A second method, called scrubbing, is based on the principle of removing information from a data set so that the details of the data cannot be used

to re-identify a patient (108, 109). Thus, for example, a list of pharmacokinetic parameters can be either rounded off to reduce precision or can be provided as ranks instead of absolute values. Each of these maneuvers would serve to increase the pool of subjects from whom these values could come and thus protect the privacy of individual subjects. Crucial to scrubbing is the idea of “bin size,” which is the minimum number of people (in some defined population) who match a certain query. Thus, we may say that in a study with 500 patients, we will never answer a query with results containing less than 10 subjects, so that individual subject data cannot be teased out. The bin size has been used by other organizations, such as the social security administration in the United States, and the U.S. census bureau (106). The obvious disadvantage to scrubbing is the loss of precision, which can make certain statistical analyses much more expensive. Although not within the scope of this review, there are associated implications about how study subjects should exercise informed consent when participating in pharmacogenomic studies. The other critical issue that is outside the technical scope of this paper, but deserves mention, is the problem of having pharmacogenomic information used to discriminate (either in terms of the research agenda, insurance, or employment) against groups within the population based on statistical associations.

PHARMACOGENOMICS: A NEW CHALLENGE FOR BIOMEDICAL INFORMATICS

The focus of biomedical informatics has, in the past, been fragmented into informatics to meet the challenges of genomics (sequence analysis, structure analysis, biochemical and regulatory pathway analysis) and informatics to meet the challenges of organizing clinical data (medical records, information extraction, database integration). In the post-genome period, there will be an increasing number of applications that require the combination of basic bioinformatics with clinical informatics. Pharmacogenomics is an excellent example of such an application area. Many different branches of biomedical informatics clearly will play a critical role in gathering, organizing, and analyzing pharmacogenomic data. The National Institutes of Health has recently formed a Pharmacogenetics Research Network and Database²⁵ program whereby several research groups are cooperating to gather pharmacogenomic data (genomic, molecular, cellular and clinical) and deposit it in a common database for public use. The database, PharmGKB²⁶, is intended to gather data not just from these groups, but from all groups worldwide wishing to disseminate their data of pharmacogenomic relevance. The initial focus is on building representations and XML standards for the submission of basic genomic variation data, enzyme kinetic data, and clinical pharmacokinetic data, but work is being done in all the areas reviewed here. As the basic infrastructure is created,

²⁵<http://www.nigms.nih.gov/pharmacogenetics/>

²⁶<http://www.pharmgkb.org/>

there will be opportunities for the community to create and distribute informatics methodologies that address each of the challenges outlined here. There is no doubt that other, perhaps unexpected, informatics challenges will arise in the course of creating these resources.

Visit the Annual Reviews home page at www.AnnualReviews.org

LITERATURE CITED

1. Tarter ME. 1979. Biocomputational methodology: an adjunct to theory and applications. *Biometrics* 35:9–24
2. Besemer J, Lomsadze A, Borodovsky M. 2001. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res.* 29:2607–18
3. Burge CB, Karlin S. 1998. Finding the genes in genomic DNA. *Curr. Opin. Struct. Biol.* 8:346–54
4. Lewis S, Ashburner M, Reese MG. 2000. Annotating eukaryote genomes. *Curr. Opin. Struct. Biol.* 10:349–54
5. Mural RJ. 1999. Current status of computational gene finding: a perspective. *Methods Enzymol.* 303:77–83
6. Landgraf R, Xenarios I, Eisenberg D. 2001. Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *J. Mol. Biol.* 307:1487–502
7. Al-Lazikani B, Jung J, Xiang Z, Honig B. 2001. Protein structure prediction. *Curr. Opin. Chem. Biol.* 5:51–56
8. Altman RB, Raychaudhuri S. 2001. Whole-genome expression analysis: challenges beyond clustering. *Curr. Opin. Struct. Biol.* 11:340–47
9. Vohradsky J. 2001. Neural model of genetic network. *J. Biol. Chem.* 6:6
10. Reis BY, Butte AS, Kohane IS. 2001. Extracting knowledge from dynamics in gene expression. *J. Biomed. Inform.* 34:15–27
11. Jenssen TK, Laegreid A, Komorowski J, Hovig E. 2001. A literature network of human genes for high-throughput analysis of gene expression. *Nat. Genet.* 28:21–28
12. Legrain P, Wojcik J, Gauthier J. 2001. Protein-protein interaction maps: a lead towards cellular functions. *Trends Genet.* 17:346–52
13. Hasty J, McMillen D, Isaacs F, Collins JJ. 2001. Computational studies of gene regulatory networks: in numero molecular biology. *Nat. Rev. Genet.* 2:268–79
14. Salazar-Ciudad I, Newman SA, Sole RV. 2001. Phenotypic and dynamical transitions in model genetic networks. I. Emergence of patterns and genotype-phenotype relationships. *Evol. Dev.* 3:84–94
15. Terwilliger JD, Goring HH, Besemer J, Lomsadze A, Borodovsky M. 2000. Gene mapping in the 20th and 21st centuries: statistical methods, data analysis, and experimental design. *Hum. Biol.* 72:63–132
16. Jansen RC, Nap J. 2001. Genetical genomics: the added value from segregation. *Trends Genet.* 17:388–91
17. Chen RO, Altman RB. 1999. Automated diagnosis of data-model conflicts using metadata. *J. Am. Med. Inform. Assoc.* 6:374–92
18. Broder S, Venter JC. 2000. Sequencing the entire genomes of free-living organisms: the foundation of pharmacology in the new millennium. *Annu. Rev. Pharmacol. Toxicol.* 40:97–132
19. Bartlett J. 2001. Technology evaluation: SAGE, Genzyme molecular oncology. *Curr. Opin. Mol. Ther.* 3:85–96
20. Hu Y. 2001. An integrated approach

- for genome-wide gene expression analysis. *Comput. Methods Programs Biomed.* 65:163–74
21. Tucker CL, Gera JF, Uetz P. 2001. Towards an understanding of complex protein networks. *Trends Cell Biol.* 11:102–6
 22. Cubitt AB, Heim R, Adams SR, Boyd AE, Gross LA, Tsien RY. 1995. Understanding, improving and using green fluorescent proteins. *Trends BioChem. Sci.* 20:448–55
 23. Papac DI, Shahrokh Z. 2001. Mass spectrometry innovations in drug discovery and development. *Pharm. Res.* 18:131–45
 24. Teichmann SA, Murzin AG, Chothia C. 2001. Determination of protein function, evolution and interactions by structural genomics. *Curr. Opin. Struct. Biol.* 11:354–63
 25. Weber W. 1997. *Pharmacogenetics*. Oxford, UK: Oxford Univ. Press
 26. Evans WE, Relling M. 1999. Pharmacogenomics: translating functional genomics into rational therapeutics. *Science* 286:487–91
 27. Rusnak JM, Kisabeth RM, Herbert DP, McNeil DM. 2001. Pharmacogenomics: a clinician's primer on emerging technologies for improved patient care. *Mayo Clin. Proc.* 76:299–309
 28. Meyer UA. 1991. Genotype or phenotype: the definition of a pharmacogenetic polymorphism. *Pharmacogenetics* 1:66–67
 29. McLeod HL, Evans WE. 2001. Pharmacogenomics: unlocking the human genome for better drug therapy. *Annu. Rev. Pharmacol. Toxicol.* 41:101–21
 30. Murphy MP. 2000. Current pharmacogenomic approaches to clinical drug development. *Pharmacogenomics* 1:115–23
 31. Murphy MP, Beaman ME, Clark LS, Cayouette M, Benson L, et al. 2000. Prospective CYP2D6 genotyping as an exclusion criterion for enrollment of a phase III clinical trial. *Pharmacogenetics* 10:583–90
 32. Leushner J, Chiu NH. 2000. Automated mass spectrometry: a revolutionary technology for clinical diagnostics. *Mol. Diagn.* 5:341–48
 33. Hess P, Cooper D. 1999. Impact of pharmacogenomics on the clinical laboratory. *Mol. Diagn.* 4:289–98
 34. Meisel C, Roots I, Cascorbi I, Brinkmann U, Brockmoller J, et al. 2000. How to manage individualized drug therapy: application of pharmacogenetic knowledge of drug metabolism and transport. *Clin. Chem. Lab. Med.* 38:869–76
 35. Yan L, Otterness DM, Weinshilboum RM. 1999. Human nicotinamide N-methyltransferase pharmacogenetics: gene sequence analysis and promoter characterization. *Pharmacogenetics* 9:307–16
 36. Glatt CE, DeYoung JA, Delgado S, Service SK, Giacomini KM, et al. 2001. Screening a large reference sample to identify very low frequency sequence variants: comparisons between two genes. *Nat. Genet.* 27:435–38
 37. Kuehl P, Zhang J, Lin Y, Lamba J, Assem M, et al. 2001. Sequence diversity in CYP3A promoters and characterization of the genetic basis of polymorphic CYP3A5 expression. *Nat. Genet.* 27:383–91
 38. Israel E, Drazen JM, Liggett SB, Boushey HA, Cherniack RM, et al. 2001. Effect of polymorphism of the beta(2)-adrenergic receptor on response to regular use of albuterol in asthma. *Int. Arch. Allergy Immunol.* 124:183–86
 39. Ewesuedo RB, Iyer L, Das S, Koenig A, Mani S, et al. 2001. Phase I clinical and pharmacogenetic study of weekly TAS-103 in patients with advanced cancer. *J. Clin. Oncol.* 19:2084–90
 40. Furuya H, Fernandez-Salguero P, Gregory W, Taber H, Steward A, et al. 1995. Genetic polymorphism of CYP2C9 and its effect on warfarin maintenance dose requirement in patients undergoing anticoagulation therapy. *Pharmacogenetics* 5:389–92
 41. O'Brien SJ, Menotti-Raymond M, Murphy WJ, Nash WG, Wienberg J, et al.

1999. The promise of comparative genomics in mammals. *Science* 286:458–62, 479–81
42. Clark MS. 1999. Comparative genomics: the key to understanding the Human Genome Project. *Bioessays* 21:121–30
43. Bray MS, Boerwinkle E, Doris PA. 2001. High-throughput multiplex SNP genotyping with MALDI-TOF mass spectrometry: practice, problems and promise. *Hum. Mutat.* 17:296–304
44. Brookes AJ. 1999. The essence of SNPs. *Gene* 234:177–86
45. Schork NJ, Fallin D, Lanchbury JS. 2000. Single nucleotide polymorphisms and the future of genetic epidemiology. *Clin. Genet.* 58:250–64
46. Judson R, Stephens JC. 2001. Notes from the SNP vs. haplotype front. *Pharmacogenomics* 2:7–10
47. Mann M, Hendrickson RC, Pandey A. 2001. Analysis of proteins and proteomes by mass spectrometry. *Annu. Rev. BioChem.* 70:437–73
48. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA, Wheeler DL. 2000. GenBank. *Nucleic Acids Res.* 28:15–8
49. Cuticchia AJ. 2000. Future vision of the GDB human genome database. *Hum. Mutat.* 15:62–67
50. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, et al. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 29:308–11
51. Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, et al. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409:928–33
52. Hamosh A, Scott AF, Amberger J, Valle D, McKusick VA. 2000. Online Mendelian Inheritance in Man (OMIM). *Hum. Mutat.* 15:57–61
53. Boyadjiev SA, Jabs EW. 2000. Online Mendelian Inheritance in Man (OMIM) as a knowledgebase for human developmental disorders. *Clin. Genet.* 57:253–66
54. Takai-Igarashi T, Nadaoka Y, Kaminuma T. 1998. A database for cell signaling networks. *J. Comput. Biol.* 5:747–54
55. Wingender E, Chen X, Hehl R, Karas H, Liebich I, et al. 2000. TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res.* 28:316–19
56. Smith CM, Shindyalov IN, Veretnik S, Gribskov M, Taylor SS, et al. 1997. The protein kinase resource. *Trends BioChem. Sci.* 22:444–46
57. Jain KK. 2000. Applications of biochip and microarray systems in pharmacogenomics. *Pharmacogenomics* 1:289–307
58. Scherf U, Ross DT, Waltham M, Smith LH, Lee JK, et al. 2000. A gene expression database for the molecular pharmacology of cancer. *Nat. Genet.* 24:236–44
59. Slonim DK. 2001. Transcriptional profiling in cancer: the path to clinical pharmacogenomics. *Pharmacogenomics* 2:123–36
60. Lindberg DA, Humphreys BL, McCray AT. 1993. The Unified Medical Language System. *Methods Inf. Med.* 32:281–91
61. Saier MH, Jr. 2000. A functional-phylogenetic classification system for transmembrane solute transporters. *MicroBiol. Mol. Biol. Rev.* 64:354–411
62. Nelson DR, Kamataki T, Waxman DJ, Guengerich FP, Estabrook RW, et al. 1993. The P450 superfamily: update on new sequences, gene mapping, accession. *DNA Cell Biol.* 12:1–51
63. Nelson DR, Koymans L, Kamataki T, Stegeman JJ, Feyereisen R, et al. 1996. P450 superfamily: update on new sequences, gene mapping, accession numbers. *Pharmacogenetics* 6:1–42
64. Harold E, Means W. 2001. *XML in a Nutshell: A Desktop Quick Reference*. Cambridge, MA: O'Reilly
65. Abernethy N, Wu J, Hewett M, Altman R. 1999. SOPHIA: a flexible, web-based knowledge server. *IEEE Intell. Sys. Appl.* 14:79–85

66. Musen MA. 1998. Domain ontologies in software engineering: use of Protege with the EON architecture. *Methods Inf. Med.* 37:540–50
67. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. 2000. The Protein Data Bank. *Nucleic Acids Res.* 28:235–42
68. Chung SY, Wong L, Blaschke C, Andrade MA, Ouzounis C, et al. 1999. Kleisli: a new tool for data integration in biology. *Trends Biotechnol.* 17:351–55
69. McEntyre J. 1998. Linking up with Entrez. *Trends Genet.* 14:39–40
70. Allen J. 1995. *Natural Language Understanding*. Redwood City, CA: Benjamin/Cummings
71. Yoshida M, Fukuda K, Takagi T. 2000. PNAD-CSS: a workbench for constructing a protein name abbreviation dictionary. *Bioinformatics* 16:169–75
72. Jenssen TK, Vinterbo S. 2000. A set-covering approach to specific search for literature about human genes. *Proc. AMIA Symp.* pp. 384–88. Philadelphia, PA: Hanley & Belfus
73. Marcotte EM, Xenarios I, Eisenberg D. 2001. Mining literature for protein-protein interactions. *Bioinformatics* 17: 359–63
74. Rindflesch TC, Tanabe L, Weinstein JN, Hunter L. 2000. EDGAR: extraction of drugs, genes and relations from the biomedical literature. *Pac. Symp. Biocomput.* pp. 517–28
75. Drawid A, Jansen R, Gerstein M. 2000. Genome-wide analysis relating expression level with protein subcellular localization. *Trends Genet.* 16:426–30
76. Marcotte EM. 2000. Computational genetics: finding protein function by nonhomology methods. *Curr. Opin. Struct. Biol.* 10:359–65
77. Masys DR, Welsh JB, Lynn Fink J, Gribkov M, Klacansky I, Corbeil J. 2001. Use of keyword hierarchies to interpret gene expression patterns. *Bioinformatics* 17:319–26
78. Kurella M, Hsiao LL, Yoshida T, Randall JD, Chow G, et al. 2001. DNA microarray analysis of complex biologic processes. *J. Am. Soc. Nephrol.* 12:1072–78
79. Ideker T, Thorsson V, Ranish JA, Christman R, Buhler J, et al. 2001. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* 292:929–34
80. Raychaudhuri S, Sutphin PD, Chang JT, Altman RB. 2001. Basic microarray analysis: grouping and feature reduction. *Trends Biotechnol.* 19:189–93
81. Srivastava M, Eidelman O, Pollard HB. 1999. Pharmacogenomics of the cystic fibrosis transmembrane conductance regulator (CFTR) and the cystic fibrosis drug CPX using genome microarray analysis. *Mol. Med.* 5:753–67
82. Kawanishi Y, Tachikawa H, Suzuki T. 2000. Pharmacogenomics and schizophrenia. *Eur. J. Pharmacol.* 410:227–41
83. Ekins S, de Groot MJ, Jones JP. 2001. Pharmacophore and three-dimensional quantitative structure activity relationship methods for modeling cytochrome p450 active sites. *Drug Metab. Dispos.* 29:936–44
84. Blundell TL, Mizuguchi K. 2000. Structural genomics: an overview. *Prog. Biophys. Mol. Biol.* 73:289–95
85. Ewing TJ, Makino S, Skillman AG, Kuntz ID. 2001. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *J. Comput. Aided Mol. Des.* 15:411–28
86. Pang YP, Kozikowski AP. 1994. Prediction of the binding site of 1-benzyl-4-[(5,6-dimethoxy-1-indanon-2-yl)methyl]piperidine in acetylcholinesterase by docking studies with the SYSDOC program. *J. Comput. Aided Mol. Des.* 8: 683–93
87. Sun Y, Ewing TJ, Skillman AG, Kuntz ID. 1998. CombiDOCK: structure-based combinatorial docking and library design. *J. Comput. Aided Mol. Des.* 12:597–604
88. Sanchez R, Sali A. 2000. Comparative

- protein structure modeling. Introduction and practical examples with modeller. *Methods Mol. Biol.* 143:97–129
89. Sunyaev S, Lathe W, Bork P III. 2001. Integration of genome data and protein structures: prediction of protein folds, protein interactions and “molecular phenotypes” of single nucleotide polymorphisms. *Curr. Opin. Struct. Biol.* 11:125–30
90. Chasman D, Adams RM. 2001. Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *J. Mol. Biol.* 307: 683–706
91. Claussen H, Buning C, Rarey M, Lengauer T. 2001. FlexE: efficient molecular docking considering protein structure variations. *J. Mol. Biol.* 308:377–95
92. Goldman BB, Wipke WT. 2000. QSD quadratic shape descriptors. 2. Molecular docking using quadratic shape descriptors (QSDock). *Proteins* 38:79–94
93. Moont G, Gabb HA, Sternberg MJ. 1999. Use of pair potentials across protein interfaces in screening predicted docked complexes. *Proteins* 35:364–73
94. Morris GM, Goodsell DS, Huey R, Olson AJ. 1996. Distributed automated docking of flexible ligands to proteins: parallel applications of AutoDock 2.4. *J. Comput. Aided Mol. Des.* 10:293–304
95. Ritchie DW, Kemp GJ. 2000. Protein docking using spherical polar Fourier correlations. *Proteins* 39:178–94
96. Sternberg MJ, Aloy P, Gabb HA, Jackson RM, Moont G, et al. 1998. A computational system for modelling flexible protein-protein and protein-DNA docking. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 6:183–92
97. Dubchak I, Brudno M, Loots GG, Pachter L, Mayor C, et al. 2000. Active conservation of noncoding sequences revealed by three-way species comparisons. *Genome Res.* 10:1304–6
98. Delcher AL, Kasif S, Fleischmann RD, Peterson J, White O, Salzberg SL. 1999. Alignment of whole genomes. *Nucleic Acids Res.* 27:2369–76
99. Karp PD, Riley M, Saier M, Paulsen IT, Paley SM, Pellegrini-Toole A. 2000. The EcoCyc and MetaCyc databases. *Nucleic Acids Res.* 28:56–59
100. Wixon J, Kell D. 2000. The Kyoto encyclopedia of genes and genomes—KEGG. *Yeast* 17:48–55
101. Werner T. 2001. Cluster analysis and promoter modelling as bioinformatics tools for the identification of target genes from expression array data. *Pharmacogenomics* 2:25–36
102. Zien A, Küffner R, Zimmer R, Lengauer T. 2000. Analysis of gene expression data with pathway scores. *ISMB* 2000:407–17
103. Nolan PM. 2000. Generation of mouse mutants as a tool for functional genomics. *Pharmacogenomics* 1:243–55
104. Nadkarni PM, Marengo L, Chen R, Skoufos E, Shepherd G, Miller P. 1999. Organization of heterogeneous scientific data using the EAV/CR representation. *J. Am. Med. Inform. Assoc.* 6:478–93
105. Collins FS, Brooks LD, Chakravarti A. 1998. A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res.* 8:1229–31
106. Sweeney L. 1998. Privacy and medical-records research. *N. Engl. J. Med.* 338: 1077; discussion-8
107. Wiederhold G, Bilello M, Sarathy V, Qian X. 1996. A security mediator for health care information. *Proc. AMIA Annu. Fall Symp.* 120–24
108. Sweeney L, Nolan PM. 1997. Guaranteeing anonymity when sharing medical data, the Datafly System. *Proc. AMIA Annu. Fall Symp.* 1:51–55. Philadelphia, PA: Hanley & Belfus
109. Malin B, Sweeney L. 2000. Determining the identifiability of DNA database entries. *Proc. AMIA Symp.* pp. 537–41. Philadelphia, PA: Hanley & Belfus