

PharmGKB: The Pharmacogenetics Knowledge Base



Daniel L. Rubin, M.D., M.S.

Stanford Medical Informatics
Stanford University School of Medicine

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

Individuals respond differently to the anti-leukemia drug 6-mercaptopurine.

Most people metabolize the drug quickly. Doses need to be high enough to treat leukemia and prevent relapses.

Others metabolize the drug slowly and need lower doses to avoid toxic side effects of the drug.

A small portion of people metabolize the drug so poorly that its effects can be fatal.

The diversity in responses is due to variations in the gene for an enzyme called TPMT, or thiopurine methyltransferase.

After a simple blood test, individuals can be given doses of medication that are tailored to their genetic profile.


Normal dose

Dose for an extra slow metabolizer (TPMT deficient)

<http://www.nigms.nih.gov/news/reports/testim99.html#pharm>

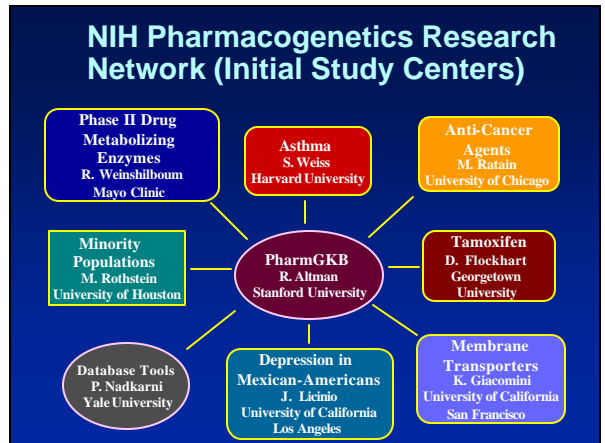
Need Integrated Resource for Pharmacogenetics

- 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

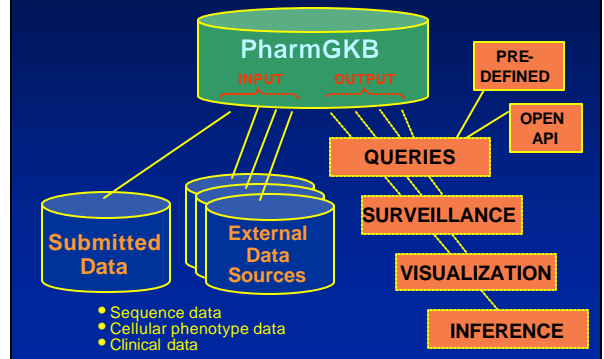
- The **Pharmacogenetics Knowledge Base (PharmGKB)** <http://pharmgkb.org>
- Part of the **Pharmacogenetics Research Network**
 - Nationwide collaborative research effort funded by NIH
- Accepting data from **10 study centers and public sources**



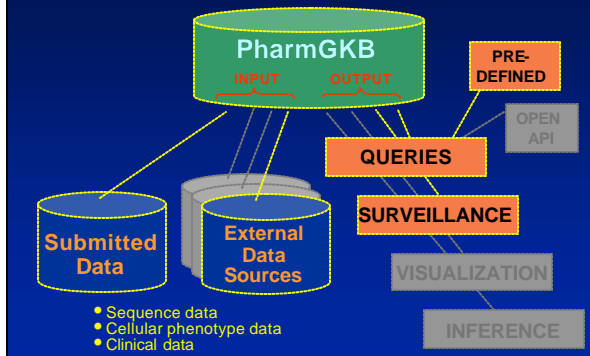
Goals of PharmGKB

- 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

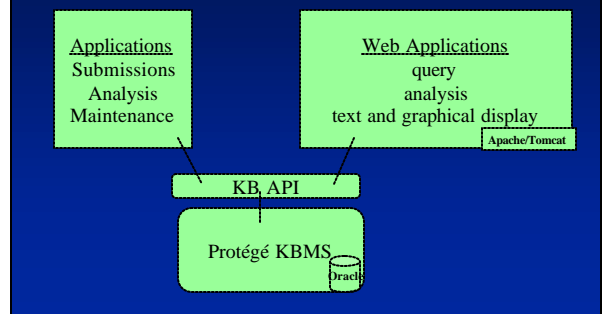
Goal State of PharmGKB



Current State of PharmGKB



PharmGKB Infrastructure



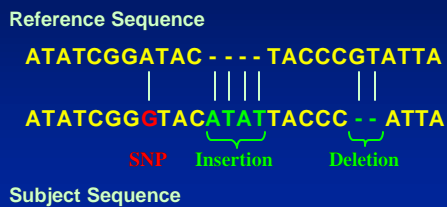
Issues in Designing PharmGKB

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

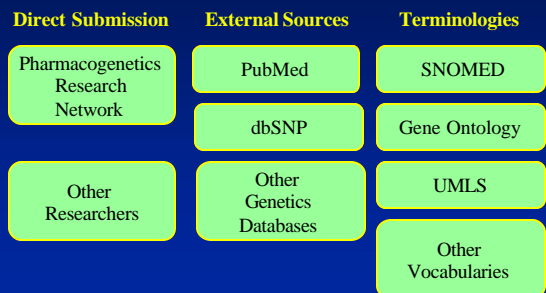
What are the Data?

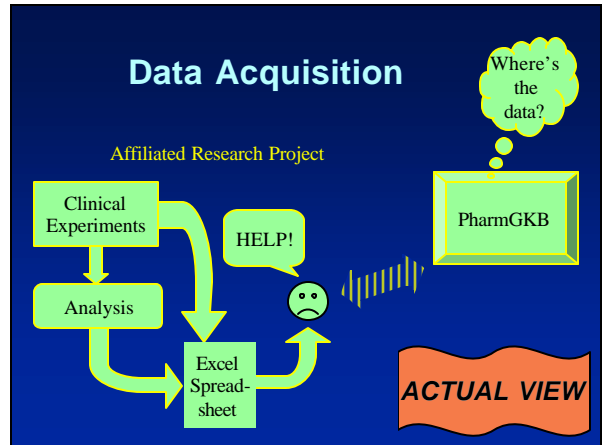
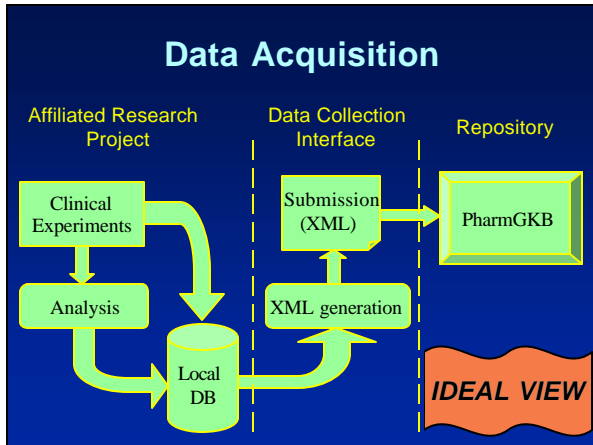
- Genotype data
 - Genetic sequences
 - Polymorphisms in individuals
- Cellular phenotype data
 - Gene expression & proteomics
 - Functional assays
 - Pharmacokinetics & pharmacodynamics
- Clinical data
 - Drug responses and clinical outcomes

What are Polymorphisms?



Sources of Data





- ### 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

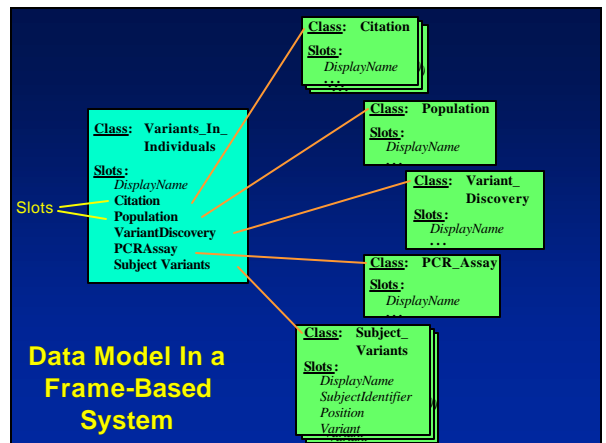
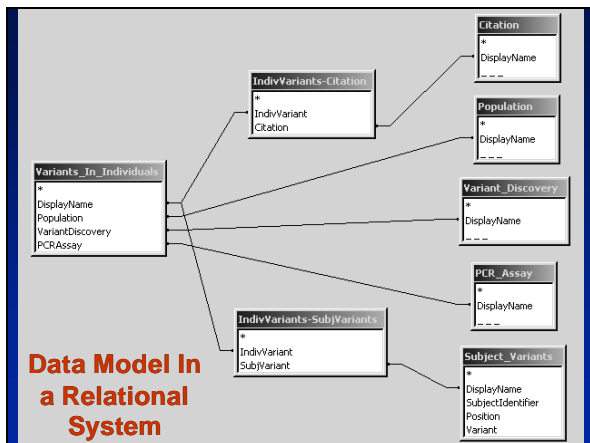
- **Data:** simple description of an observation; lowest level of known facts
- **Information:** data that has been sorted, analyzed, and interpreted so known facts have substance and purpose
- **Knowledge:** information that has been placed in the context of other information
- **KB:** a computational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

Gully A. P. C. Burns
http://www-hbp.usc.edu/_Documentation/presentation/neuroscholar_cns98/

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 for PharmGKB

Ontology is preferred for PharmGKB

- Domain complexity
 - ◆ Many entities and relationships (is-a, 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)

Genbank Accession	Locus	Definition	Version	Segment	..	Sequence
U44106	HSHNMT01	Human, HLA region, N. on chromosome 6 (HNM1) gene, exon 1	U44106.1 GI:1256584	1 of 6	..	aagggcagagctca ...

- 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...

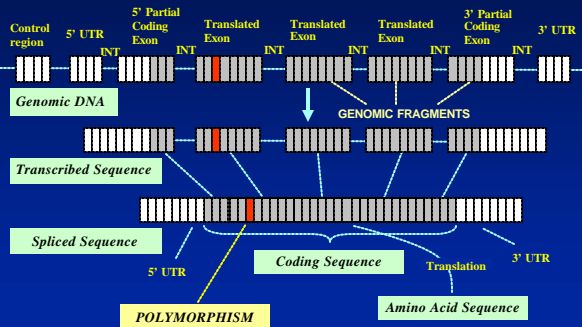
More than Letters in a Genetic Sequence

- Coding regions
- Flanking sequence
- Exons/introns
- Primer regions

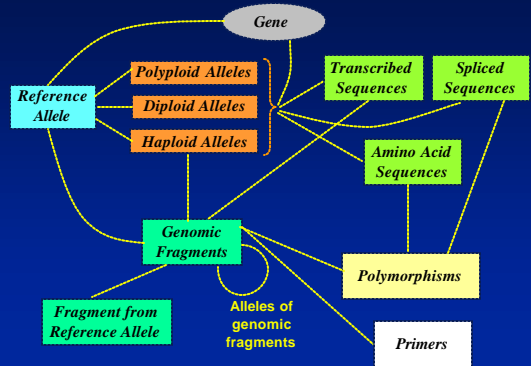
```

1825+
1 ctgcatcttc caagctccc agttaaagat tgttaatgaa taaaacctat abtttgaat 004310
61 ataactaata gatggcaata taactgalat aatgggaca tttaagtgtg gceatgttt Exon3
121 caticcattgt atttttagtc tgtctcttcc aacagacta gataatcaca ttcaacaag
181 caactbaaca atttttctaa aactacataa ttttttctt tcaqGATTGG AGACACAAA
241 TGAGAAATTA AGATTCTAAG CATAGCCGGA GGTGCAGtga tgagtaabat abttttaaag
301 tt atttaacca ttatctctgt tgaatgcaat a . . . . INTRON 3 .
2166+
1 catctttgat ttgatgaat atagtgatag atgttaaaga tcatgtaaac gaatgatgg 025508
61 cactcaacgc cctccttgag tcaacttaet atgctaactt agaacctagc tgccctgcat Exon4
121 catggcaggg cagcagtga acattattct ttatttatgt taggcttcc tagtaagggt
181 aagcaataa ataaatcaac ta atgttt tttaactatt tcaagtgga atagatgac
241 ctgtctctca tctttagctt taagtgaaa ttagacttca AATCTCICC AAGGTCAGG
301 CTCAAATACC AGGAGTTTGT ATCAACAATG AAGTTTGA GCCAAGTCT GAACAAATTG
361 CCAATACAA AGgtacctgt: aactcctggt cctctacacc agatcctatc ccaaaa
aactcaaat tcttcccttg aatgattaaa aatatagtta ctgggtatg cttttcaaca
481 gcttttggg agagaacctg aattagtctc tgggggat gactaacat ctcaaatgt
541 gaacagtgaa taataaactc ccttttctat taacaattca tcaatcccc agttgtatc
601 aatgattaac ttttggatgt ttagctttaa gtaacagctc at. . . . INTRON 4 .
    
```

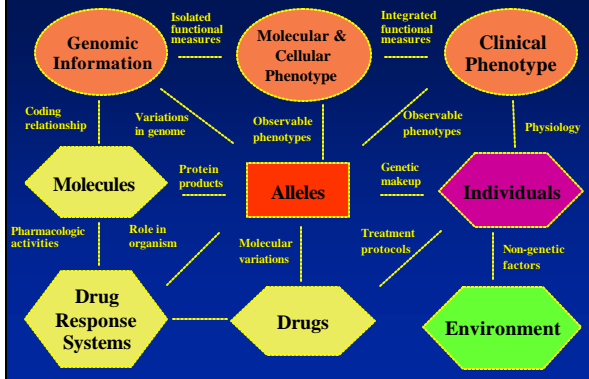
Different Entities for "Sequence"



Relationships Among Entities

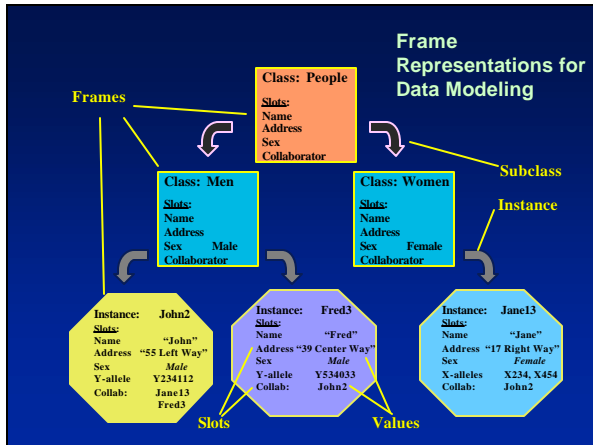


Complexity of Relationships in Pharmacogenetics



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**



Database Schema Should Match Common Data Model

- Queries are not predefined—users must interact directly with schema
 - Open API for queries
 - Need to understand database schema
- Data integration from external databases having differing schemas
- *Analysis* is as important as storage/retrieval
 - Analytical functions not predefined—users must be able to write applications

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

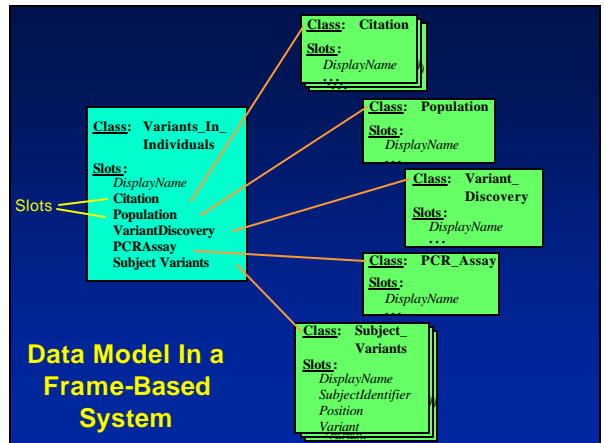
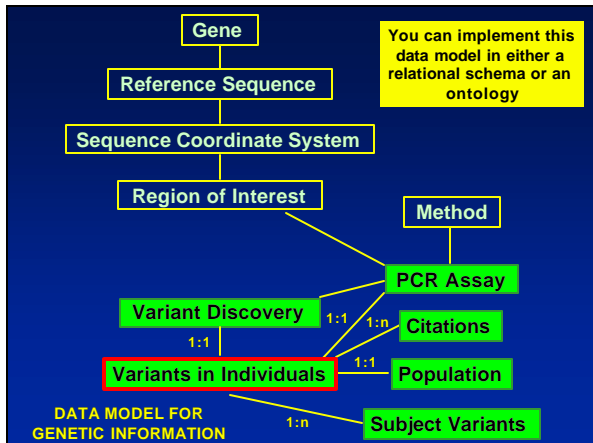
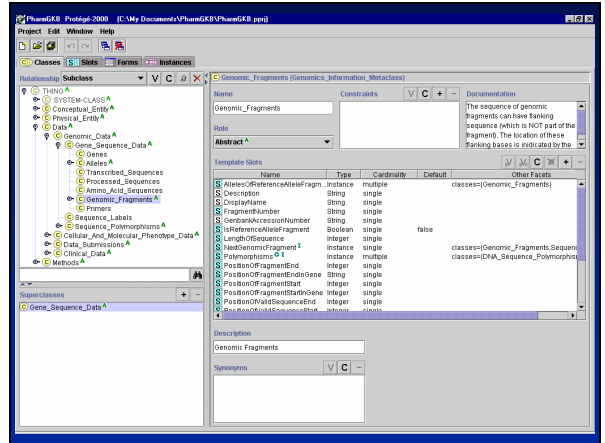
- PharmGKB data model for genetic information implemented in:
 - RDBMS: Oracle 8.1.7
 - KBMS: Protégé-2000
- Sample queries pertinent to pharmacogenetics
- Approximate timings on queries*
- Comparison of database schemas

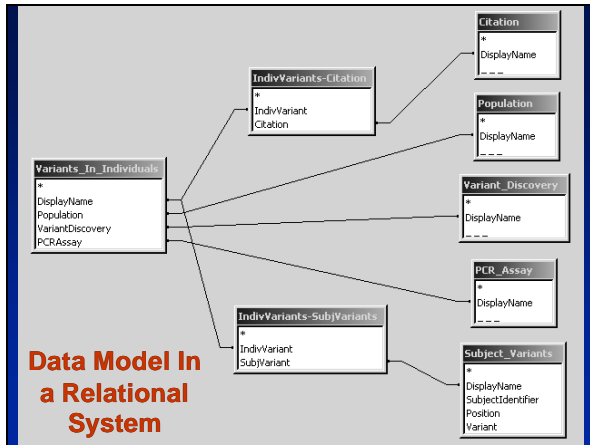
*Big grain of salt

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/>





SQL Query to RDBMS

Query: For each subject, find all the variants

```

SELECT 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, indivsndssubmission t5, indivsndvariant t6,
subjectvariant t7
WHERE t0.displayname = t1.gene AND t1.displayname = t2.refseq
AND t2.displayname = t3.seqcoord AND
t3.displayname=t4.expreion 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) AND
substr(t7.variant,3,1) = substr(t1.sequence,t7.precedingvarpos+1,1))

```

Query to KBMS (pseudocode of java program)

Query: For each subject, find all the variants

Get all instances of *Subject Variants* class
for each instance:
 get its Subject
 get its Variants
 add the variants to subject groupings
print the groupings

Query Performance

Query	Timing for Query (seconds)	
	Ontology	Relational
1 How many regions of interest are in the MDR1 gene?	2.0/0.02	0.6
2 List all regions of interest and start/stop positions relative to the reference sequence	1.3/0.04	0.6
3 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?)	338/3.8	8.4
4 Which subject has the most variants?	139/1.4	4.1
5 For each subject, find all the variants	5.5/3.0	0.6

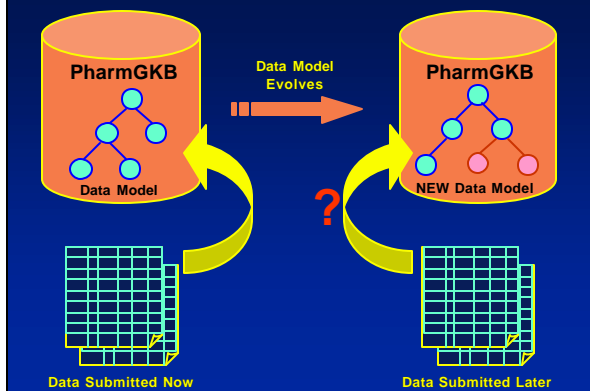
Challenges for PharmGKB

- 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

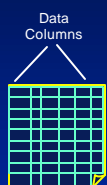
Our Work Addresses this Problem



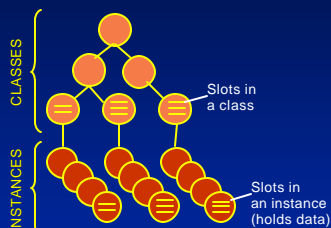
Goals

- 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

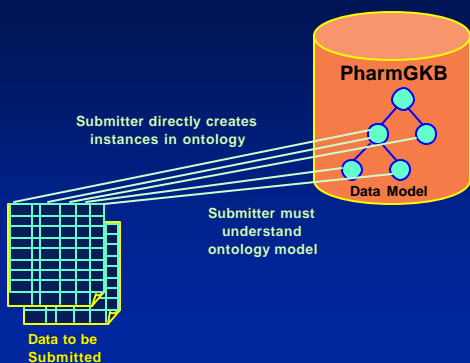


PharmGKB Ontology

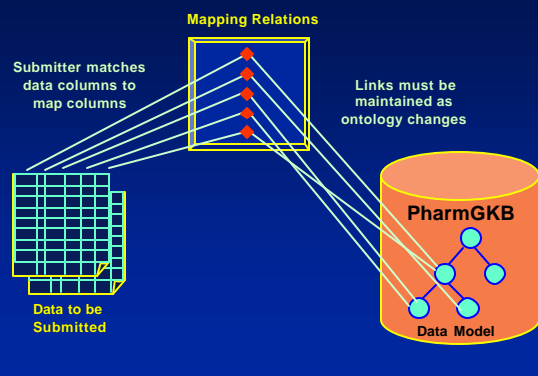
Current Approaches to Integrating Relational Data into Ontologies

- **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

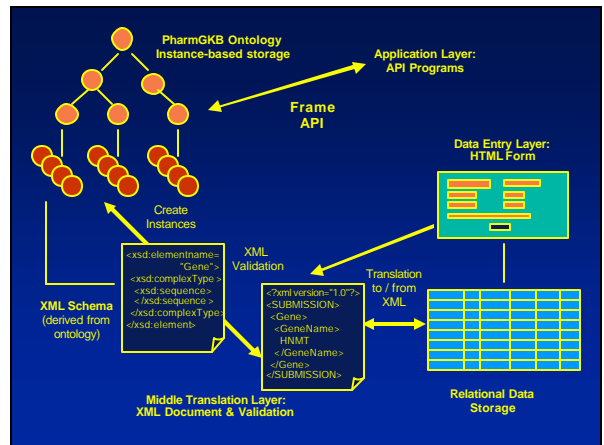


Static Mappings for Data Integration



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

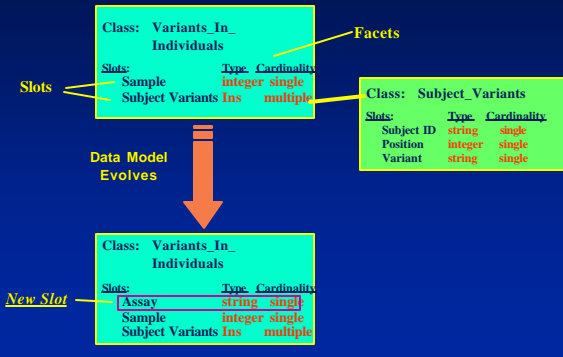
- **Self-describing syntax for defining valid XML documents**
- **Derived from ontology**
- **Updated as ontology changes**

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



Classes, Slots, and Facets in PharmGKB Ontology

PharmGKB ontology

Slots/facets for PCR Assay Class

Name	Type	Cardinality	Other Facets
DisStoid	String	single	
DisplayName	String	required single	
ExperimentalRegionSubmission	Instance	required mult.	classes={Region_Of_Interest_Submissions}
FirstPositionInterrogatedRange	Integer	required single	
FivePrimeFlankingSequence	String	single	
ForwardPcrPrimer	Instance	required single	classes={Forward_Pcr_Primer}
HasBeenValidated	Boolean	single	default={false}
LastPositionInterrogatedRange	Integer	required single	
MethodSubmission	Instance	required mult.	classes={Method_Submissions}
PharmGKBInstances	Instance	multiple	classes={THING}
ReversePcrPrimer	Instance	required single	classes={Reverse_Pcr_Primer}
SNDAssaySubmissionOr	Instance	multiple	classes={Data_Submissions}
Stoid	String	single	
ThreePrimeFlankingSequence	String	single	

Ontology Operations V C + -

XmlSchemaElements V C + -

- Get value of DataSubmissionsOf slot
- Comment
- DisplayName
- ExperimentalRegionSubmission
- ForwardPcrPrimer

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

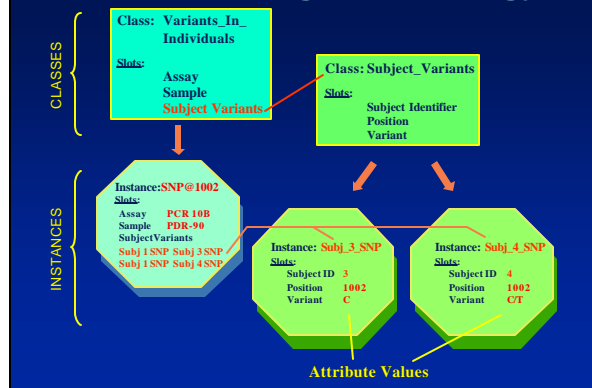
Reference Sequence	Assayed SNP Positions		
	U44106	U44106	U44106
NT Position in GenBank Sequence	1002	1034	1088
"Wild Type" Nucleotide	T	G	T
Variant Nucleotide	C	A	C
Subject 1	C/T	G	T
Subject 2	T	G	T
Subject 3	C	G	C/T
Subject 4	C/T	G	T

Experimental Data in XML

```

<Variants_In_Individuals>
  <DisplayName>SNP@1002</DisplayName>
  <Assay>PCR 10B</Assay>
  <Sample>PDR-90</Sample>
  <Subject_Variants>
    <DisplayName>Subj_3_SNP</DisplayName>
    <SubjectID>3</SubjectID>
    <Position>1002</Position>
    <Variant>C</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>
  
```

Internal Storage in Ontology

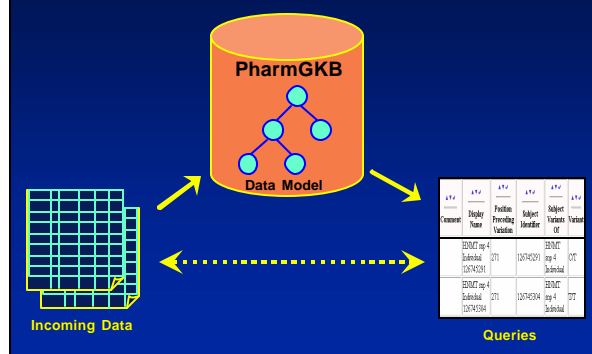


Data in Ontology Viewed in Relational Form

Reference Sequence	Assayed SNP Positions		
	U44106	U44106	U44106
Subject 1	C/T	G	T
Subject 2	T	G	T

Comment	Display Name	Position Preceding Variation	Subject Identifier	Subject Variants Of	Variant
	HNMT snp 4 Individual 126745291	271	126745291	HNMT snp 4 Individual	C/T
	HNMT snp 4 Individual 126745304	271	126745304	HNMT snp 4 Individual	T/T

Result: A Transparent Interface Between Ontology and Data



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/>