Data Integration and Genomic Medicine
Brenton Louie, Peter Mork, Fernando Martin-Sanchez, Alon Halevy, Peter Tarczy-Hornoch

Bio2RDF: Towards a mashup to build bioinformatics knowledge systems
Francois Belleau, Marc-Alexandre Nolin, Nicole Tourigny, Philippe Rigault, Jean Morissette
What is Bioinformatics ???

Computer systems

BIOINFORMATICS

Biological systems
Why Bioinformatics ???

Descriptive, observational science → Hypothesis driven

Predictive information Science → “Discovery” driven

Internet

RESEARCH

SEARCH
If not for Bioinformatics !!

- Structural Plasticity of the Human Genome
  (Copy number variants)
- Individual Human Variation (when a mutation is not a mutation!)
- Alternate Splicing
- Non-Coding RNAs (genes?)

None of these most important genetic discoveries would have been possible !!!
Data Integration.. Why is that important ??

NIH

GenBank

Entrez

NCBI

• Submissions
• Updates

DDBJ

CIB

• Submissions
• Updates

SRS

getentry

EBI

• Submissions
• Updates

NIG

EMBL

• Submissions
• Updates

EMBL

• Submissions
• Updates
Common Data Integration Architectures

Data Warehouses
- Fast queries and clean data
- Stale Data, Complex Schema

Database Federation
- Current Data, Flexible architecture
- Slower queries, Complex Schema, unclean Data

Database federation with mediated schema
- Current Data, Flexible architecture, schema tailored to users
- Slower queries, complex schema, unclean data, mapping from source schema to mediated schema required
Peer data management systems

👍 Current Data, Flexible Architecture, Schema
Tailored to users, Mapping between schemas
distributed across peers

👎 Experimental, slower queries, unclean
data
Two Dimensions of Data Integration
The Integration Axis
(Where the data resides)

Data and Knowledge Representation
Integration Architecture

- Data Warehouse
  - Faster Queries – non trivial for biologists since performance is often the key
  - Handling Volumes: The volume of data in this field is simply too high to handle. Updates suffer and Maintenance becomes an issue
  - Schema Restrictions: The restriction of inability to create a global schema is a deterrent since data is extremely rich
  - Best suited for specific and narrow areas of research. Eg. UCSC Genome Browser, BioMolQuest..
Database Federations

- Common Data Model – maintains a common data model and relies on schema mapping for integration
- Federations relieve the temporal problems of a data warehouse since they reside at the source and are updated constantly
- Some of the extremely difficult queries could be solved using database federations
Database Federations with Mediated Schema

- Dealing with Various Source Schema - This drawback of database federations is dealt by having a database federation with mediated schema
- Federations as Middleware – The federations with mediated schema act as middleware, where data sources are mapped to mediated schema
- Best suited to situations when researchers need to ask complex questions spanning disparate knowledge resources.
Peer Data Management Systems

- Tailored and Focused Mediated Schema – Developing such schemas and integrating is PDMS.
- Each Data source provides a semantic mapping to one or more peers
- Addresses the problem of creating a global mediated schema
- Technology still in evolutionary stage
Data and Knowledge Representation

- Relational Schemas
  - Traditional model of table with tuples and attributes
  - Well understood and robust, but is modeling complex
  - Hierarchically structured biological data is difficult to model
  - Most common and ubiquitous
Semi Structured Data

- Free from rigid structures
- Data with a series of labels and associated values
- More natural modeling of Biological data due to features like nesting
- Complex relationships are still difficult to model
- XML, RDF are examples
Ontology

- Defined as a “specification of a conceptualization”
- Best suited to represent semantic web
- Specify objects classes, relationships and functions
- Well suited for representing biological data
<?xml version="1.0"?>
<GeneList>
  <Gene symbol="CREB-17A" organism="D. melanogaster">
    <Sequence>ACTGCGACCGCCCTGC/GCTGCT</Sequence>
    <Sequence>ACTGCGACGGCCCGAGCGCT</Sequence>
    <Product>cAMP-regulated enhancer-binding protein</Product>
    <Function id="0007616" status="confirmed">Term>long-term memory</Term></Function>
  </Gene>
  <Gene symbol="CER" organism="D. melanogaster">
    <Sequence>TTCGTC/GCTAGCT</Sequence>
    <Function id="0007616" status="inferred">Term>long-term memory</Term></Function>
  </Gene>
</GeneList>
Genomic Medicine with relevance to Data Integration

- Modern Human Genetics
  - Researchers “Swim a sea of data” to study diseases and their links to genes
  - Lack of Standards, Presence of huge number of data sources makes it even more difficult
  - Queries often vague and highly complex, require join of multiple databases
  - Difficulties in combining clinical and genetic information
Microarray Studies
- Genes represented as spots on microarrays
- For each experiment, external annotation needed which often come from public databases
- Need integrated information to perform studies effectively
Application of Data Integration concepts to genomic medicine
BioBanks

- Also known as a biorepository
- A place that collects, stores, processes and distributes biological materials and the data associated with those materials
- Stored as Relational Tables
- [http://www.ukbiobank.ac.uk](http://www.ukbiobank.ac.uk) – a public biobank
MicroArrays

- A multiplex technology used in molecular biology and in medicine
- It consists of an arrayed series of thousands of microscopic spots of DNA oligonucleotides, called features
- A repository containing microarray gene expression data is the Microarray database
Genomics in Clinical Practice and Rational Drug Design

- Technologies of Future !!! Not yet completely developed
- Some breakthrough achieved.. Drugs like Relenza to treat influenza
- Rational drug design is the creation of drugs based on the structure of the drug receptor
- Drug Design is based on the structure of the protien
Gaps in DI research to facilitate genomic medicine

- Data Availability
  - Clinical data still scarce in comparison to bioinformatics data

- Privacy
  - Issues of “De-Identification” still an issue
  - Every DNA is a Unique fingerprint

- Data issues
  - Most data available as Natural Text, More mining required
Lack of Standards

- Too much data, Too little standards
- Integration of diverse complex data types including genomic, proteonomics, clinical, pharmological and chemical requires standards for proper semantic integration of heterogenous data
Questions  ??????