Pacific Symposium on Biocomputing 7:65-76 (2002)

ONTOLOGY DEVELOPMENT FOR A PHARMACOGENETICS KNOWLEDGE BASE

DIANE E. OLIVER, DANIEL L. RUBIN, JOSHUA M. STUART, MICHEAL HEWETT, TERI E. KLEIN, AND RUSS B. ALTMAN

Stanford Medical Informatics, Stanford University School of Medicine, 251 Campus Drive, MSOB X-215, Stanford, CA 94305-5479

oliver@smi.stanford.edu, rubin@smi.stanford.edu, stuart@smi.stanford.edu, hewett@smi.stanford.edu, klein@smi.stanford.edu, altman@smi.stanford.edu

Research directed toward discovering how genetic factors influence a patient's response to drugs requires coordination of data produced from laboratory experiments, computational methods, and clinical studies. A public repository of pharmacogenetic data should accelerate progress in the field of pharmacogenetics by organizing and disseminating public datasets. We are developing a pharmacogenetics knowledge base (PharmGKB) to support the storage and retrieval of both experimental data and conceptual knowledge. PharmGKB is an Internet-based resource that integrates complex biological, pharmacological, and clinical data in such a way that researchers can submit their data and users can retrieve information to investigate genotype–phenotype correlations. Successful management of the names, meaning, and organization of concepts used within the system is crucial. We have selected a frame-based knowledge-representation system for development of an ontology of concepts and relationships that represent the domain and that permit storage of experimental data. Preliminary experience shows that the ontology we have developed for gene-sequence data allows us to accept, store, and query data submissions.

1 Introduction

In the quest to understand the impact of genetic factors on drug response, researchers must integrate data produced from laboratory experiments, computational methods, and clinical studies. Different individuals have different responses to the same medications. With the draft sequence of the human genome and increased understanding of metabolic enzymes, drug transporters, and drug receptors, there is great promise for improving our understanding of how genetic variation affects variation in drug efficacy and toxicity. The sheer volume of biological data, the uncertainties associated with those data, and the complexity of the relationships among concepts present challenges for structuring and managing pharmacogenetic data and knowledge.

We are participating in the Pharmacogenetic Research Network and Knowledge Base consortium, a group of investigators funded by the National Institutes of Health (NIH) to study pharmacogenetics. The goal of NIH in funding this consortium is two-fold: (1) to create a network of multidisciplinary, collaborative research groups that study pharmacologically significant genetic variation, and (2) to build a knowledge base that is available to the research community and that can

stimulate hypothesis-driven research[1]. Investigators in the consortium conduct studies to identify genetic polymorphisms, to assess functional variation of variant proteins, and to relate clinical drug responses to genetic variation[2]. The pharmacogenetics knowledge base (PharmGKB) stores the data, and is publicly accessible over the Internet (http://www.pharmgkb.org).

A principal challenge for PharmGKB is to integrate complex biological data, pharmacological data, and clinical data in such a way that researchers can contribute results. There are also ethical issues associated with studying populations from different ethnic backgrounds, maintaining confidentiality of data, and addressing issues of intellectual property[3]. However, the focus of this paper is limited to the problem of modeling biological, pharmacological, and clinical data to support the goal of linking genotype to phenotype.

Developing such a resource is challenging because (1) the data are complex, (2) the data come from diverse sources and must be integrated into a single system, (3) terms that identify clinical and biological concepts must be used consistently throughout the software modules and by different users, (4) knowledge is constantly changing, (5) a mixture of experimental data and knowledge about the domain must coexist in the knowledge base, and (6) data and knowledge within the system must be consistent with data stored in external public databases.

These problems are exacerbated by a lack of standard terminologies for clinical and biological concepts and by a lack of standard representations for the experimental data being submitted. However, successful management of names, meaning, and organization of concepts is crucial, and hence, an ontology of formally specified concepts and relationships is central to the design of our system.

2 What is an ontology?

Due to ambiguity associated with the term *ontology*, we first address the question of what an ontology is, and clarify our use of the term for PharmGKB. The word *ontology* was originally used by philosophers to describe a branch of metaphysics concerned with the nature and relations of being[4]. The artificial-intelligence community later adopted the term, but has debated its meaning. Guarino[5] states that in the most prevalent use of the term, an ontology refers to an engineering artifact, constituted by a specific vocabulary used to describe a certain reality, plus a set of explicit assumptions regarding the intended meaning of the vocabulary words. He says that in the simplest case, an ontology describes a hierarchy of concepts related by subsumption relationships, whereas in more sophisticated cases, suitable axioms are added to express other relationships between concepts and to constrain interpretation of those concepts. We adopt this definition, and recognize that it leaves open a number of questions regarding how an ontology should be designed, implemented, and used.

Two types of ontologies that are relevant to PharmGKB are (1) controlled vocabularies and (2) domain ontologies. A controlled vocabulary is a collection of terms organized in a hierarchy intended to serve as a standard nomenclature[6,7,8]. The purpose of a controlled vocabulary is to provide a common set of terms that users of a single system can share (e.g., MeSH in Medline[9]), or that users can share across multiple systems (e.g., Gene Ontology[7]). A controlled vocabulary usually contains concepts, but no instances. A domain ontology, described by Musen[10], is a set of classes and associated slots that describe a particular domain. The purpose of this type of ontology is to serve as a knowledge-base schema, analogous to a database schema. However, unlike a database schema, the domain ontology may also contain classes that are not intended to have instances, but that represent concepts organized in a hierarchy to serve as a controlled vocabulary. When a knowledge-base developer adds instances to classes in a domain ontology, the result is a knowledge base. The domain ontology itself does not contain instances.

For PharmGKB, we need a controlled vocabulary to specify shared names, synonyms, and meanings of concepts, and we need a knowledge-base schema to specify how experimental data are represented and stored. We also need other domain-relevant classes and instances that support various PharmGKB applications. Thus, we need to represent experimental data and domain knowledge, and all of the classes modeled for these purposes contribute to the domain ontology.

3 Methods

We discuss here the two primary tasks that are important in content development for the PharmGKB ontology: (1) modeling experimental data, and (2) modeling domain knowledge. The distinction between data and knowledge is fuzzy, but for our purposes, the data that result from experimental studies are *data*, and the controlled-vocabulary concepts that are used as values for the experimental data, that provide classification or synonym information, or that provide supportive relationship information for applications are *knowledge*. In general, we are taking a bottom-up approach for developing the ontology for the experimental data, and a top-down approach for developing the knowledge.

We have chosen to use a frame-based knowledge-representation system to store both experimental data and domain knowledge. We are using Protégé to build the PharmGKB knowledge base. Protégé[11] is a frame-based knowledge-representation system that offers classes, slots, facets, instances, and slot values as the building blocks for representing knowledge. *Classes* are data structures that may or may not have instances; they have *slots* (sometimes called attributes or roles) that establish relationships between classes. Classes are organized in a hierarchy, and each class has at least one parent (except the root, which has no parent). In PharmGKB, we make the restriction that non-root classes have only one parent. Slots have *slot values* that may or may not be inherited. Slots also have *facets* that specify

cardinality and data-type constraints on the slot value (e.g., string, integer, enumerated symbols, or instance of another class).

3.1 Modeling Experimental Data

The research goals and data requirements of the first five research groups submitting data to PharmGKB provide an initial framework for modeling experimental data. Each research group is studying a set of genes, and each gene of interest codes for a protein that is thought to have an effect on a phenotypic response to a drug. Each protein (e.g., enzyme, transporter, or receptor) affects one or more drugs studied by the research group. To characterize genotypes, investigators look for polymorphisms in the genes of interest in human DNA samples. To link genotype to phenotype, the researchers select phenotypic observations that they can measure at the molecular, cellular, or clinical level, and that they can correlate with genotype.

Table 1 shows the five groups that are providing initial data, and summarizes their research interests[12,13,14]. The groups focus their research in different ways—for example, one group may focus on particular enzymes or transporters, and another may focus on a particular drug class or disease. However, all groups collect data to link genotype to phenotype. In our bottom-up approach to ontology development, we are modeling the PharmGKB ontology to fit the data collected by these research centers. We will expand the model as needed to accommodate additional kinds of experimental data provided by other research centers.

3.2 Modeling Domain Knowledge

Analysis of the areas of interest and data of the five groups suggests broad categories that are appropriate in modeling domain knowledge in pharmacogenetics. Table 2 shows several high-level categories that are useful for organizing controlled-vocabulary concepts in PharmGKB, and gives examples of entities from the researchers' areas of interest that fall into these general categories[12,13,14]. Additional modeling is required to refine the entities into a multi-level classification hierarchy.

Naming conventions are required for biological entities such as genes, alleles, and proteins, for pharmacological entities such as drugs and metabolites, and for clinical entities such as diseases, symptoms, laboratory tests, and test results. Fortunately, standards do exist in certain areas. The Enzyme Commission has established an enzyme nomenclature[15], and the Human Gene Nomenclature Committee has established rules for gene names and maintains the HUGO gene nomenclature[16]. However, a gene sequence may have multiple accession numbers in Genbank, a sequence can be identified by either a GenBank accession number or a LocusLink ID, and although there may be names for certain alleles of a particular

Table 1. Research groups that are providing data initially to PharmGKB and their areas of research. Each research group is identified by the principal investigator of the project and by the primary institutional affiliation of the group.

Research Group	Research Areas
Richard Weinshilboum (Mayo Clinic)	Enzymes involved in phase II metabolism of drugs, especially in methylation and sulfate conjugation reactions, and the impact of genetic variation in genes that encode these enzymes (e.g., TPMT, HNMT, COMT)
Kathleen Giacomini (University of California at San Francisco	Transporter genes, including genes that encode the serotonin transporter (SERT) and vesicular monoamine transporter (VMAT2), and the impact of variation in these genes on efficacy and toxicity of antidepressants
Mark Ratain (University of Chicago)	Pharmacogenetics of anticancer agents, with an emphasis on topoisomerase inhibitor drugs, such as irinotecan and etoposide
David Flockhart (University of Indiana)	Metabolism of tamoxifen by cytochrome P450 enzymes, and effects of genetic variation on pharmacokinetics, clinical efficacy, and adverse effects of tamoxifen
Scott Weiss (Harvard University)	Genetic factors that influence patient response to three classes of drugs used in asthma: (1) inhaled beta agonists, (2) inhaled steroids, and (3) leukotriene modifiers

gene, there may not be a name for every sequence variant of a gene that occurs in the population. In clinical medicine, standards are even less clear. Nevertheless, standards are emerging, and we are evaluating controlled vocabularies maintained by others to determine their suitability for PharmGKB. When necessary, we will develop our own approach, but will do this only when no standards exist.

The two parts of the ontology—the knowledge-base schema for experimental data and the domain conceptual knowledge—are integrated in PharmGKB to support user queries. The classes and instances that form the knowledge may be

Table 2. Examples of categories and entities in these categories, based on data supplied by the five groups.

Category	Entities
Genes	TPMT, HNMT, COMT, SLC6A4, SLC18A2, CYP3A4, CYP2C9, CYP2D6, UGT1A1, IL2, IL4, IL13, IL2RG
Proteins	
Enzymes	cytochrome P450 isoenzymes, methyltransferases, sulfotransferases, UDP-glucuronosyltransferases
Transporters	serotonin transporter (SERT), vesicular monoamine transporter (VMAT2)
Receptors	interleukin receptors, cholinergic receptor
Drugs	selective estrogen receptor modulators (SERMs), inhaled beta agonists, inhaled corticosteroids, leukotriene antagonists, topoisomerase inhibitors, selective serotonin receptor antagonists (SSRIs)
Diseases	asthma, depression, breast cancer, leukemia, colon cancer
Functional Studies	
In vitro	enzyme kinetic studies, measurement of levels of immunoreactive protein
In vivo	pharmacokinetic studies, clinical studies of drug efficacy and adverse effects

used in a variety of ways. For example, the domain conceptual knowledge offers the following features:

1. *Controlled-vocabulary terms.* Researchers who enter experimental data as PharmGKB instances must use names of entities in slot values that are the same as the names used by others who enter or query data. Thus, the domain knowledge provides a shared nomenclature, enforced by our data-acquisition methods.

2. *Alternative names*. Entities may have synonyms or near synonyms, and maintenance of alternative names in the system assists a users in searching for a concept.

3. *Accession numbers*. Accession numbers that are unique identifiers for entities in external databases are stored in PharmGKB to facilitate communicate with those databases. Examples of relevant external databases that have their own coding schemes of identifiers are Genbank, LocusLink, Refseq, PubMed, and Online Mendelian Inheritance in Man.

4. *Classification hierarchy*. The classification hierarchy allows users to browse for terms of interest by navigating up or down the hierarchy. It also permits the user to formulate a query in terms of a single high-level concept, and to apply that query automatically to multiple lower-level concepts that are subsumed by the concept in the original query.

5. *Nonhierarchical relationships between concepts*. Slots can be considered nonhierarchical links between classes. Such links can provide additional knowledge that supports browsing or querying. In PharmGKB, examples of useful associations include gene–enzyme links (a particular gene encodes a particular enzyme), gene–transporter links (a particular gene encodes a particular transporter), drug–metabolite links (a particular drug has a particular set of metabolites), and drug–enzyme links (a particular drug is metabolized by a particular enzyme).

4 Scenario of Use

To demonstrate use of the domain-knowledge hierarchy in conjunction with queries for experimental data, we present a scenario in which a researcher enters experimental data and a user later retrieves the data. Suppose the researcher has completed a pharmacokinetic study on the drug irinotecan and has collected genotype data from individuals in the study. Pharmacokinetic data collected include blood levels of the drug and its metabolites, as well as summary parameters that describe the rise and fall of drug and metabolite levels in the blood. The researcher selects the drug and metabolites by navigating through a display of the controlled-vocabulary hierarchy. Alternatively, he enters the drug and metabolite names as text, and the system confirms whether or not the drug and metabolites are known to PharmGKB. In addition, the system verifies that the metabolites are indeed metabolites associated with irinotecan.

Later, a user querying PharmGKB might search for information on studies of topoisomerase inhibitors. Since irinotecan is categorized in the knowledge hierarchy as a topoisomerase inhibitor, the system returns data from an irinotecan pharmacokinetic study. Another user might search for data on the drug Camptosar, which is the trade name for irinotecan. Since the knowledge hierarchy stores information about trade names that correspond to generic names, the system again returns information about the irinotecan pharmacokinetic study.

Figure 1 shows a portion of the ontology that reflects the knowledge-base schema for experimental data. Figure 2 shows a portion of the ontology that reflects domain conceptual knowledge. In the pharmacokinetic study, the drug is administered to an individual. The individual is identified by a PharmGKB subject identifier. Information about the event in which the subject is given a dose of the drug is stored in an instance of the class *DrugDosingEvent*. The value of the slot *Drug* in this instance would be *Irinotecan*. The entity *Irinotecan* is part of the

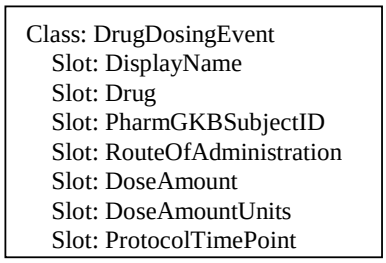

Figure 1. Information about a drug-dosing event. The *DrugDosingEvent* class is part of the knowledge-base schema for experimental data.

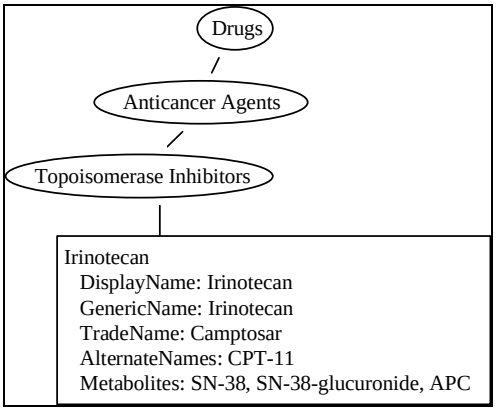

Figure 2. Domain knowledge about the drug irinotecan. In this fragment of the ontology, *Irinotecan* is an instance of the class *Topoisomerase Inhibitors*, and that class exists in a hierarchy of *Drugs*.

controlled vocabulary of drugs stored in the domain knowledge. As shown in Figure 2, *Irinotecan* is classified as one of the *Topoisomerase Inhibitors*, and its trade name, *Camptosar*, is stored in a slot. There is also a slot for metabolites of irinotecan, which enables the system to validate the fact that the metabolites entered as text by the user are indeed metabolites of irinotecan. Thus, in this example, the domain knowledge provides (1) values for experimental data, (2) constraints for data validation, (3) categorical classification support for queries, and (4) synonym support for queries.

5 Preliminary Experience

The initial task that we addressed in our ontology development was the creation of classes that would support submission of gene-sequence data. Network investigators submit data that describe the experiments performed, as well as the associated results. Data that describe experiments include information about the investigators who performed the study, genes studied, primers and methods used, and gene regions analyzed. Results include polymorphisms discovered, the frequency of those polymorphisms in particular populations, and summary single nucleotide polymorphism (SNP) data from pooled samples of DNA. The ontology contains all classes and slots necessary to support the automatic submission of SNPs to the NIH-supported dbSNP resource. When a user submits SNP data to PharmGKB and the SNP is not yet present in dbSNP, PharmGKB automatically sends a new SNP submission to dbSNP.

Prior to release of our sequence-submission software, we conducted a test with collaborators from the Mayo Clinic. The Mayo group produced a sample set of submissions to describe experimental data collected in their studies of the HNMT gene. They incorporated their data into the XML format required for direct XML submissions. The required format is specified by the PharmGKB XML schema (http://www.pharmgkb.org/xml-schemas.html), and that schema corresponds directly to the PharmGKB ontology.

In the development of the ontology for gene-sequence submissions, we encountered a number of subtleties of definition that had to be clarified before ontology developers and data submitters reached a shared understanding. We describe here several of the most important constructs in the resulting ontology.

A *reference sequence* is a specified sequence of bases against which variations are compared. A reference sequence is associated with a gene and may (but need not) correspond exactly to a sequence already deposited in GenBank. However, a reference sequence is not required to contain the entire gene structure associated with a gene. Reference sequences can be different molecule types (DNA, RNA, or protein). The only restriction is that they consist of a contiguous series of monomers. Gaps in the sequence or fragments pasted together are not allowed.

A *sequence coordinate system* is required to ensure agreement about how to identify a particular position in a reference sequence. The sequence coordinate system specifies which base is labeled +1, and indicates whether the base that precedes position +1 is numbered 0 or -1.

A *region of interest* identifies a segment of a reference sequence that is of interest to the investigators. It may be a subsequence of the reference sequence, or it may be the entire reference sequence.

```
<ForwardPcrPrimer>
<DisplayName>Exon 5 Forward Primer</DisplayName>
<FirstAnnealingPositionInPrimer>19</FirstAnnealingPositionInPrimer>
<FirstAnnealingPositionInRegion>6</FirstAnnealingPositionInRegion>
<LastAnnealingPositionInPrimer>41</LastAnnealingPositionInPrimer>
<LastAnnealingPositionInRegion>28</LastAnnealingPositionInRegion>
<Sequence>TGTAAAACGACGGCCAGTAGGAGTATCTAGCCCAAGCAATA</Sequence>
</ForwardPcrPrimer>
```

Figure 3. Sample input data from Mayo test submission in XML format. This sample input shows the submission of a forward PCR primer used, and is stored as experimental data in PharmGKB.

A *simple nucleotide difference* (SND) defines a position in a region of interest of a reference sequence where bases differ from the bases in the corresponding location of a tested sequence. A SND is *simple* because the bases in the variant segment must be contiguous, rather than located in different parts of the genome. A SND differs from a single nucleotide polymorphism (SNP) in that there is no frequency restriction in the definition of a SND. In contrast, when scientists perform SNP detection assays, it is common practice to filter out SNPs that have allele frequencies that are less than some threshold percentage (e.g., 10 percent). Also, in the spirit of dbSNP, a SND can refer to an insertion, a deletion, or a variable number of repeats, as well as to a single nucleotide difference. The convention in PharmGKB for specifying where the difference is located is to identify the position in the reference sequence that precedes the variant site (the position upstream in the 5' direction). This approach provides a consistent method for describing variant positions across all polymorphism types.

Figure 3 shows a representative sample of data from the Mayo test submission. It shows the submission of a forward PCR primer used in an experiment. Annealing positions are based on a numbering scheme that was previously specified in a sequence coordinate system for the reference sequence.

6 Concluding Remarks and Future Work

Given the potential impact of pharmacogenetic research and the vast quantities of data that are likely to result from efforts to link genotype to phenotype, the NIH has begun a program that encourages collaboration among investigators and that mandates public sharing of data. The value of PharmGKB as a resource for sharing pharmacogenetic experimental data and knowledge lies not only in its commitment to public dissemination of data, but also in its demonstration of the use of knowledge representation techniques to organize pharmacogenetic knowledge and data. There is currently no standard data model for pharmacogenetic knowledge, and without standards for names and meanings of terms, it is difficult to share information in computer-based systems. Thus, the ontology effort is essential to the

success of this project, and may contribute to ontology development done by others who work in this area.

Our ontology development process is a process of iterative development and communication between bioinformatics professionals and other collaborators, including molecular biologists, chemists, clinical pharmacologists, and clinicians. Our bottom-up approach to modeling experimental data allows us to take a staged-delivery approach in software development. We can provide software that is usable to a few groups initially, and then extend it in a controlled fashion. However, our top-down approach to knowledge modeling also encourages us to consider the broader picture in the early stages.

Our ontology is comprised of the data model for experimental data, and the domain conceptual knowledge that provides controlled-vocabulary information and other knowledge that supports queries. These two parts are integrated in PharmGKB, but it is useful to distinguish them because the former is essential for communication with our collaborators who submit data, and the latter is essential for management of shared concepts in the system. Together, these two parts form the ontology that may be reusable in other settings in the field of pharmacogenetics.

Future work on the PharmGKB ontology includes (1) expansion of content to broaden the scope, (2) enhancement of constraint representation in the ontology to support automated or semi-automated data validation, (3) extension of change logging features to facilitate change management, (4) development of merging techniques to support the process of merging the production version of the knowledge base with the development version when a new version is released, and (5) enhancement of methods that help users to query PharmGKB in an intuitive manner to obtain genotype–phenotype associations.

Acknowledgements

PharmGKB is financially supported by grants from the National Institute of General Medical Sciences (NIGMS), the Human Genome Research Institute (NHGRI) and the National Library of Medicine (NLM) within the National Institutes of Health (NIH). This work is supported by the NIH/NIGMS Pharmacogenetics Research Network and Database grant U01GM61374, and by Stanford University's Children's Health Initiative. JMS is supported by National Library of Medicine grant LM07033.

References

1. Long RM, Giacomini KM. Announcement. June 1, 2001 http://www.nigms.nih.gov/pharmacogenetics/editors.html
2. RFA GM-00-003, April 7, 2000 http://grants.nih.gov/grants/guide/rfa-files/RFA-GM-00-003.html
3. MA Rothstein, PG Epps "Ethical and legal implications of pharmacogenomics" *Nature Review Genetics*, **2**, 228-231 (2001)
4. *Webster's New Collegiate Dictionary*, 9th edition, Ontology, p. 825 (Springfield, MA: Merriam-Webster, 1991)
5. N Guarino, "Formal ontology and information systems" *Proceedings of FOIS '98*, Trento, Italy, June 6-8, 1998. (Amsterdam, IOS Press, 1998) pp. 3-15
6. C Price, M O'Neil, TE Bentley, PJB Brown, "Exploring the ontology of surgical procedures in the Read Thesaurus" *Methods of Information in Medicine* **37**, 420-5 (1998)
7. The Gene Ontology Consortium, "Gene ontology: tool for the unification of biology" *Nature Genetics* **25**, 25-9 (2000)
8. D Fensel, "Ontologies and electronic commerce" *IEEE Intelligent Systems* January/February, 8 (2001)
9. "Medical Subject Headings" http://www.nlm.nih.gov/mesh/meshhome.html
10. MA Musen, "Domain ontologies in software engineering: Use of Protégé with the EON architecture" *Methods of Information in Medicine* 37(4–5), 540-50 (1998)
11. "Welcome to the Protégé project" http://www.smi.stanford.edu/projects/protege/
12. "PharmGKB Investigators" http://www.pharmgkb.org/investigators.html
13. "Query PharmGKB" http://www.pharmgkb.org/PharmGKB/query
14. "Pharmacogenetics Research Network and Knowledge Base First Annual Scientific Meeting" April 25, 2001 http://pub.nigms.nih.gov/pharmacogenetics
15. "Enzyme nomenclature" http://www.chem.qmw.ac.uk/iubmb/enzyme
16. "HUGO Gene Nomenclature Committee" http://www.gene.ucl.ac.uk/nomenclature