Drug Response and Genotype

- Patient responses to drugs are variable and sometimes unpredictable
- Adverse drug reactions account for more than 2 million hospitalizations and 100,000 deaths in 1994
- Current approach: historical; risk stratification (clustering; classification)
- Response to some drugs has a genetic basis
- Desired approach: individualized treatment based on genotype

Genotype and Phenotype

- Genotype
  - Genetic makeup
  - Genetic sequence of DNA in an individual
- Phenotype
  - Visible trait (eye color, disease, etc.)
  - Manifestation of a genotype

Pharmacogenetics

- Discipline to understand how genetic variation contributes to differences in drug responses
- Methods: genotype-phenotype studies
- Goal: drug treatment tailored to individual patients
- Promises: new drug discovery and treatments by mining genome & SNP databases
Pharmacogenetics: A Case Study

Need Integrated Resource for Pharmacogenetics

- Proliferation of experimental data
  - Gene sequencing studies
  - Biological and clinical studies of phenotype
- Need to connect genotype \(\leftrightarrow\) phenotype
- Gives insight into gene–drug relationships
- Understand how genetic variation contributes to differences in drug responses


PharmGKB: Pharmacogenetics Knowledge Base of the NIH

- The Pharmacogenetics Knowledge Base (PharmGKB - http://pharmgkb.org)
- Part of the Pharmacogenetics Research Network
  - Nationwide collaborative research effort funded by NIH
- Accepting data from 10 study centers and public sources

NIH Pharmacogenetics Research Network (Initial Study Centers)

Goals of PharmGKB

- National data resource linking genetic, laboratory data, and clinical data
- Contain high quality publicly-accessible data
- Link with complementary databases (Medline, dbSNP, Genbank, etc.)
- Assist researchers discover genetic basis for variation in drug response
- Receive genotype/phenotype data from participating study centers
- Analytical functionality to link genotype and phenotype

Goal State of PharmGKB

Current State of PharmGKB

PharmGKB Infrastructure
Issues in Designing PharmGKB

- Data acquisition from study centers
- Data integration with external sources
- Data model
- Data storage (DBMS/KBMS)
- Query support
- User tools (visualization, etc.)

What are the Data?

- Genotype data
  - Genetic sequences
  - Polymorphisms in individuals
- Cellular phenotype data
  - Gene expression & proteomics
  - Functional assays
  - Pharmacokinetics & pharmacodynamics
- Clinical data
  - Drug responses and clinical outcomes

What are Polymorphisms?

Reference Sequence

ATATCGGATAC - - - TACCCGTATTA
ATATCGGGGTACATATTACC - - ATTA

Subject Sequence

SNP Insertion Deletion

Sources of Data

<table>
<thead>
<tr>
<th>Direct Submission</th>
<th>External Sources</th>
<th>Terminologies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pharmacogenetics</td>
<td></td>
<td>SNO-MED</td>
</tr>
<tr>
<td>Research Network</td>
<td></td>
<td>Gene Ontology</td>
</tr>
<tr>
<td></td>
<td></td>
<td>dbSNP</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UMLS</td>
</tr>
<tr>
<td>Other Researchers</td>
<td></td>
<td>Other Vocabularies</td>
</tr>
<tr>
<td>Other Genetics Databases</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Page 4
Challenges for PharmGKB
- DB vs. KB (relational model vs. ontology)
- Data integration
  - Data from study centers
  - Data from external databases
- Ontology evolution
  - Maintain mapping from external data input/output formats to internal representation
  - Change management between development & production versions (schema update problem in databases)
- Data validation, data editing/audit trail

Biomedical Databases
- Paper
- Electronic versions of paper (pdf, img files)
- Spreadsheets
- Text files or other formats
- RDBMS
- OODBMS
- KBMS (e.g., frame systems)
Definitions

- **Data**: simple description of an observation; lowest level of known facts
- **Information**: data that has been sorted, analyzed, and interpreted so known facts have substance and purpose
- **Knowledge**: information that has been placed in the context of other information
- **KB**: a computational repository of knowledge, and the information and data that the knowledge is built upon

KB vs. DB: The Difference is the Data Model

- In many ways, KB & DB are interchangeable
  - Data model can be implemented in RDBMS or KBMS
  - “KB” can be implemented in RDBMS
- Difference in data model
  - DB: relations, relational schema
  - KB: frames, ontology (locality of information)
- Data model for DB in form to facilitate retrieval
- Data model for KB in form to facilitate reasoning

Data Model In a Relational System

Data Model In a Frame-Based System
Data Model for PharmGKB

**Ontology** is preferred for PharmGKB

- **Domain complexity**
  - Many entities and relationships (i.e., hierarchical)
  - Multi-valued attributes (simple & object types)
- **Rapid evolution of data model** → changing database schema
- Storage schema can closely parallel “common data model”
- Support applications relying on inheritance & other relationships in ontology
- Reasoning over information in KB

---

Data Models in Genetic Databases

- Data can be described in “flat” tabular representations (entry + attributes)

<table>
<thead>
<tr>
<th>Genbank Accession</th>
<th>Locus</th>
<th>Definition</th>
<th>Version</th>
<th>Segment</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>U44106</td>
<td>HSNMT1</td>
<td></td>
<td></td>
<td>1 of 6</td>
<td>angggcagagtca</td>
</tr>
</tbody>
</table>

- Relational schema appropriate
- Fine for pre-defined functionality (BLAST, etc.)
- Goal: storage/retrieval; less so for analysis

---

Domain Complexity in Pharmacogenetics

- Different distinctions in the same data
e.g., for sequences:
  - String of letters making up the sequence
  - Genomic structure of the sequence
  - Polymorphisms in the sequence
  - Haplotypes of the sequence
- Many relationships
  - Genes have sequences; sequences have genomic structure; individuals have polymorphisms in sequences...

---

More than Letters in a Genetic Sequence

- Coding regions
- Flanking sequence
- Exons/introns
- Primer regions
Different Entities for “Sequence”

Genomic DNA

Control region
5' UTR
INT
INT
INT
INT

5' Partial Coding
Transcribed
Sequences

3' Partial Coding
Transcribed
Sequences

Translated
Translated
INT
INT
INT
INT
INT
INT

3' UTR

5' UTR

POLYMORPHISM

Amino Acid Sequence

Translation

Coding sequence

Spliced sequence

Transcribed sequence

Relationships Among Entities

Gene

Reference Allele

Polyploid Alleles

Diploid Alleles

Haploid Alleles

Amino Acid Sequences

Spliced Sequences

Transcribed Sequences

Polymorphisms

Fragment from Reference Allele

Alleles of genomic fragments

Primers

Complexity of Relationships in Pharmacogenetics

Genomic Information

Molecula & Cellular Phenotype

Clinical Phenotype

Molecules

Alleles

Individuals

Drugs

Environment

Our Approach to Modeling Genetic Information for Pharmacogenetics

- Data Model: Ontology
  - Well-suited to complex/diverse data types
  - Specifies:
    - the classes of information in the domain
    - the attributes for these concepts
    - the relationships among these concepts
    - Intuitive connection to real objects in the world
- Flexible; suitable for evolving databases
- Implementation: frame-based systems
Pre-defined Queries vs. Open API to DB

- Predefined queries & functionality
  - e.g., free-text/keyword search; BLAST
  - User does not directly see DB schema (if at all)
  - DB schema understood only by administrator
    - Can be optimized for performance
    - Hard to understand by external user
- Open API for queries
  - Users can formulate customized queries
  - User must understand the data schema

A Comparison Study

- PharmGKB data model for genetic information implemented in:
  - RDBMS: Oracle 8.1.7
  - KBMS: Protégé-2000
- Sample queries pertinent to pharmacogenetics
- Approximate timings on queries*
- Comparison of database schemas

*Big grain of salt
What is Protégé-2000*?

- A tool that allows you to create and maintain an ontology by:
  1. Constructing a domain model using classes and slots
  2. Customizing forms for acquiring instances of classes
  3. Entering data as instances
  4. Querying for instances that match your criteria

- A platform on which you can build applications

- A library you can use from other applications

*http://protege.stanford.edu*
**Data Model In a Relational System**

**SQL Query to RDBMS**

Query: For each subject, find all the variants

```sql
SELECT t0.displayname, t7.precedingvarpos+1, t7.variant, t0.displayname, t7.precedingvarpos+1, t7.variant,
       substr(t1.sequence, t7.precedingvarpos+1, 1), t7.subjectident
FROM genesubmission t0, refseqsubmission t1, seqcoordsubmission t2, expregionsubmission t3,
     pcrassaysubmission t4, indivsndsubmission t5, indivsndvariant t6, subjectvariant t7
WHERE t0.displayname = t1.gene AND t1.displayname = t2.refseq AND t2.displayname = t3.seqcoord AND t3.displayname = t4.expregion AND t4.displayname = t5.sndassay AND t5.displayname = t6.indivsnd AND t6.subvariant = t7.displayname AND NOT (substr(t7.variant, 1, 1) = substr(t1.sequence, t7.precedingvarpos+1, 1)),1)
```

**Query to KBMS (pseudocode of java program)**

Query: For each subject, find all the variants

1. Get all instances of Subject Variants class
2. For each instance:
   - Get its Subject
   - Get its Variants
   - Add the variants to subject groupings
3. Print the groupings

**Query Performance**

<table>
<thead>
<tr>
<th>Query</th>
<th>Timing for Query (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>How many regions of interest are in the MDR1 gene?</td>
<td>2.0/0.6</td>
</tr>
<tr>
<td>List all regions of interest and start/stop positions relative to the reference sequence</td>
<td>1.3/0.04</td>
</tr>
<tr>
<td>For each variant, what is the base at that same position in the reference sequence? (e.g. for 97 G/G variant, what is position 98 in reference sequence?)</td>
<td>333/8.8</td>
</tr>
<tr>
<td>Which subject has the most variants?</td>
<td>139/1.4</td>
</tr>
<tr>
<td>For each subject, find all the variants</td>
<td>5.5/0.5</td>
</tr>
</tbody>
</table>
Challenges for PharmGKB

- DB vs. KB (relational model vs. ontology)
- Data integration
  - Data from study centers
  - Data from external databases
- Ontology evolution
  - Maintain mapping from external data input/output formats to internal representation
  - Change management between development & production versions (schema update problem in databases)
- Data validation, data editing/audit trail

Need to Integrate Different Data Models

- Ontology (PharmGKB data model)
  - Describes pharmacogenetics concepts & relationships among them
  - Flexible and highly expressive
  - Suitable for rapidly evolving knowledge bases
- Relational (incoming study center data)
  - Tabular
  - Predominant in most biology databases
- Data Integration Task:
  - Import study center data into PharmGKB

Our Work Addresses this Problem

Goals

- Interface ontology models with external relational data sources
- Import raw sequence data (relational) into ontology of pharmacogenetics
- Automate updating links between ontology and data acquisition when ontology changes
### Relational Data vs. Ontologies

- **Study Center data in Relational Format**
- **PharmGKB Ontology**

### Current Approaches to Integrating Relational Data into Ontologies

- **Direct data entry into ontology**
  - Requires understanding of ontology structure
  - Usually different from “intuitive” view of data
- **Static mappings**
  - Map each slot in ontology to column in table
  - Difficult to maintain as ontology changes
- **The challenge: maintaining the links as the ontology changes**

### Direct Data Entry Into Ontology

- **Submitter directly creates instances in ontology**
- **Submitter must understand ontology model**

### Static Mappings for Data Integration

- **Submitter matches data columns to map columns**
- **Links must be maintained as ontology changes**
Our Approach

- Declarative interface between relational data acquisition and ontology
  - XML schema
    - Defines mapping & constraints on incoming data
  - Ontology stores information needed to specify XML schema
  - Automated update of XML schema when ontology changes
- Incoming data in XML
  - Existing relational tables mapped to XML schema

XML Schema

- Self-describing syntax for defining valid XML documents
- Derived from ontology
- Updated as ontology changes

```xml
<xsd:element name="PCR_Assay_Submissions">
  <xsd:complexType>
    <xsd:sequence>
      <xsd:element name="Comment" type="xsd:string" minOccurs="0" maxOccurs="1"/>
    </xsd:sequence>
  </xsd:complexType>
</xsd:element>
```

The XML Schema is Defined by the Ontology

- Facets on slots define data constraints
  - Range of legal values
  - Data type (string, number, Instance, or Class)
  - Required or optional
  - Single or multiple cardinality
- When ontology changes, facets change too!
  - Updated XML schema immediately available
- Code handling XML remains unchanged
Storing Information Needed to Specify the XML Schema in Ontology

Data Model Evolves

New Slot

Classes, Slots, and Facets in PharmGKB Ontology

Evaluation

- Study center mapped sequence data to XML schema
- Data submitted to PharmGKB in XML
  - PharmGKB internal storage format: ontology
  - Output (query) format: relational, like original data
- Ontology changed—XML schema rapidly updated
- No change needed in processing code

Input Experimental Data in Relational Format

<table>
<thead>
<tr>
<th>Reference Sequence</th>
<th>Assayed SNP Positions</th>
</tr>
</thead>
<tbody>
<tr>
<td>U44106</td>
<td>NT Position in GenBank Sequence</td>
</tr>
<tr>
<td></td>
<td>1002</td>
</tr>
<tr>
<td>&quot;Wild Type&quot; Nucleotide</td>
<td></td>
</tr>
<tr>
<td>Variant Nucleotide</td>
<td></td>
</tr>
<tr>
<td>Subject 1</td>
<td>C</td>
</tr>
<tr>
<td>Subject 2</td>
<td>T</td>
</tr>
<tr>
<td>Subject 3</td>
<td>C</td>
</tr>
<tr>
<td>Subject 4</td>
<td>C/T</td>
</tr>
</tbody>
</table>
Experimental Data in XML

```xml
<Variants_In_Individuals>
  <DisplayName>SNP@1002</DisplayName>
  <Assay>PCR 10B</Assay>
  <Sample>PDR</Sample>
  <Subject_Variants>
    <DisplayName>Subj_3_SNP</DisplayName>
    <SubjectID>3</SubjectID>
    <Position>1002</Position>
    <Variant>CC</Variant>
  </Subject_Variants>
  <Subject_Variants>
    <DisplayName>Subj_4_SNP</DisplayName>
    <SubjectID>4</SubjectID>
    <Position>1002</Position>
    <Variant>C/T</Variant>
  </Subject_Variants>
</Variants_In_Individuals>
```

Data in Ontology Viewed in Relational Form

<table>
<thead>
<tr>
<th></th>
<th>Assayed SNP Positions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference Sequence</td>
<td>U44106</td>
</tr>
<tr>
<td>Subject 1</td>
<td>C/T</td>
</tr>
<tr>
<td>Subject 2</td>
<td>T</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Comment</th>
<th>Display Name</th>
<th>Position Preceding Variation</th>
<th>Subject Identifier</th>
<th>Subject Variants OT</th>
<th>Variant</th>
</tr>
</thead>
<tbody>
<tr>
<td>BNMTtep-4</td>
<td>126745291</td>
<td>271</td>
<td>U44106</td>
<td>C/T</td>
<td></td>
</tr>
<tr>
<td>BNMTtep-4</td>
<td>126745304</td>
<td>271</td>
<td>U44106</td>
<td>C/T</td>
<td></td>
</tr>
</tbody>
</table>

Result: A Transparent Interface Between Ontology and Data

[PharmGKB Data Model]
Conclusions (1)

- An ontology provides a flexible data schema
- Built ontology of pharmacogenetics information
- Model is expandable; permits broad range of queries
- Data model close to the biological model is useful
- Tradeoffs between RDBMS/KBMS
- Practical issues of importing data and data integration overwhelm theoretical issues

Conclusions (2)

- Method for integrating ontology and relational data
- XML schema interface
  - Simplifies mapping to relational data
  - Shields user from ontology structure
- XML for data exchange--keeps the data in clear, human-readable format
- Can rapidly update XML schema interface even after ontology changes

Future Work

- Develop improved database back end for KBMS
- Provide graphical views
- Develop open API for querying KB
- Develop analytic routines

Acknowledgments

- Russ Altman, M.D., Ph.D.
- Teri Klein, Ph.D.
- Micheal Hewett, Ph.D.
- Diane Oliver, M.D., Ph.D.
- Mark Woon
- Steve Lin
- Katrina Easton
- NIH/NIGMS Pharmacogenetics Research Network and Database (U01GM61374)
Thank you.

Contact info:
rubin@smi.stanford.edu
help@pharmgkb.org
http://www.pharmgkb.org/