Biological Data Management, part 1

H. V. Jagadish
University of Michigan

Acknowledgments

- Adriane Chapman,
- Aaron Elkiss,
- Magesh Jayapandian,
- Bin Liu,
- Arnab Nandi,
- Louiqa Raschid,
- Wing-Kin Sung,
- Glenn Tarcea,
- Limsoon Wong,
- Cong Yu

Outline

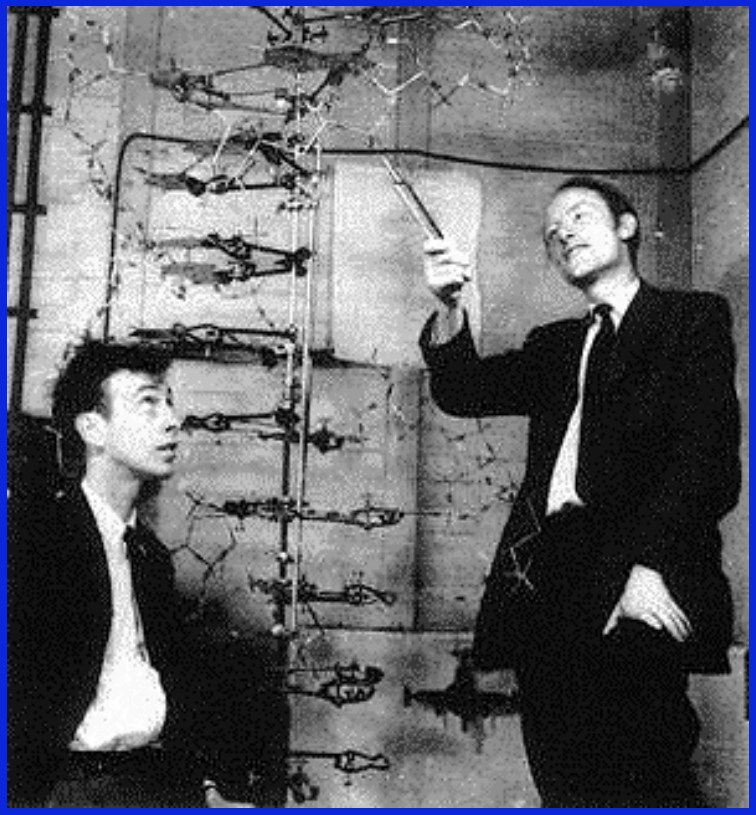
- Introduction to Biology and Bioinformatics
  - Biology 100
  - Major classes of bioinformatics studies
- Case Study of a Biological Data Management System
- Technical Challenges
  - Provenance
  - Ontology
  - Usability

Cell

- A *cell* is the basic unit of life
- Cells perform two types of function
  - Chemical reactions needed to maintain our life
  - Pass info for maintaining life to next generation
- In particular
  - Protein performs chemical reactions
  - DNA stores & passes info
  - RNA is intermediate between DNA & proteins

DNA

- Stores instructions needed by the cell to perform daily life function
- Consists of two strands interwoven together to form a double helix
- Each strand is a chain of some small molecules called nucleotides



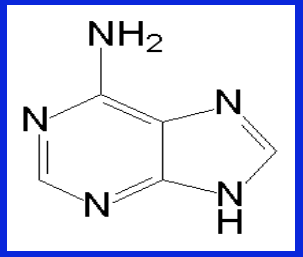
Francis Crick shows James Watson the model of DNA in their room number 103 of the Austin Wing at the Cavendish Laboratories, Cambridge

Classification of Nucleotides

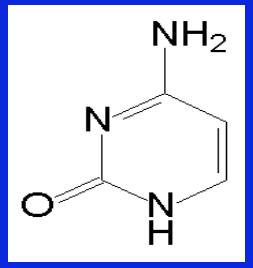
- 5 different nucleotides: adenine(*A*), cytosine(*C*), guanine(*G*), thymine(T), & uracil(U)

- *A*, *G* are purines. They have a 2-ring structure
- *C*, T, U are pyrimidines. They have a 1-ring structure

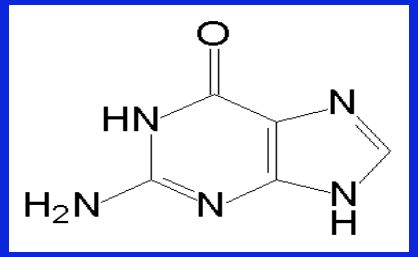
- DNA only uses *A*, *C*, *G*, & T



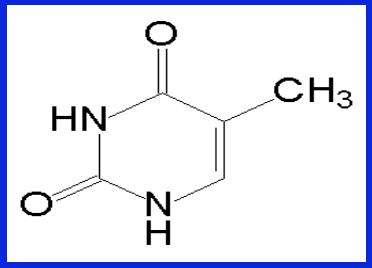
A



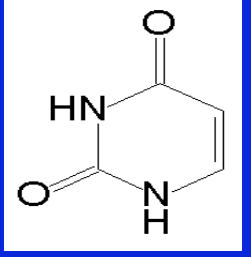
C



G



T



U

Chromosome

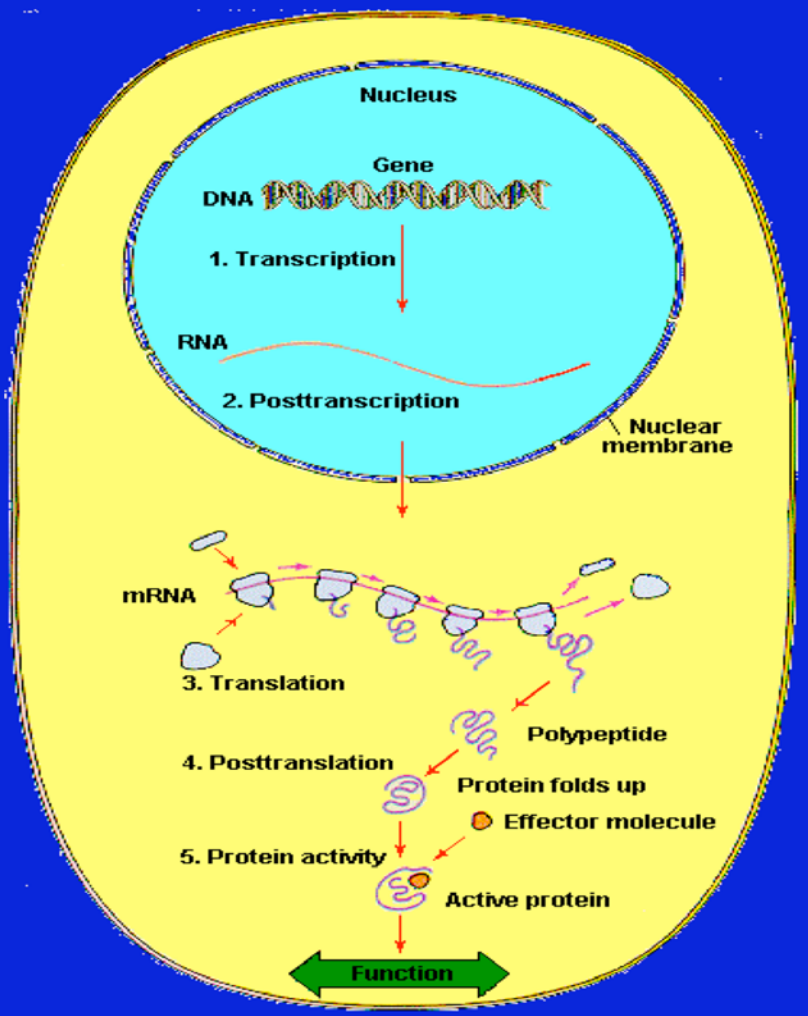
- A chromosome is a molecular unit of DNA
- The genome is the complete set of genetic information in all chrosmosomes
- In most multi-cell organisms, every cell contains the same complete genome
- Human genome has 3 giga bases, organized in 23 pairs of chromosomes

Gene

- A gene is a sequence of DNA that encodes a protein or an RNA molecule
  - Notice vagueness in definition
  - Scientists often disagree on what exactly comprises a gene
- About 30,000 – 35,000 (protein-coding) genes in human genome
- Most genes encode for one protein

Central Dogma

- *A gene is expressed* when it is directing protein production
- *Transcription* of DNA to mRNA is the first step in expression
- *Translation* of mRNA into protein is net major step.



Genetic Code

- Start codon:

  ATG (code for M)

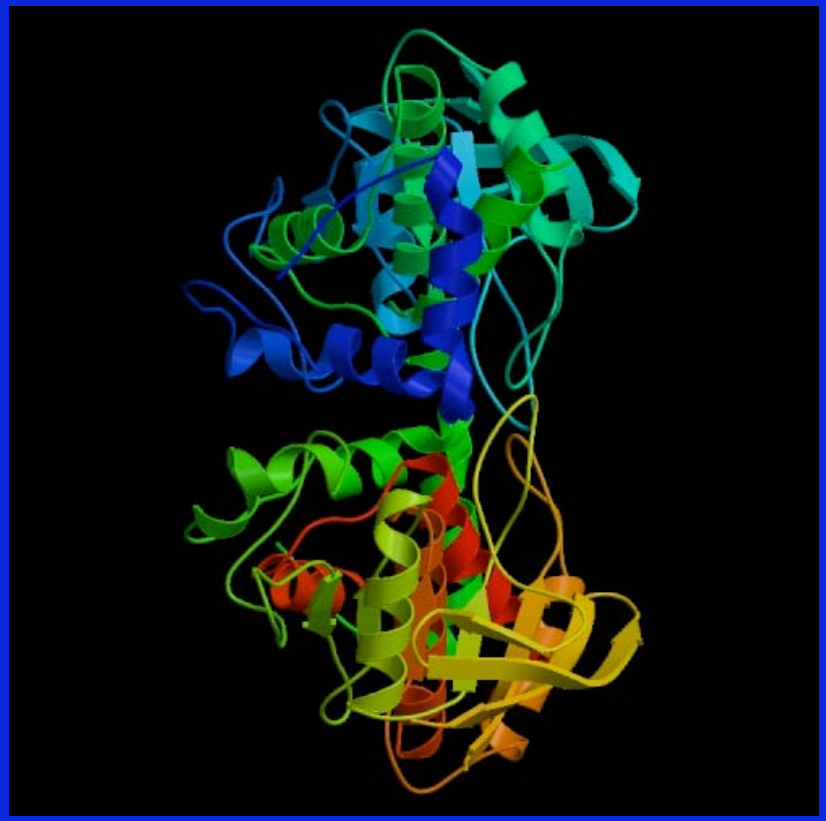
- Stop codon:

  TAA, TAG, TGA

First Position	Second Position of Codon: T	Second Position of Codon: C	Second Position of Codon: A	Second Position of Codon: G	Third Position
T	TTT Phe [F]	TCT Ser [S]	TAT Tyr [Y]	TGT Cys [C]	T
	TTC Phe [F]	TCC Ser [S]	TAC Tyr [Y]	TGC Cys [C]	C
	TTA Leu [L]	TCA Ser [S]	TAA *Ter* [end]	TGA *Ter* [end]	A
	TTG Leu [L]	TCG Ser [S]	TAG *Ter* [end]	TGG Trp [W]	G
C	CTT Leu [L]	CCT Pro [P]	CAT His [H]	CGT Arg [R]	T
	CTC Leu [L]	CCC Pro [P]	CAC His [H]	CGC Arg [R]	C
	CTA Leu [L]	CCA Pro [P]	CAA Gln [Q]	CGA Arg [R]	A
	CTG Leu [L]	CCG Pro [P]	CAG Gln [Q]	CGG Arg [R]	G
A	ATT Ile [I]	ACT Thr [T]	AAT Asn [N]	AGT Ser [S]	T
	ATC Ile [I]	ACC Thr [T]	AAC Asn [N]	AGC Ser [S]	C
	ATA Ile [I]	ACA Thr [T]	AAA Lys [K]	AGA Arg [R]	A
	ATG Met [M]	ACG Thr [T]	AAG Lys [K]	AGG Arg [R]	G
G	GTT Val [V]	GCT Ala [A]	GAT Asp [D]	GGT Gly [G]	T
	GTC Val [V]	GCC Ala [A]	GAC Asp [D]	GGC Gly [G]	C
	GTA Val [V]	GCA Ala [A]	GAA Glu [E]	GGA Gly [G]	A
	GTG Val [V]	GCG Ala [A]	GAG Glu [E]	GGG Gly [G]	G

Protein

- A sequence composed from an alphabet of 20 amino acids
  - Length is usually 20 to 5000 amino acids
  - Average around 350 amino acids
- Folds into 3D shape, forming the building block & performing most of the chemical reactions within a cell



Outline

- Introduction to Biology and Bioinformatics
  - Biology 100
  - Major classes of bioinformatics studies
    - Sequence alignment
    - Gene expression microarrays
    - Mass Spectrometry
- Case Study of a Biological Data Management System
- Technical Challenges

Motivations for Sequence Comparison

- DNA is blue print for living organisms
- Evolution is related to changes in DNA
- By comparing DNA sequences we can infer evolutionary relationships between the sequences w/o knowledge of the evolutionary events themselves
- Foundation for inferring function, active site, and key mutations

Guess function for a new protein T

Compare *T* with seqs of known function in a db

Poor Sequence Alignment

- Poor seq alignment shows few matched positions

⇒ The two proteins are not likely to be homologous

Alignment by FASTA of the sequences of amicyanin and domain 1 of ascorbate oxidase

```
                    60        70        80        90       100
Amicyanin          MPHNVHFVAGVLGEAALKGPMMKKEQAYSLTFTEAGTYDYHCTPHPFMRGKVVV
                                          :..:  . ::. ::
Ascorbate Oxidase ILQRGTPWADGTASISQCAINPGETFFYNFTVDNPGTFFYHGHLGMQRSAGLYG
                     70        80        90       100       110
```

No obvious match between Amicyanin and Ascorbate Oxidase

Discard this function as a candidate

Good Sequence Alignment

- Good alignment usually has clusters of extensive matched positions

⇒ The two proteins are likely to be homologous

```
>gi|134476732|ref|NP_108301.1|  unknown protein [Mesorhizobium loti]
gi|14027493|dbj|BAB53762.1|  unknown protein [Mesorhizobium loti]
          Length = 105

 Score =  105 bits (262), Expect = 1e-22
 Identities = 61/106 (57%), Positives = 73/106 (68%), Gaps = 1/106 (0%)

Query: 1   MKPGRLASIALAIIFLPMAVPAHAATIEITMENLVISPTEVSAKVGDTIRWVNKDVFAHT 60
           MK G L  ++        MA PA AATIE+T++ LV SP  V AKVGDTI WVN DV AHT
Sbjct: 1   MKAGALIRLSWLAALALMAAPAAAATIEVTIDKLVFSPATVEAKVGDTIEWVNNDVVAHT 60
```

good match between Amicyanin and u[illegible]wn M. loti protein

Assign to *T* same function as homologs

Confirm with suitable wet experiments

Phylogenetic Tree/Network

- Phylogenetic tree is a tree whose leaves are labeled by some species
- Represented by a rooted tree, distinctly leaf-labeled
- Phylogenetic network, with DAG structure is more realistic

Outline

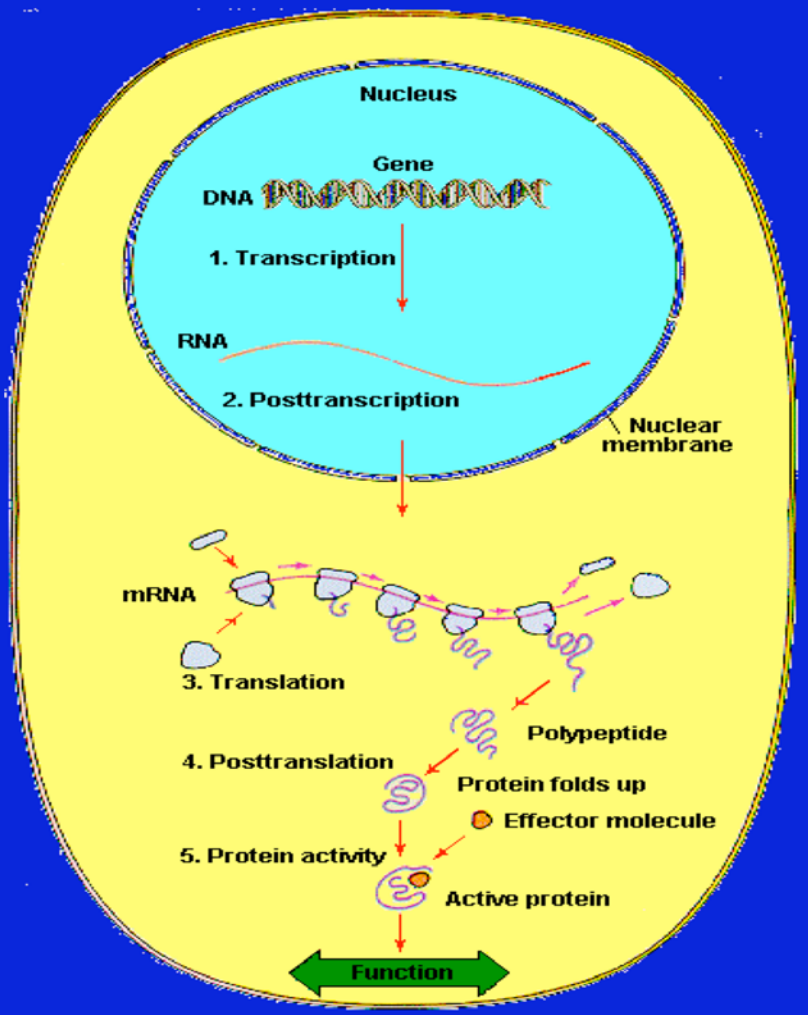
- Introduction to Biology and Bioinformatics
  - Biology 100
  - Major classes of bioinformatics studies
    - Sequence alignment
    - Gene expression microarrays
    - Mass Spectrometry
- Case Study of a Biological Data Management System
- Technical Challenges

Microarrays

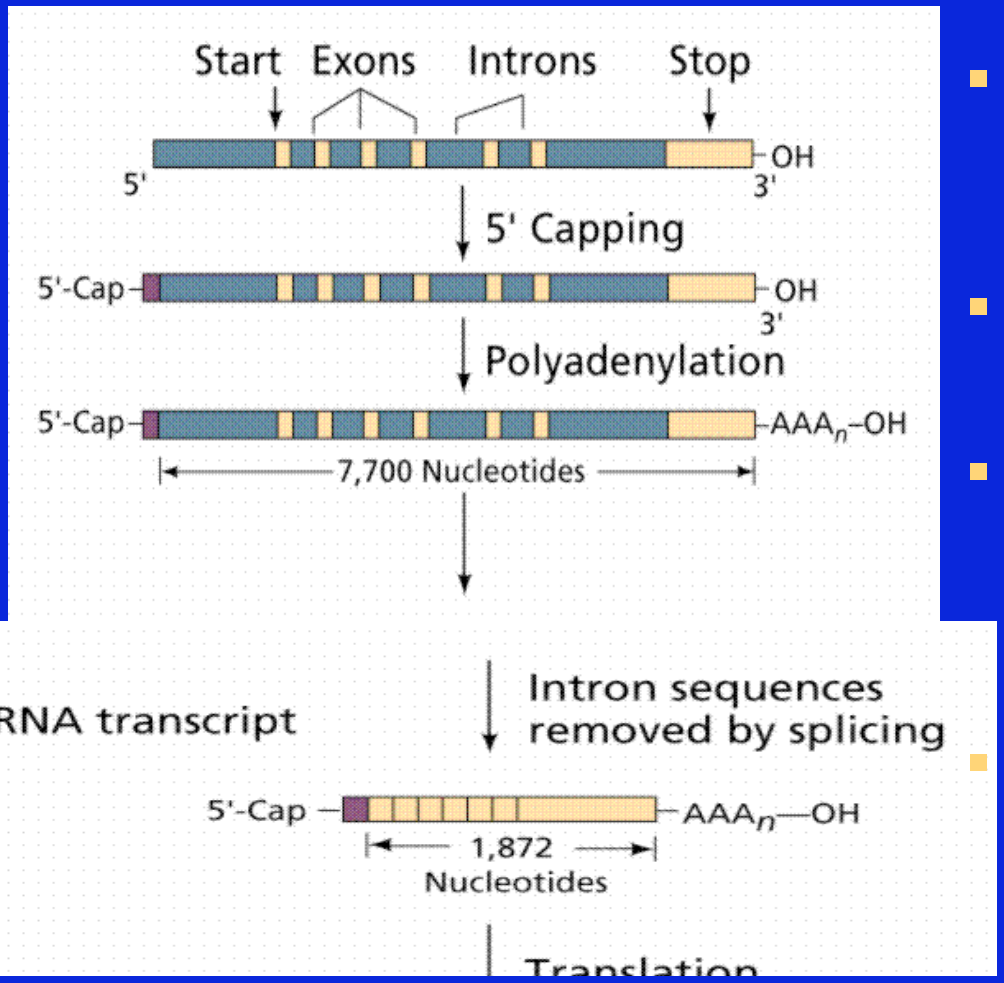
- An assay with a large number of probes for molecular phenomena of interest tethered to specific locations.
- Many uses of microarrays, depending on the probes:
  - Gene expression (most frequent)
  - Genotypes (SNPs)
  - Tissues (few antibodies on many tissues)
  - Protein (antibodies to many proteins)
  - Small molecules (for binding affinity to target)

Quick review of gene expression

- *A gene is expressed* when it is directing protein production
- *Transcription* of DNA to mRNA is the first step in expression
- By measuring the products of transcription, we can assay gene expression



A more nuanced view



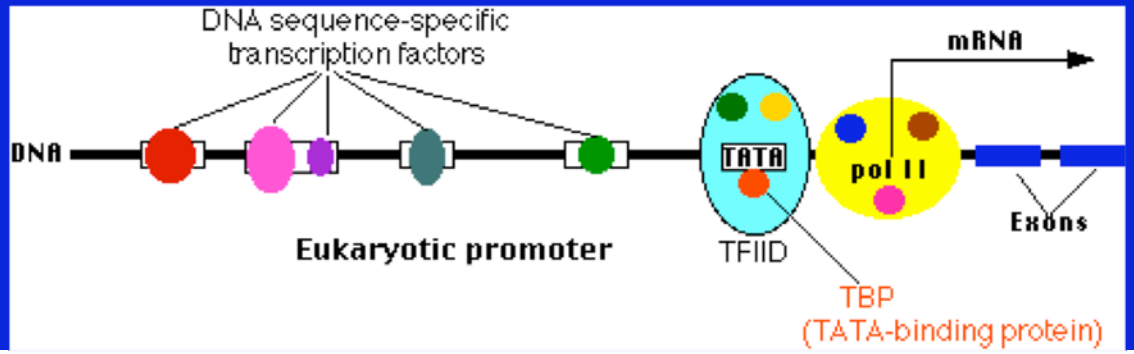
- Genes are expressed at varying levels (not just on/off)
- mRNA isn't just copied, but processed
- Mature mRNA has
  - Introns removed
  - PolyA tail, 5' cap
- Alternative splicings

Expression is central because...

- Differentiation: All cells in a body have the same genome. Expression is what differentiates, e.g. brain cells from liver.
- Physiology: Cells do their business (dividing, sending signals, digesting, etc.) largely via changes in expression
- Response to stimuli: Environmental changes (like drugs or disease) often cause changes in expression
- Disease markers and drug targets: changes in expression associated with disease can be diagnostic markers and/or suggest novel pharmaceutical approaches.

Control of expression

- Which genes are expressed and at what levels is under molecular control
- Proteins that influence gene expression are *transcription factors*.
- Non-coding regions contain transcription factor binding sites



Array technology

- **Basic idea: mRNA hybridizes best to exactly complementary sequences.**
- **Method:**
  - Probes are attached to a substrate in a known location
  - mRNA in one or more samples are fluorescently labeled
  - samples are hybridized to probe array, excess is washed off, and fluorescence reading are taken for each position
- **Two major classes:**
  - "custom" spotted arrays (probes printed on slides)
  - "Affymetrix" probes built up on silicon by photolithography

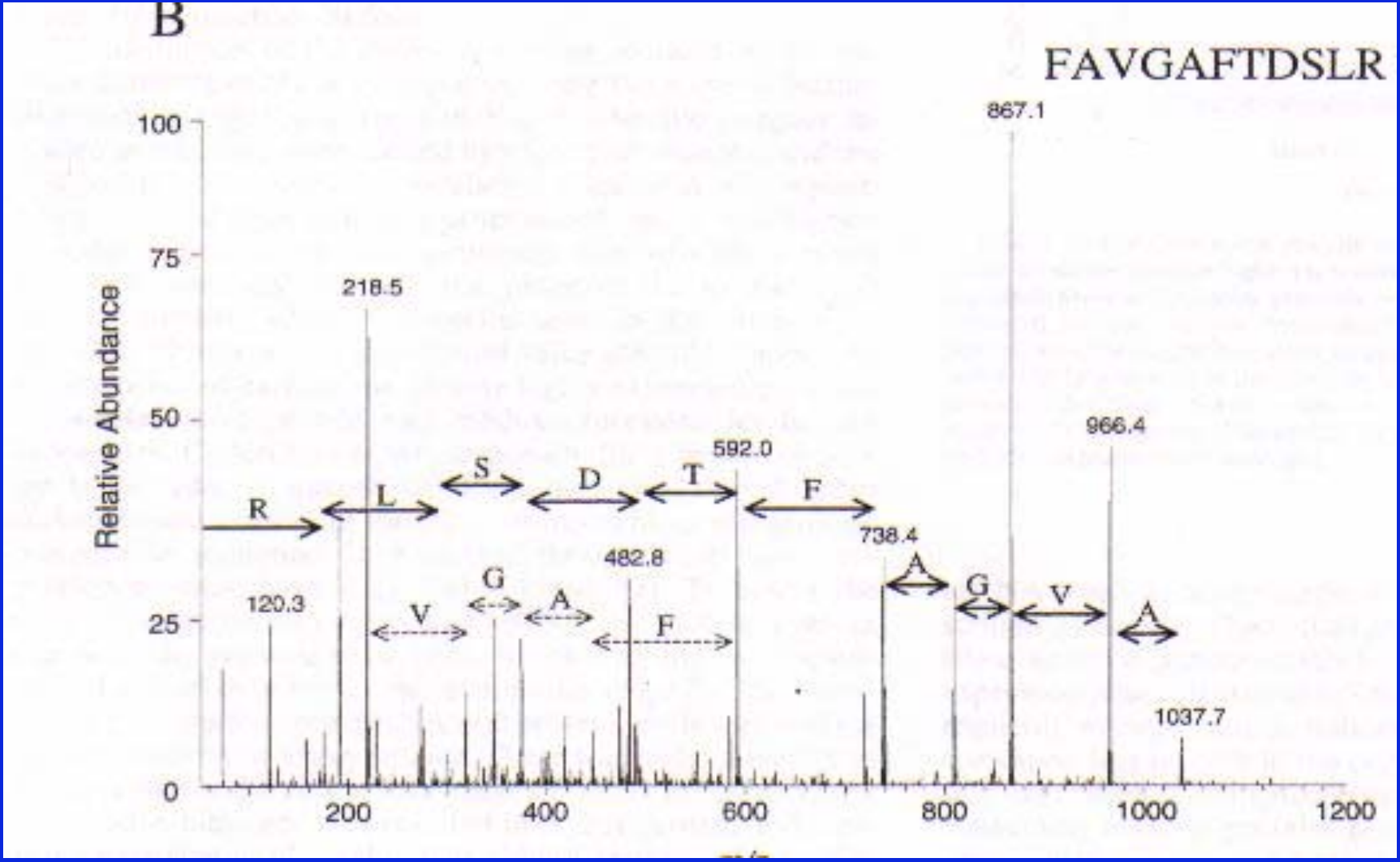
Outline

- Introduction to Biology and Bioinformatics
  - Biology 100
  - Major classes of bioinformatics studies
    - Sequence alignment
    - Gene expression microarrays
    - Mass Spectrometry
- Case Study of a Biological Data Management System
- Technical Challenges

Peptide Sequencing

- Unlike DNA, deducing the amino acid sequence of a protein peptide is not easy
- The problem of finding the amino acid sequence of a protein peptide is known as the Peptide Sequencing Problem
- One solution is to use mass spectrometry

An Example MS/MS Spectrum



Two Ways for Identifying the Amino Acid Sequence

- Given the spectrum M, there are two ways to identify the amino acid sequence
  - De Novo sequencing
    - Among all possible peptides, find a peptide which is best explaining the spectrum M
  - Database searching
    - Select a peptide from the database which is best explaining the spectrum M

Outline

- Introduction to Biology and Bioinformatics
- Case Study of a Biological Data Management System: Integrating Information on Protein Interactions
  - Overview of information integration
  - Specific challenges with protein interaction
  - Details of MiMI system
- Technical Challenges

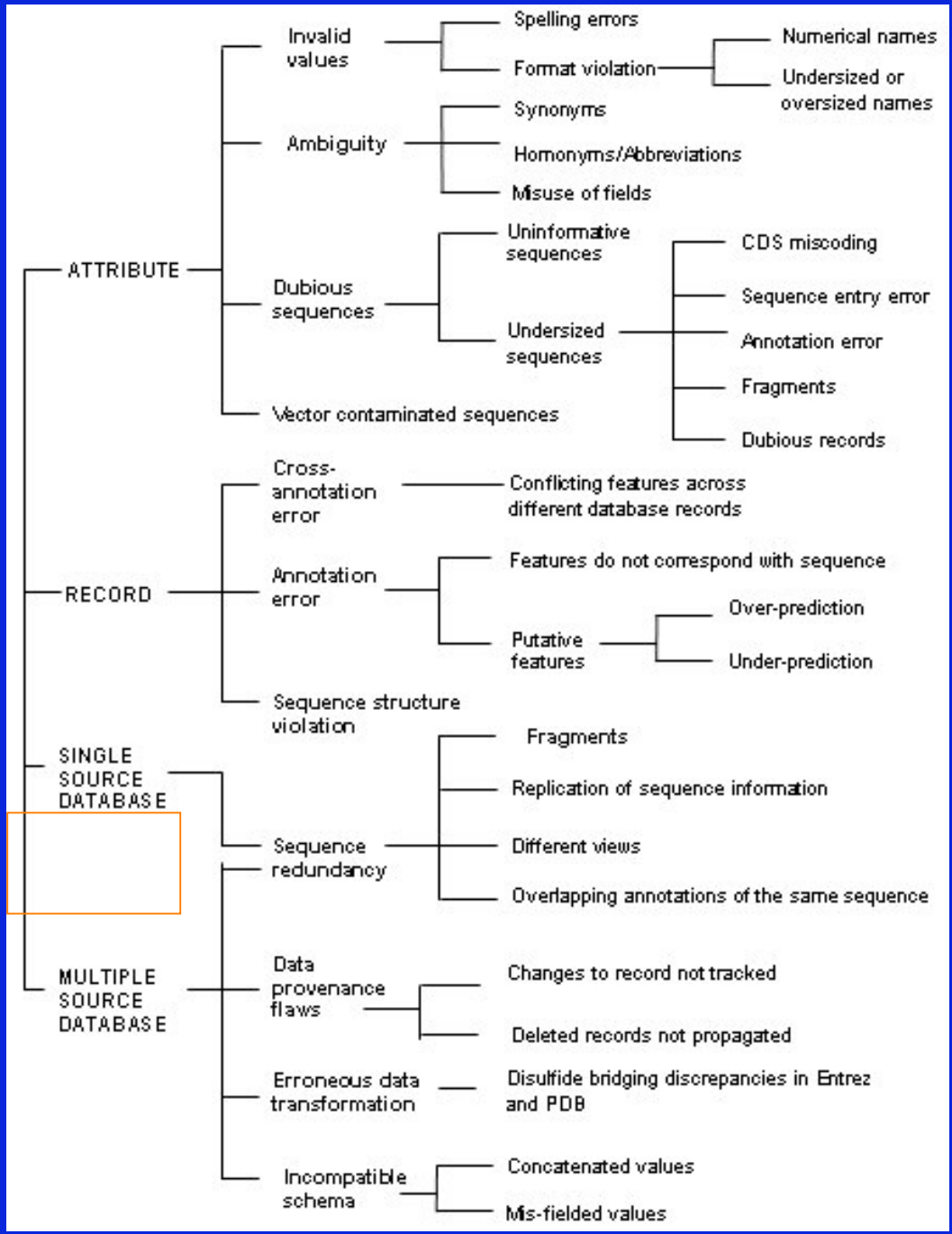
MiMI Motivation

- Copious amounts of protein data exist online
- Some of it is repeated across sources, some of it is contradictory between sources
- Experiments used to furnish data have varying levels of false positive and negatives
- Researchers must get pieces from disparate sources and piece them together manually, making judgments about the quality of each source as they work.

Some Common Sources of Error

- Diverse sources of data
  - **Repeated submissions of sequences to databases**
  - **Cross-updating of databases**
- Data Annotation
  - **Databases have different ways to annotate data**
  - **Different interpretations**
- Lack of standardized nomenclature

A Classification of Errors



Outline

- Introduction to Biology and Bioinformatics
- Case Study of a Biological Data Management System: Integrating Information on Protein Interactions
  - Overview of information integration
  - Specific challenges with protein interaction
  - Details of MiMI system
- Technical Challenges

Currently

```
<node uid="DIP:5601N" id="5564" name="RL23_YEAST“
        class="protein">
 <feature name="SWP:P04451" class="cref">
  <src>SwissProt</src>
  <val>RL23_YEAST</val>
 </feature>
 <feature name="PIR:R5BY17" class="cref">
  <src>PIR</src>
 </feature>
 <feature name="GI:603356" class="cref">
  <src>NCBI</src>
 </feature>
 <att name="taxon">
  <val>4932</val>
 </att>
 <att name="kwds">
  <val>protein biosynthesis; ribosome</val>
 </att>
 <att name="descr">
  <val>ribosomal protein L23.e, cytosolic</val>
 </att>
 <att name="organism">
  <val>Saccharomyces cerevisiae (budding yeast)</val>
 </att>
</node>
```

```
<node uid="DIP:6527N" id="6474"
     name="RL23_YEAST" class="protein">
 <feature name="SWP:P04451" class="cref">
  <src>SwissProt</src>
  <val>RL23_YEAST</val>
 </feature>
 <feature name="GI:603356" class="cref">
  <src>NCBI</src>
 </feature>
 <att name="taxon">
  <val>4932</val>
 </att>
 <att name="kwds">
  <val>Multigene family; Ribosomal protein</val>
 </att>
 <att name="descr">
  <val>60S ribosomal protein L23 (L17)</val>
 </att>
 <att name="organism">
  <val>Saccharomyces cerevisiae (budding yeast)</val>
 </att>
</node>
```

Overlapping data records for RL23_YEAST from DIP

DIP 1

```
<node uid="DIP:5601N" id="5564"
      name="RL23_YEAST"
       class="protein">
 <feature name="SWP:P04451"
      class="cref">
  <src>SwissProt</src>
  <val>RL23_YEAST</val>
 </feature>
 <feature name="PIR:R5BY17"
      class="cref">
  <src>PIR</src>
 </feature>
 <feature name="GI:603356"
      class="cref">
  <src>NCBI</src>
 </feature>
 <att name="taxon">
  <val>4932</val>
 </att>
 <att name="kwds">
  <val>protein biosynthesis;
      ribosome</val>
 </att>
 <att name="descr">
  <val>ribosomal protein L23.e,
      cytosolic</val>
 </att>
 <att name="organism">
  <val>Saccharomyces cerevisiae
      (budding yeast)</val>
 </att>
</node>
```

DIP 2

```
<node uid="DIP:6527N" id="6474“
      name="RL23_YEAST"
class="protein">
  <feature name="SWP:P04451"
class="cref">
   <src>SwissProt</src>
   <val>RL23_YEAST</val>
  </feature>
  <feature name="GI:603356"
class="cref">
   <src>NCBI</src>
  </feature>
  <att name="taxon">
   <val>4932</val>
  </att>
  <att name="kwds">
   <val>Multigene family;
Ribosomal protein</val>
  </att>
  <att name="descr">
   <val>60S ribosomal protein L23
(L17)</val>
  </att>
  <att name="organism">
   <val>Saccharomyces cerevisiae
(budding yeast)</val>
  </att>
 </node>
```

And worse…

Swiss Prot

Entry name **RL23_YEAST**
Accession number **P04451**
Description 60S ribosomal protein L23 (L17).
Gene name(s) (RPL23A OR RPL17A OR YBL087C OR YBL0713)
AND (RPL23B OR RPL17B OR YER117W).
Organism source Saccharomyces cerevisiae (Baker's yeast).
NCBI TaxID 4932
Length: **137 aa**, molecular weight: **14473 Da**,

MSGNGAQGTK FRISLGLPVG AIMNCADNSG
ARNLYIIAVK GSGSRLNRLP AASLGDMVMA
TVKKGKPELR KKVMPAIVVR QAKSWRRRDG
VFLYFEDNAG VIANPKGEMK GSAITGPVGK
ECADLWPRVA SNSGVVV

NCBI

```
LOCUS       AAC03215               137 aa
linear   PLN 05-APR-2002
DEFINITION  Rpl17bp: Ribosomal protein, large
subunit
ACCESSION   AAC03215
VERSION     AAC03215.1  GI:603356
SOURCE      Saccharomyces cerevisiae (baker's
yeast)
FEATURES    product="Rpl17bp: Ribosomal protein,
large subunit"
            gene="RPL17B“
ORIGIN
1 msgngaqgtk frislglpvg aimncadnsg
   arnlyiiavk gsgsrlnrlp aaslgdmvma
61 tvkkgkpelr kkvmpaivvr qakswrrrdg
```

Need Deep Integration

- User desires a comprehensive answer to a query, rather than a confusing mishmash of information from multiple sources.
- So, fuse information from multiple sources.
- Desirable for human users, but essential for machine "users".
  - E.g. when further analysis is to be performed on the results of the query.

Context

- Users often require context for interaction data that they see:
  - type of experiment used,
  - the organism,
  - the tissue,
  - disease state
- And also information that may really be associated with one of the interactors:
  - Cellular location
  - Putative function
- These are often obtained from additional sources, compounding the data integration problem.

A Deeply Integrated Example

```
<object>
  <object-names>
      <name>RL23_YEAST</name>
      <name>Rpl23ap</name>
      <name>Rpl17bp</name>
  </object-names>
<object-descr>
  <Val>
     60S ribosomal protein L23 (L17)
  </Val>
   <Val>
     ribosomal protein L23.e, cytosolic
   </Val>
 </object-descr>
```

```
<object-ids>
 <object-ext-id>
  <database> <DIP/> </database>
   <Dist>
      <Val prob="0.5">DIP:6527N </Val>
      <Val prob="0.5">DIP:5601N </Val>
   </Dist>
 </object-ext-id>
</object-ids>
<seq>
  msgngaqgtk frislglpvg aimncadnsg
  arnlyiiavk gsgsrlnrlp aaslgdmvma
  tvkkgkpelr kkvmpaivvr qakswrrrdg
  vflyfednag vianpkgemk gsaitgpvgk
  ecadlwprva snsgvvv
<seq>
</object>
```

But …

- Deep Integration means the original source data has been transformed, and is not directly available.
- This is an issue because some times the transformation may have made an error. At other times, we may have conflicting information in different sources, and our resolution between these may be wrong.
- This is also of concern because a scientist often wishes to dig deep into a specific pathway, interaction, or protein. Needs a clear path to be able to do this.
- Address, by maintaining:
  - Provenance
  - Probability

Provenance

- For each unit of data, keep track of where it came from.

- What is a unit of data?
  - Individual attributes and element content.
- How far back do record where it came from?
  - 1-2 steps – practical decision
  - Data source used, and possibly publication from which data source got its data.

Provenance Storage

- The nearest ancestor contains the provenance for a node
- Any node can over-ride the general provenance by stipulating its own
- Factor out provenance detail so that only provenance ids need be stored repeatedly

```
<object>
  <provenance>
    <provid> 1 </provid>
    <provid> 2  </provid>
  </provenance>
  <name> RL23_YEAST
    <provenance>
      <provid> 1 </provid>
    </provenance>
  </name>
  <descr> 60S ribosomal protein
          L23 (L17) </descr>
```

Conflicting Values in Integration

Probability Annotation

- Annotate how likely a piece of data is.
- Even for a single source, we have issues:
  - Experiments are flawed
  - Computational methods are error prone
  - Hand annotation is never exact
- Even more errors with multiple sources, due to errors data fusion and object identification.
- When values from multiple sources conflict, record all, and assign probabilities

Probability Model

- Probability associated with an element (or attribute) is the probability of its value being correct with respect to its parent.
  - For non-leaf element, needs some care to define properly.
- Absolute probability is obtained by multiplying a sequence of conditional probabilities up to the root.
- See paper in VLDB 2002.

A Deeply Integrated Example with Provenance and Probability

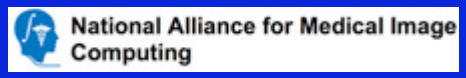
```
<object>
  <provenance>
    <provid>2</provid>
  </provenance>
  <object-names>
    <name>RL23_YEAST</name>
    <name>Rpl23ap</name>
    <name>Rpl23bp</name>
  </object-names>
 <object-descr>
    <Dist>
      <Val prob="0.5">
        60S ribosomal protein L23 (L17) </Val>
      <Val prob="0.5">
        ribosomal protein L23.e, cytosolic </Val>
    </Dist>
  </object-descr>
```

```
<object-ext-id>
      <database> <DIP/> </database>
      <ext-db-id>
        <Dist>
          <Val prob="0.5">DIP:6527N </Val>
          <Val prob="0.5">DIP:5601N </Val>
        </Dist>
      </ext-db-id>
    </object-ext-id>
  </object-ids>
<seq>
  msgngaqgtk frislglpvg aimncadnsg
  arnlyiiavk gsgsrlnrlp aaslgdmvma
  tvkkgkpelr kkvmpaivvr qakswrrrdg
  vflyfednag vianpkgemk gsaitgpvgk
  ecadlwprva snsgvvv
<seq>
    <provenance>
      <provid> 3 </provid>
      <provid> 4 </provid>
    </provenance>
  </object>
```

Outline

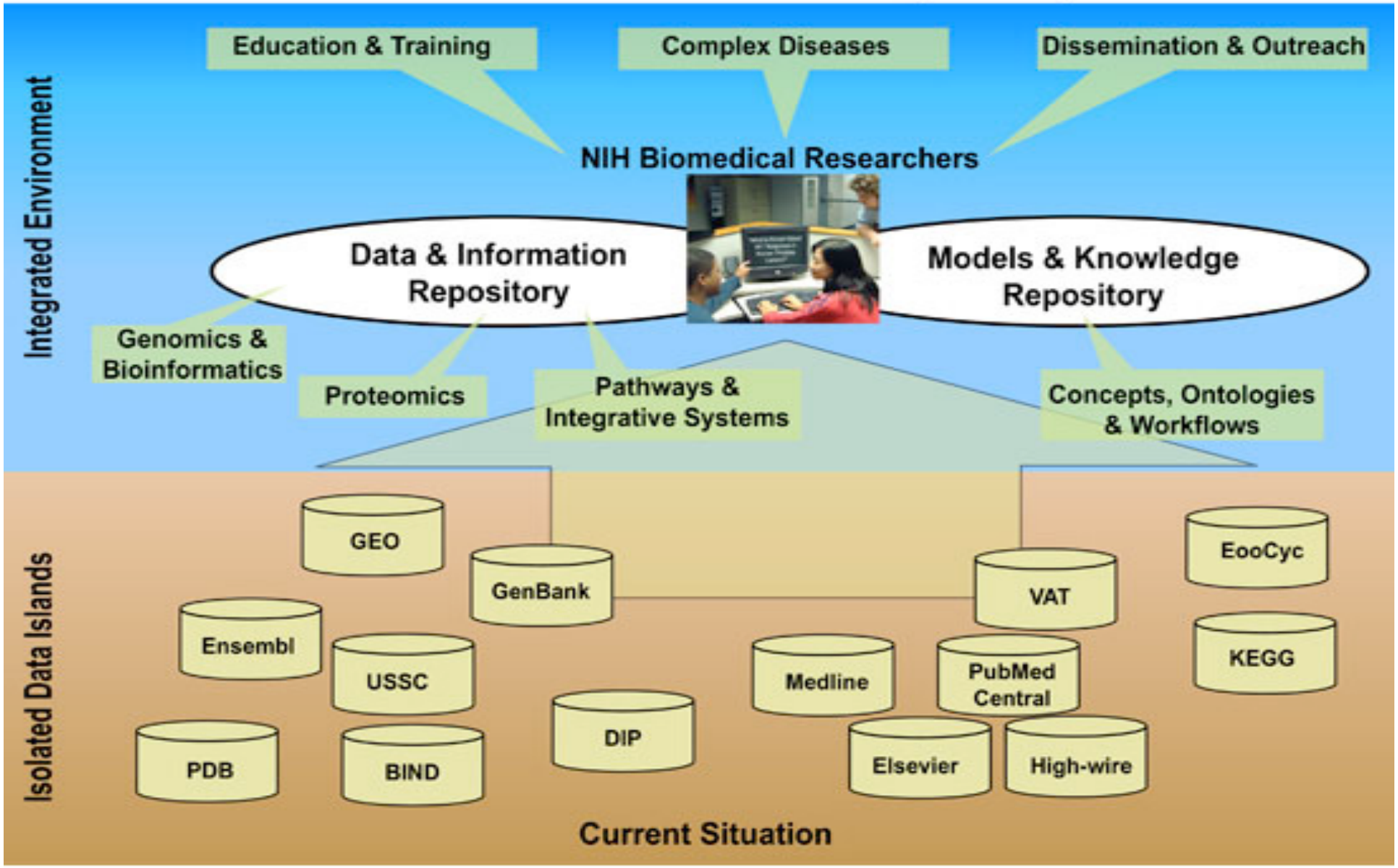
- Introduction to Biology and Bioinformatics
- Case Study of a Biological Data Management System: Integrating Information on Protein Interactions
  - Overview of information integration
  - Specific challenges with protein interaction
  - Details of MiMI system
  - M. Jayapandian, A P Chapman, et al, "Michigan Molecular Interactions (MiMI): Putting the Jigsaw Puzzle Together,'' *Nucleic Acids Research*, vol 35, D566-D571, Jan. 2007.
- Technical Challenges

National Centers for Biomedical Computation



NCIBI

Vision of the NIH National Center for Integrative Biomedical Informatics (NCIBI)



The modern biomedical scientist needs access to a rich information environment.

Michigan Molecular Interactions

- http://mimi.ncibi.org

Combines:
IntAct HPRD DIP
GO BIND BioGRID
CCSB-HI1 InterPro
IPI MDC PPI
Organelle DB
OrthoMCL Pfam
ProtoNet

Molecules:	107,884
Interactions:	246,559

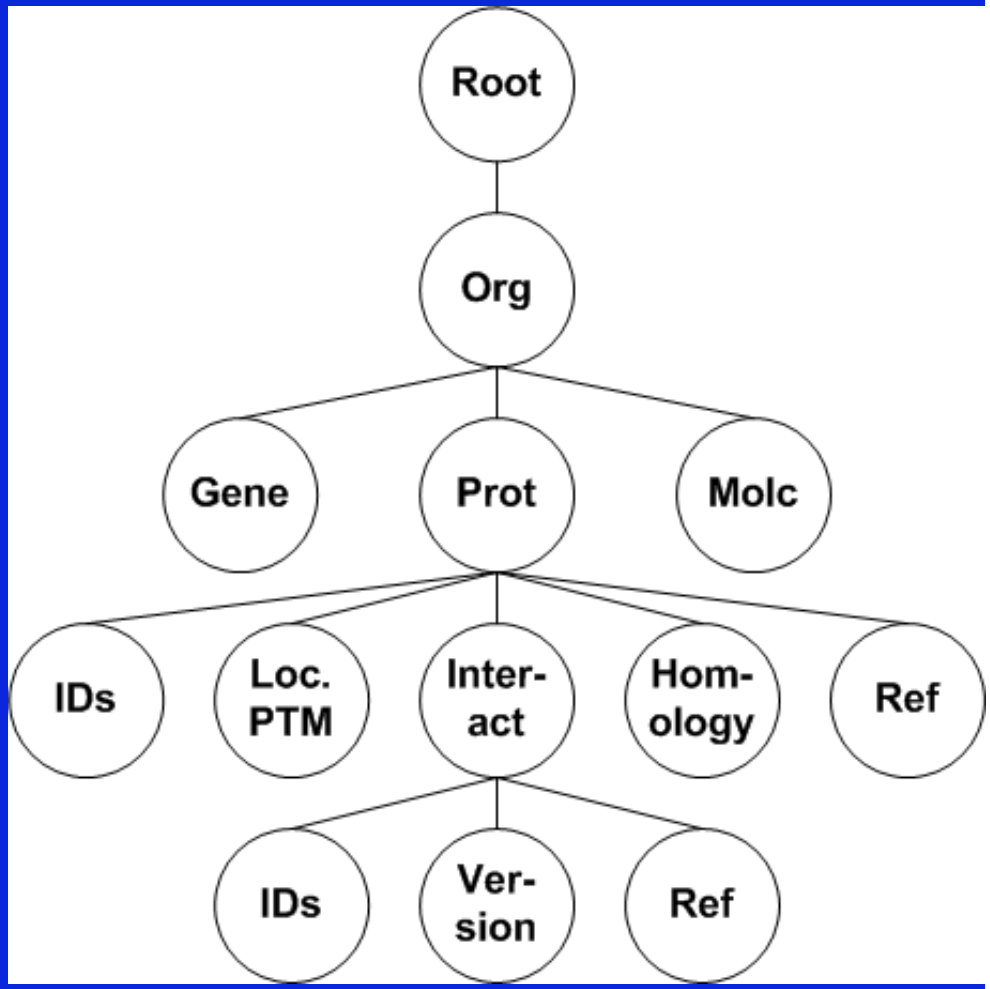
MiMI Schema

- Centered around "molecule" and "interaction"
- Molecule:
  - Identification: Internally generated ID, External references to other databases, name(s)
  - Basic attributes: type, sequence, structure, description
  - GO properties: cell locations, functions, processes
  - Homology: family, method of determination
- Interaction:
  - Participating molecules
  - Experimental system: two hybrid, etc.
  - Molecular conditions: PTM, etc.
  - Cell location, domain, residues
  - Supporting publications

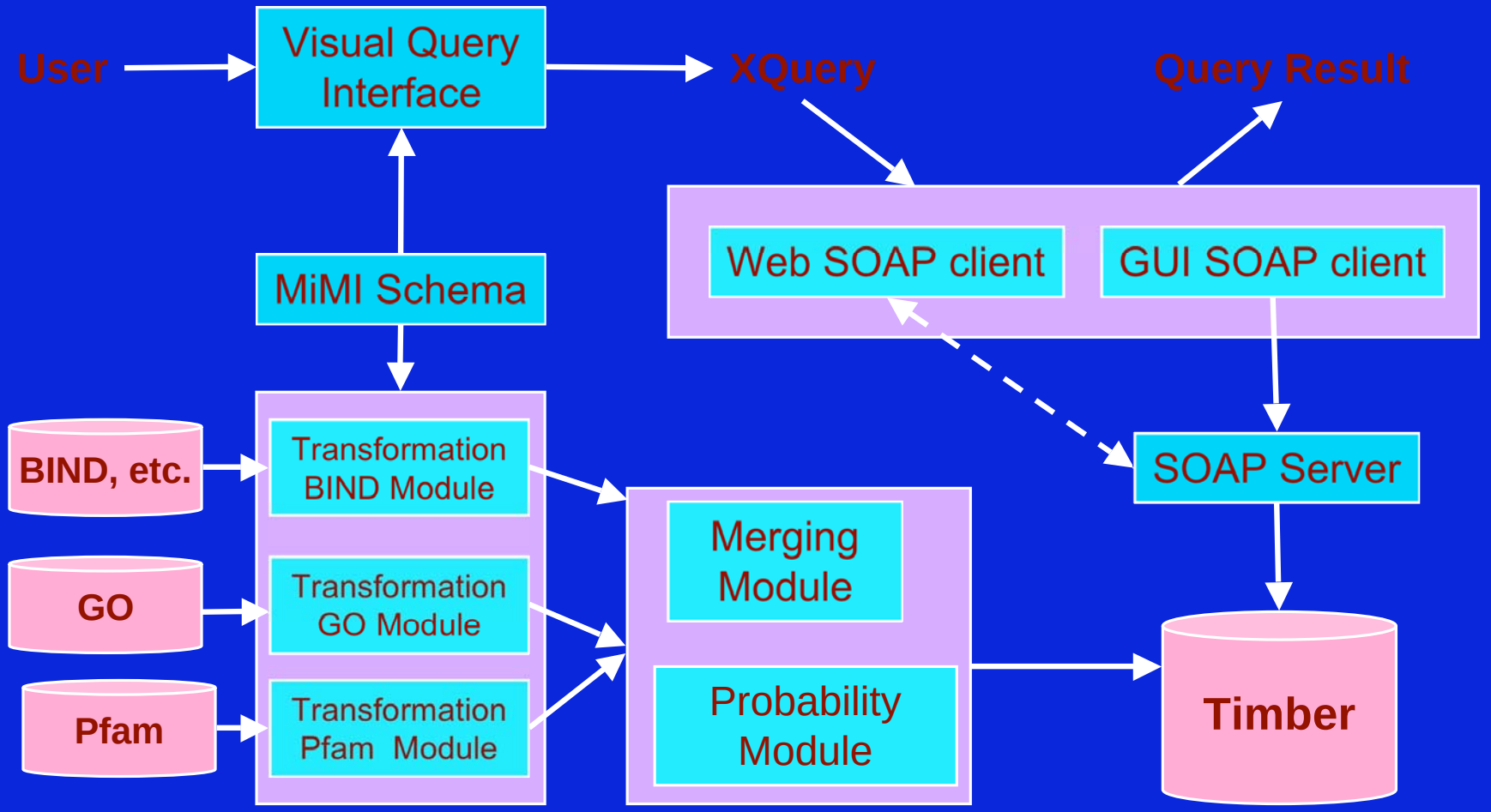
MiMI Data Model

- Provenance is always recorded
- Allows conflicting data to be represented
- Applies a probability field to attributes
- Usable via TIMBER
- Allows usage of XQuery

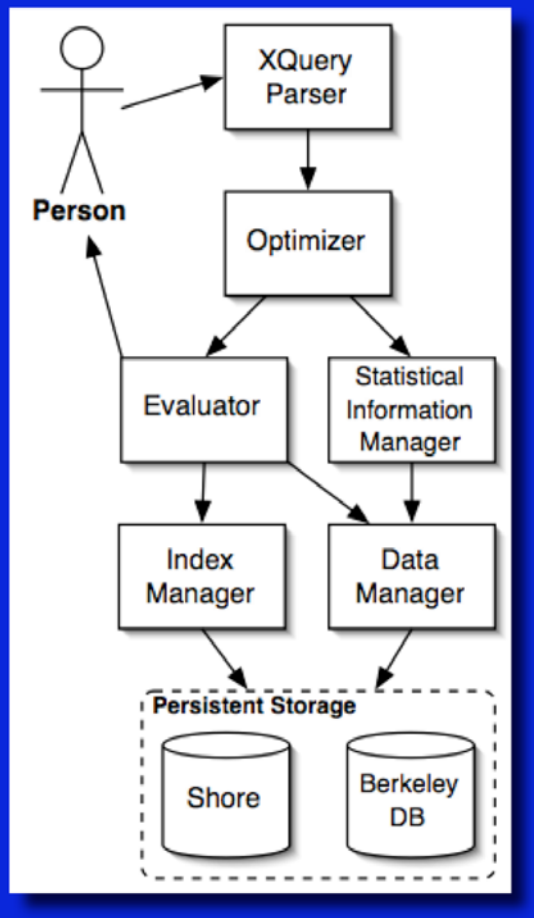
Simplified view of MiMI Schema



System Architecture



Timber



- Native XML database system
- Architecture is similar to a relational database system
- Underlying techniques differ
  - Queries are based on trees
  - XQuery/XPath
  - Basic storage unit is a node
  - Determining ancestor and descendant relationships is an efficient operation

Integration Mechanism

- Data compilation & Merging
  - Transformation scripts are written for each input data source
  - Entities are identified based on information within the source, and also other ID maps
  - Similar entities from different databases are juxtaposed (e.g. protein, interaction, etc.)
  - Probabilistic measures are associated with uncertain data

Value Added by MiMI

- Allows a user to obtain a cohesive view of knowledge about a protein
  - Includes deeply integrated data from multiple sources
- Full XQuery support provides the ability to ask complex queries
  - Minimizes the need to write Perl scripts on a database dump
- Provenance annotations to credit original source, and provide users with source information
- Probability annotations to manage unreliability

Some PUMA interactions in Entrez Gene

- Puma interacts with Bcl-2.
- PUMA interacts with Bcl-2.
- PUMA-alpha interacts with Bcl-2.
- Puma interacts with A1.
- Puma interacts with Bcl-XL.
- PUMA interacts with Bcl-xL.
- Bcl-XL interacts with Puma.

PUMA interactions in MiMI

PUMA interacts with Bcl-2.
PubMed:15574335; BIND:193182;
PubMed:11463392; BIND:196458; BIND:196459;
PubMed:15694340; BIND:210088;
PUMA-alpha interacts with Bcl-2.
PubMed:11463392; BIND:196458;
PUMA-beta interacts with Bcl-2.
PubMed:11463392; BIND:196459;
Puma interacts with Bcl-2.;
PubMed:15694340; BIND:210088;

This mol

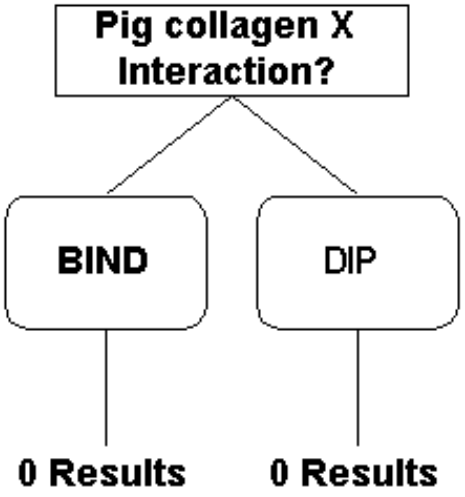
- bfl1_r
- p9728
- bcl2_
- BCLW
- baxa_human (View interaction)
- bclx_human (View interaction)
- BCL2 related protein A1 (View interaction)
- MCL1_HUMAN (View interaction)
- Basonuclin 1 (View interaction)

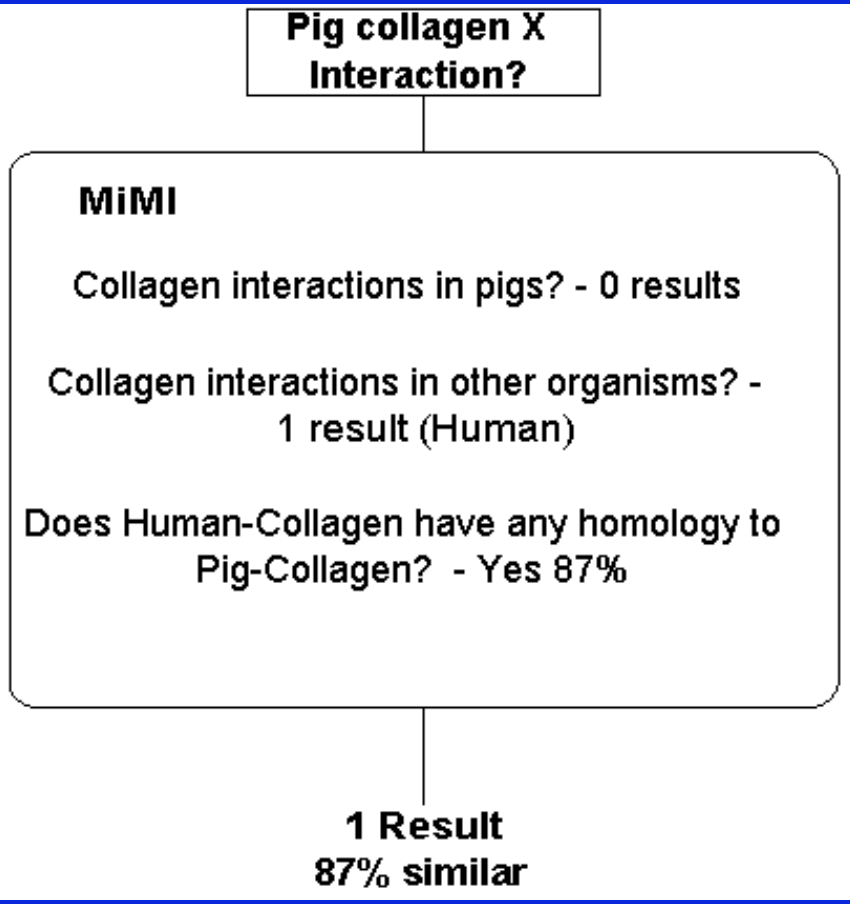
Harder to answer queries

- Query: "I want to know all of the collagen X protein-protein interactions in pig. However, none are currently reported."
- Query: "I want to know all protein-protein interactions in cows, but the dataset is cow-poor."

Putative Interactions

Query: "I want to know all of the collagen X protein-protein interactions in pig. However, none are currently reported."





1 result from MiMI supported by a literature search:

Nielson Vivi, C. Bendixen, J. Arnbjerg, C. Sorensen, H. Jensen, N. Shukri, B. Thomsen. (2000) Abnormal growth plate function in pigs carrying a dominant mutation in type X collagen. *Mammalian Genome*. 11, pg 1087-1092.

Putative Interactions – A Closer Look

Collagen X Interactions in Pigs?

MiMI

Collagen X interactions in other organisms with a known homology with a Pig-collagen X?

```
for $b in document("mimi.xml")//object
where $b/org-ref_taxname = "Sus scrofa"
and $b/name = "collagen X"
return <result>
        {$b/interactions}
        {for $a in document("mimi.xml")//object
        where $a/homolog-with = $b/mimi-id
        return <homolog> {$a/interactions}
                        {$a/homolog-similarity} </homolog>
        }
</result>
```

```
<result>
  <homolog>
    <interaction> Collagen X
    </interaction>
     <homolog-similarity>
      87%
     </homolog-similarity>
  </homolog>
<result>
```

Challenges

- Matching is hard
  - notion of identity
- Merging is hard
  - the current solution with provenance and probability provides a good framework, but not the full solution.
- Usage is hard
  - Good data models should not get in the way of non-technical users
  - MiMI uses a global "hide provenance" button
  - Need effective way to capture and present provenance and probabilities.
  - Graphical tools make this really hard to do.

Matching is Hard

- Match by name
  - Same object known by many very different names – "polysemy"
  - Very different objects may have very similar names – "synonymy"
- Match by sequence
  - Much more robust
  - But not all objects have an associated sequence
  - Often need to do approximate match.
- Match by identifier
  - Standard identifiers often available, e.g. from Entrez
  - Should provide perfect match, but…

Notion of Object is fuzzy

- Orthologs in multiple species
- Polymorphism across individuals in same species
- Post translation modifications, e.g. phosphorylation

Merging is Hard

- Some issues dealt with through provenance and probability.
- Provides a good framework for information-rich merging.
- But values still need to be populated.
  - Much hard work to get reasonable probability values.

- Similar variants, though not identical repeats, may some times not carry much additional information.
- Need to summarize these.

Some values for "Description" Attribute of a Protein

- Tumor protein p73; p53-related protein. This protein shares sequence homology with p53 DNA binding regions. Multiple isofroms arise by multiple promotors and alternative splicing. The identifier listed below corresponds to isoform alpha.
- Tumor protein p73, also called p53-related protein, isoform alpha. OMIM:601990
- Tumor Protein p73; p53-related protein. This protein shares sequence homology to p53 DNA binding regions. OMIM:601990
- Tumor Protein p73, also called p53-related protein, isoform alpha. OMIM:601990.
- Tumor protein p73, p53-related protein.
- tumor protein p73, alpha isoform.

Outline

- Introduction to Biology and Bioinformatics
- Case Study of a Biological Data Management System
- Technical Challenges
  - Provenance
  - Ontology
  - Usability

Provenance

- The origin or source from which something comes
- The history of an item including amendments

- From the Latin *provinir* – to come forth

Provenance – a simple idea?

~~This axe was made in 1861 at the Allegheny US Arsenal.~~

- This axe HEAD was made in 1861 at the Allegheny US Arsenal.
- This axe HAFT was made in 1980 in Mr. Smith's workshop.

Provenance – a simple idea?

- This axe was made in 1861 at the Allegheny US Arsenal.

- The axe was traded by the US army to people of the Seneca tribe.
- The axe haft broke in 1890, and the axe head was discarded.
- The axe head was discovered by Mr. Smith in his backyard in 1978.

Type 1 Diabetes Example

- Find one of the proteins that has different ptms in different situations. Go to different sites, copy information.
- Find error – where did it come from?

Type 1 Diabetes Example

- Use same protein. Copy url information with info.
- Find error – where did it come from?
- Click link – broken link, still no answers

Consensus – There is none

Something Provenance

- actor provenance
- data provenance
- disclosed provenance
- false provenance
- inform provenance
- infrastructure provenance
- input provenance
- interaction provenance
- logical provenance
- logical redo provenance
- process provenance
- observed provenance
- prospective provenance
- redo provenance
- retrospective provenance
- runtime provenance
- stream provenance
- stream-related provenance
- the provenance of interactions
- where provenance
- why provenance
- workflow provenance

Usage of '

Luc Moreau. *IPAW 2006,* Chicago.

Two High-level Views

- "Where" provenance
  - Where did the information for Keap1 come from?
    - -> NCBI
    - -> HPRD
  - Answers:
    - Origin
    - Modifications
- "Why" provenance
  - Why is Keap1 in MiMI?
    - -> It satisfied the query:
      select * from HPRD
  - Answers:
    - What query and underlying dataset generated this field

Prov. for Biologists - Where

- Want to know:
  - Where originated
  - How modified
  - Reproduce results
  - Execute workflows using same setup
  - View previous incarnations of data

Where provenance options

- Provenance Tracking Systems
  - Chimera
  - myGrid
  - MiMI
  - CMCS
- Provenance embedded in Workflow Systems
  - Kepler
  - ESSW
  - myGrid

Why Provenance

- Has been interpreted to mean the complete set of base data used to derive the result in question.
- Much nice theoretical work.
- Trio system.
- But not very useful in practice…

Why Example

"Return books that cost more than average"

ABC, Bar, Foo, PQR, XXX.

"Why is "Foo" in the result?"

Why provenance answer is the set of prices for all books.

Unexpected Difficulties

- Real systems will produce unexpected results at times.
- Good systems must be able to explain why.

Unexpected Behavior

- Unable to query
- Inconsistent results using two query paths

- E.g. (in MiMI)

  "For the query ovo AND organism:dro*, I get back a result;

  For the query organism:dro*, I get back a long list, but if I search for ovo within that list, it is not present."

Unexpected Results

- Often important (lead to discovery)
- But more often anomalous

- E.g. (in MiMI)
  - The molecule record of p53 says that it interacts with 308 other molecules.
  - But only 298 interaction records involving p53 exist

Adequate Explanation

- Losing his tail was probably painful and unexpected for the lizard. Why did it happen?

Explanation: Someone wanted him for lunch, so his tail detached allowing him to escape. Therefore, while painful and unexpected, the behavior was reasonable.

- A query for “cheap flights” returns: Los Angelos $75, Boston $100, San Francisco $400. Why is SF in this list?

Explanation: $400 was less than half the average price for a ticket to San Francisco.

How to capture provenance

- Alice is copying information from different sites
- Bob is actively searching annotation and repository sites for sequences
- Carol has to track sources and scripts run to establish provenance for what Alice and Bob did.
- How to alleviate the user burden?
  - Buneman, P., Chapman, A., and Cheney, J. (June, 2006) Provenance management in curated databases. *ACM SIGMOD*.

How to efficiently store prov.

- Provenance size can easily grow larger than the data size.
  - MiMI 1.1 ~300MB
  - Provenance ~ 4 GB
- How can we shrink the size of the store while still being able to use provenance information?
  - Chapman, A and Jagadish H.V. Efficient Provenance Storage. *In Preparation.*

How to query provenance

- Can we present a huge, unreadable series of manipulations succinctly to the user?
  - Users won't care about some details, can we be smart?
- Once provenance is compressed, can we access it efficiently?

How to easily add provenance to relational databases

- Lots of people use relational databases (mySQL, Access, Oracle, etc).
- Is there a standardized set of rules that will automatically capture sufficient provenance without growing too large?

Conclusions

- Provenance tracking is challenging
- There is no consensus on what to store, how to capture it, or how to store it.
- Need a theory of explanations – which go beyond mere provenance.