PERSPECTIVE

Integrating genotype and phenotype information: an overview of the PharmGKB project

TE Klein, JT Chang, MK Cho, KL Easton, R Fergerson, M Hewett, Z Lin, Y Liu, S Liu, DE Oliver, DL Rubin, F Shafa, JM Stuart and RB Altman

Stanford Medical Informatics, Stanford, CA, USA

Pharmacogenetics seeks to explain how people respond in different ways to the same drug treatment. A classic example of the importance of pharmacogenomics is the variation in individual responses to the anti-leukemia drug, 6-mercaptopurine. Most people metabolize the drug quickly. Some individuals, with a genetic variation for the enzyme thiopurine methyltransferase (TPMT),[1] do not. Consequently, they need lower doses of 6-mercaptopurine for effective treatment as normal doses can be lethal. One of the many promises of the human genome project is an ability to pharmacologically treat individuals in a more personalized rather than statistical manner.

Two independent sequence drafts of the human genome were recently completed and reported simultaneously in the literature.[2,3] With these drafts in hand, and further refinement of these sequence maps expected in the future, scientists are presented with many informatics opportunities and challenges including defining methods to analyze and relate these data to observations from genetics, biology, chemistry, and clinical medicine. A principal challenge in the analysis of these data is the difficulty in linking information about the variation in human genes to the variation in drug response (pharmacogenetics) and to understand how interacting systems of genes determine individual drug responses (pharmacogenomics).

Pharmacogenetics and pharmacogenomics are interdisciplinary and collaborative fields requiring the cooperative efforts of research and clinical scientists (ie, geneticists and physicians). Ideally, the data from all research would be made public so that hypotheses could be proposed and tested using computational techniques. Sample inquiries might include the following: (1) for gene X, show all observed polymorphisms in its sequence; (2) for drug Y, show the variability in pharmacokinetics; and (3) for phenotype Z, show the variability in association with drug Y and/or gene X. Such queries require a database that can model key elements of the data, acquire data efficiently, provide query tools for analysis and deliver the resulting system to the scientific community.

The National Institutes of Health, recognizing: (1) the need for fully-disclosable and publishable research in this domain; and (2) the need to store the results of this research in a public database, has funded the Pharmacogenetics Research Network and Knowledge Base (PharmGKB: http://www.pharmgkb.org).* PharmGKB will become a national resource containing high quality structured data linking genomic information, molecular and cellular phenotype information, and

clinical phenotype information. The ultimate product of this project will be a knowledge base that will provide a public infrastructure for understanding how variations in the human genome lead to variations in clinical response to medications.

PharmGKB users may include scientists from the public and private sectors seeking access to polymorphism data or phenotype data about variants, clinical scientists with focused questions about how a drug may be metabolized and the general public with an interest in pharmaceutical science or specific clinical phenotypes. The database will also provide publicly accessible data for young scientists wishing to familiarize themselves with the detailed ways in which the data are collected and analyzed.

The data model for PharmGKB will describe both pharmacokinetic and pharmacodynamic data. Pharmacokinetics includes absorption, distribution, metabolism and elimination of a drug. Pharmacodynamics includes pharmacological effects and clinical response leading to toxicity and efficacy for a particular drug or metabolite. Figure 1 illustrates the complexity of relationships that are of interest for this knowledge base. The success of PharmGKB will depend on its ability to accurately represent the classes of information in each domain of scientific interest, the attributes for each of the concepts and the relationships among these concepts. PharmGKB may use or create structured vocabularies to describe the data in precise ways. These structured vocabularies will provide the foundation for intelligent reasoning and inference (including semi-automated validation and verification of data, and the generation of exploratory scientific hypotheses) from the knowledge base.

The PharmGKB requires a detailed model of genomic sequence, in order to represent accurately DNA sequence data, gene structure and polymorphisms in sequence—much more than

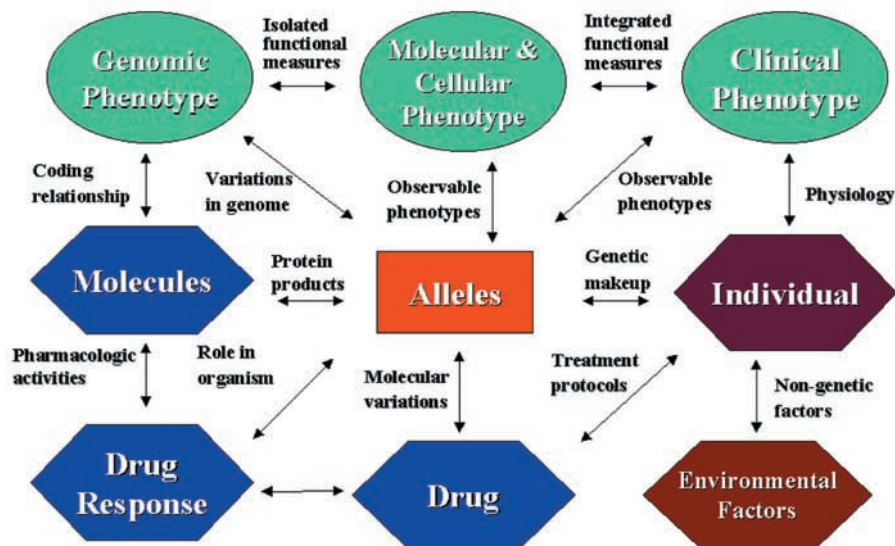## Complexity of Relationships in Pharmacogenetics

Figure 1 Data elements to be modeled in PharmGKB (©Stanford University, 2000).

simply storing the DNA sequence.† Important distinctions must be maintained in order to track haploid, diploid or polyploid alleles, alternative splice sites, and polymorphisms observed as common variants. There are also methods for representing sample populations, such as the Coriell Institute of Medical Research sample sets (see http://arginine.umdnj.edu/). Molecular and cellular phenotype data include enzyme kinetic measurements, such as binding, catalysis and inhibition constants for particular drugs, cellular drug processing rates, homology modeling of three-dimensional structures and pharmacodynamic assays. Again, carefully constructed data models are critical in order to capture details that users will want to use for search, report generation, data analysis and inference. Clinical phenotype is perhaps the most difficult data to model and link with genomic and molecular/cellular phenotypic data. Clinical phenotype data include basic pharmacokinetic measurements (such as drug absorption, distribution, elimination and

metabolism) as well as pharmacodynamic profiles, which currently include pulmonary, cardiac and psychological function tests, and cancer chemotherapeutic side effects. Modeling clinical information can be difficult because controlled vocabularies are not routinely employed and there are many levels of detail required to describe the relationships between clinical signs and symptoms, diseases and physiology.

Due to the sensitive nature of clinical data, particular attention must be paid to protecting the privacy and confidentiality rights of individuals who have consented to participate in pharmacogenomic research. Informed consent is the standard modality for educating research study participants, but the nature of this consent changes when medical data may exist on the world wide web.[4] The PharmGKB will respect the absolute confidentiality of genetic information of individuals. Towards that goal, data flow is limited both *into* and *out of* the knowledge base, based on evolving rules defining what can be stored in the PharmGKB and what can be disseminated. No identifying information about an individual patient will be accepted into the knowledge base, and methods will be

employed to ensure that patient identity cannot be reconstructed from publicly available data records. The PharmGKB is subject to Stanford University's institutional review and oversight, and also adheres to the principles outlined in the Health Insurance Portability and Accountability Act of 1996 (Public Law 104–191, or HIPAA). Specific policies can be found at http://www.pharmgkb.org/policies.html.

The PharmGKB software architecture consists of three layers. The top layer is a collection of knowledge manipulation and retrieval tools that operate on the knowledge ontology (data model) and instances of relevant data. Tools in development currently include single-use rules to extract information, periodic rules to examine new data, automated and human curated verification and validation methods, and intelligent reasoning for automatic deduction and reasoning by analogy.

The middle layer is a frame-based knowledge base that supports a hierarchical object-oriented organization of data.‡ PharmGKB will provide a net-

---

†For further details, the reader is referred to the XML schema found at http://pharmgkb.org/xml-schemas.html

---

‡We are currently using a derivative of the Protégé 2000, please see http://protege.stanford. edu/

PharmGKB dbSNP Surveillance Report

HGNC Gene Name: thiopurine S-methltransferase
HGNC Gene Symbol: TPMT
Locus ID: 7172
RefSeq Accessions: NM_000367, NP_000358
Genbank Accessions: U30518, S62904, NT_007180, XM_012752, AF019369, U81562, U81563, U81564, U81566, U81567, BC005339, U12387

All SNPs Reported

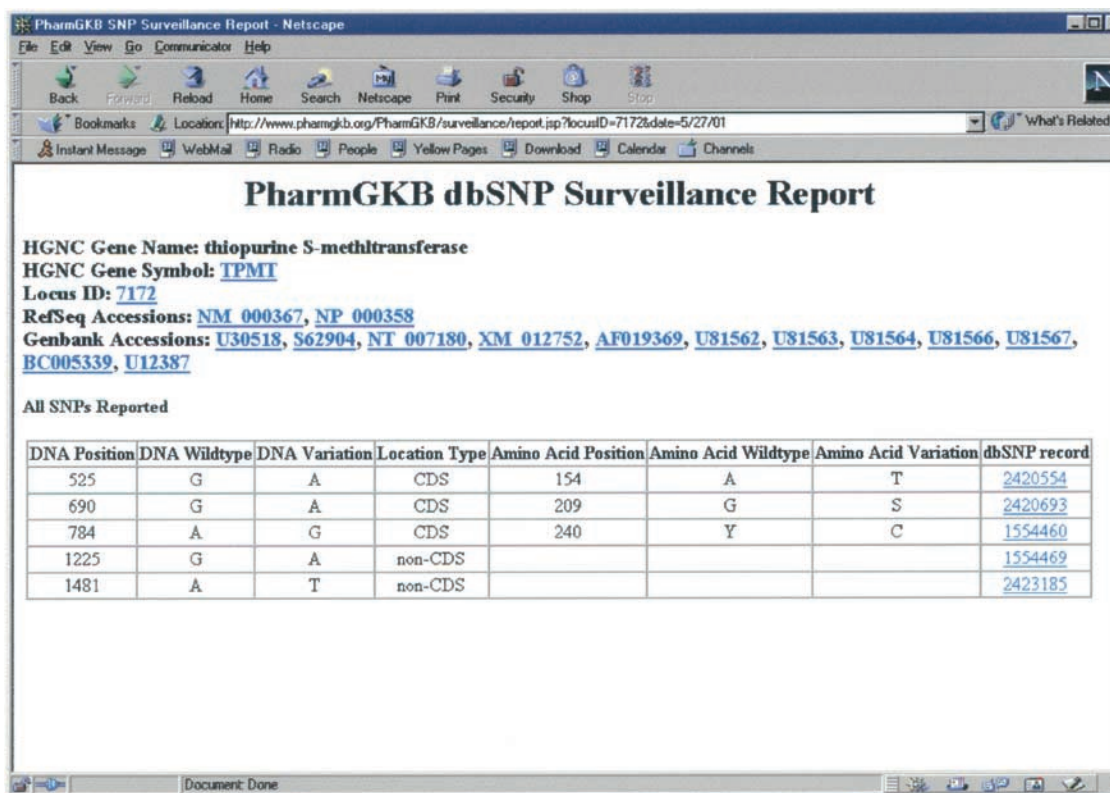| DNA Position	DNA Wildtype	DNA Variation	Location Type	Amino Acid Position	Amino Acid Wildtype	Amino Acid Variation	dbSNP record
525	G	A	CDS	154	A	T	2420554
690	G	A	CDS	209	G	S	2420693
784	A	G	CDS	240	Y	C	1554460
1225	G	A	non-CDS				1554469
1481	A	T	non-CDS				2423185

Figure 2 PharmGKB dbSNP surveillance report for TPMT.

work-accessible application programmer's interface (API) in Java that is consistent with the Open Knowledge Base Connectivity (OKBC) standard for interacting with a knowledge base.

The third layer provides a relational database for physical storage of the knowledge. The relational database provides support for transactions, secure access and mechanisms for maintaining data integrity.

Data can be deposited by registered PharmGKB submittors into the knowledge base via either web-based forms or direct XML submissions. Primary and ancillary supporting data can also be accepted to provide access to 'raw' scientific data, suitable for re-interpretation. As part of PharmGKB services, PharmGKB tools will also formulate data submissions to relevant external databases such as the Single Nucleotide Polymorphism database (dbSNP).§

§For further information about dbSNP, please see http://www.ncbi.nlm.nih.gov/SNP/

Thus, the PharmGKB developers must track the changing formats of these external databases and provide general purpose mechanisms for efficient and maintainable data transfer. Finally, PharmGKB is a public database, and the data will be available for export in text format, such as XML.

The PharmGKB contains valuable data that must be cross-referenced to a variety of other web-accessible databases. Mechanisms for surveillance of and integration with external databases are therefore critical. The ability to combine PharmGKB information with other publicly available information provides a much more powerful query capability than multiple queries to the individual databases. Towards that end, the PharmGKB is creating methods for database integration that allow multiple information sources to be combined in formulating response to a user query.

The first example of database integration for the PharmGKB is its

relationship to the NCBI resource, dbSNP. PharmGKB code monitors dbSNP, looking for new information about the genes of interest to the various associated research groups (see http://pharmgkb.org/PharmGKB/query/ The surveillance tool retrieves all SNPs in the transcripts of all the sequence entries for each locus. In the second version of the dbSNP surveillance tool, all SNPs in the genomic region will be reported. Figure 2 shows an example of the PharmGKB dbSNP Surveillance Report for TPMT.

Another tool provided to the users is a comparative genomics analysis between human and mouse. Specifically, long-range regulatory elements can be difficult to find experimentally, but are often conserved in syntenic regions between mice and humans. Identification of such segments may help focus polymorphism studies on non-coding areas that are more likely to be associated with detectable phenotypes. We use VISTA (Visualization Tool for Alignment),[5,6]

an integrated system for global alignment and visualization designed for comparative genomic analysis. VISTA identifies conserved regions in orthologous genomic sequences from two or more species.

The Pharmacogenetics Knowledge Base (PharmGKB) is in its infancy. In parallel with the basic infrastructure development, research projects are beginning that address scrubbing clinical data for patient de-identification, using natural language processing to extract pharmacogenomic information from the published literature, microarray analysis, and interactive three-dimensional modeling of structural variants. PharmGKB will assist the scientific and clinical communities in the exploration of pharmacogenetics and pharmacogenomics domains.

DUALITY OF INTEREST

None declared

Correspondence should be sent to

TE Klein, Stanford Medical Informatics, 251 Campus Drive, MSOB X-215, Stanford, CA 94305-5479, USA.
E-mail: klein@SMI.Stanford.edu

1 Lee D, Weinshilboum RM. *Drug Metab Dispos* 1995; **23**: 398–405.
2 Venter JC *et al. Science* 2001; **291**: 1304–1351.
3 Lander ES *et al. Nature* 2001; **409**: 860–921.
4 Rothstein MA, Epps PE. *Nature Rev Genet* 2001; **2**: 228–231.
5 Dubchak I, Frazer KA. *Genome Res* 2000; **10**: 1304.
6 Mayor C, Dubchak I. *Bioinformatics* 2000; **16**: 1046–1047.