Data Integration in the Life Sciences

Sarah Cohen-Boulakia
Université Paris Sud, LRI CNRS UMR 8623
cohen@lri.fr
01 69 15 32 16

https://www.lri.fr/~cohen/BIGDATA/biodata-ami2b.html
Introduction

- Understanding Life Sciences
  - Progress in multiple domains: biology, chemistry, maths, computer science...

- Emergence of new technologies: Next generation sequencing,...
  - Increasing volumes of raw data
  - All stored in Web data sources

- Raw data are not sufficient
  - Data Annotated by experts
  - Bioinformatics analysis of data
  - New data sources

- Concrete example: Querying NCBI Entrez
  (« Gquery NCBI » on google 😊)
What is known about the Long QT syndrome?

Sarah Cohen-Boulakia, Université Paris Sud
OMIM entry (Long QT)

http://omim.org/entry/611818

Several pages of (structured) text describing the Long QT9 form of the disease

Manual annotations only (few data)

Curated data (physicians)
What is known about the Long QT syndrome?
A lot of gene-centric information
Genomic context, genomic regions...

Gathering of data
What is known about the Long QT syndrome?
One GenBank entry (Long QT)

KVLQT1 - A LONG QT SYNDROME GENE WHICH ENCODES KVLQT1 WHICH COASSEMBLES WITH
GenBank: DI042621.1
FASTA Graphics

**GenBank id**

- GenBank is a deposit of sequences
- Each sequence must be uploaded to GenBank
- A GenBank entry = nucleotide sequence + one reference + a few comments

Raw data

`aris Sud`
Wrap-up

- Even if scientists use a portal, querying biological databases is not easy...
- High **heterogeneity** of the sources
  - Very different kinds of contents
    - Free text (OMIM), semi-structured data (GenBank)...
    - From free text to controlled vocabulary (free text to Ontologies)
- Diverses levels of data **quality**
  - From automatically obtained (EntrezGene) to manually annotated (OMIM)
- Different **Biological entities**
  - OMIM : Disease
  - Entrez Gene : Gene
  - GenBank : Nucleotides

⇒ A bit of history...
Robbins, R. J. (1994). "Report of the invitational DOE Workshop on **Genome Informatics I: Community Databases**." [Rob94a]

- DOE funded large parts of the **Human Genome Project**

- “Continued HGP progress will depend in part upon the ability of genome databases to answer increasingly complex queries that span multiple community databases. Some examples of such queries are given in this appendix.”

- “Note, (...), none of the queries in this appendix can be answered. The current emphasis of GenBank seems to be providing human-readable annotation for sequence information. Restricting such information to human-readable form is totally inadequate for users who require a different point of view, namely one in which the sequence is an annotation for a computer-searchable set of feature information.”
Twelve Queries Unanswerable in 1994

1. Return all sequences which map 'close' to marker M on chrom. 19, are put. members of the olfactory receptor family, and have been mapped on a contig
   - Multidatabase: Chromosome maps from GDB, sequence-contig in GenBank, annotation from elsewhere

3. Return the map location, where known, of all alu elements having homology greater than "h" with the alu sequence "S".
   - Only needs GenBank and a similarity search

4. Return all h. gene sequences for which a putative functional homologue has been identified in a non-vertebrate organism
   - Human: GenBank, non-vertebrates: species databases; how to describe function?

8. Return the number and a list of the distinct human genes that have been sequenced
   - What is a gene? Semantic heterogeneity and scientific uncertainty

11. Return all publications from the last two years about my favorite gene, accession number X####.
    - Synonyms & homonyms; naming conventions, disambiguation
Take Home Message

- The **classical problems** are all there already
- Distributed information
- Semantic heterogeneity
- Scientific uncertainty and evolving concepts
- Naming conventions on the object level
- Naming conventions on the concept level
- Inclusion of non-standard processing
Task: Find genes that play a central role in the response of a host to a pathogen
- Bacteria / viruses must attach to cells to have an influence
- Attachment is a physical binding of proteins
- This binding provokes a reaction in the cell, transmitted by more PPI (e.g. transient signaling)

Known PPI between host and pathogen

Intersect

Expand graph with neighboring PPIs

Filter for overrepresented subnetworks

GSEA to find relevant processes

Study co-regulation by shared TFBS

Hosts genes differentially expressed during infection
Known PPI between host and pathogen

Hosts genes differentially expressed during infection

Intersect

Expand graph with neighboring PPIs

Filter for overrepresented subnetworks

GSEA to find relevant processes

Study co-regulation by shared TFBS

PPI Databases (IntAct, MINT, ...)

Microarray Databases (GEO, AE)

Gene Ontology + GOA

TFBS databases (TracsFac, Jaspar)
Known PPI between host and pathogen

ID Mapping

Quality filtering

Intersect

Expand graph with neighboring PPIs

Filter for overrepresented subnetworks

GSEA to find relevant processes

Study co-regulation by shared TFBS

Unstructured annotations

Uncertainty

Microarray Databases (GEO, AE)

TFBS databases (TracsFac, Jaspar)

ID Mapping

ID Mapping

ID Mapping

Gene Ontology + GOA

ID Mapping

ID Mapping

ID Mapping

PPI Databases (IntAct, MINT, …)
Take Home Message

- The number of sources to be used has increased a lot
- The diversity of the sources has increased a lot
- The complexity of the questions to be answered has increased a lot
Emergence of New Trends

- The number of sources to be used has increased a lot
  - Scalability of integration in number of sources
  - One major goal of the Semantic Web, development of ontologies

- The diversity of the sources has increased a lot
  - Inclusion of quality as a first-class citizen
  - Ranking of integrated search results

- The complexity of the questions to be answered has increased a lot
  - Integration requires analysis and analysis requires integration
  - Scientific workflows
This Tutorial

- Part I – Data Integration for the Life Sciences
  - Biological data & biological databases
  - Some Myths, some Truths

- Part II – Presence

- Part III – Current Trends and Conclusions
Are BDB Distributed?

- > 1,000 different databases
  - Plus many data sets that are not stored in a DB
  - e.g. Supplementary material

- Content is highly redundant
  - Replica (sequence databases)
  - Large unintentional overlaps (KEGG – Reactome)
  - Large intentional overlaps (species specific data)
  - Some databases mostly copy from other sources

- Content may be curated during copying
  - Inconsistencies

Each year, the NAR (Nucleic Acid research) journal has a database issue, listing the databases available
Extreme Example: Protein-Protein Interactions

- There are >500 BDBs related to PPI and pathways
  - See http://www.pathguide.org

- Manually created “source” DBs
Inconsistent understanding of what a PPI actually is
- Binary, physical interaction
- Complexes
- Transient, functional association

Some integrated DBs have imported more data than there is in the sources

Source databases overlap to varying degrees
- Effort to sort things out in IMex consortium

Largely different reliability of content
- Literature-curated, high-throughput experiments (false positive rate), results transferred from orthologs, ...

Literature-curated DBs do not exhibit higher quality than HT [CYS08]
- Re-annotation reveals inconsistencies, subjective judgments, errors in gene assignment, ...
Are BDB Heterogeneous?

- Technical heterogeneity: a bit
  - Web services, HTML forms, ...

- Syntactic heterogeneity: not much of a problem any more
  - XML exchange, flatfiles
  - Many ready-to-use parsers are available

- **Semantic heterogeneity: terrible**
  - Objects have **several names** and IDs (and versions and states)
  - Definition of object types are heterogeneous, scientifically uncertain, and **change over time**
  - Schema element names are heterogeneous
  - **Metadata** often is not available in sufficient depth

- As usual – distribution creates (semantic) heterogeneity
A stretch of DNA (with holes) on a chromosome that at some stage gets translated into a protein.
A re-assembly of stretches of DNA that are transcribed together plus some further editing on the mRNA level
What is a Gene (3)?

- Like Def.2, plus **parts of the sequence downstream** that is necessary to regulate transcription of the gene
What is a Gene (4)? [GBR+07]

- The same gene?
  - Genes may generate different assemblies (differential splicing)
  - Gene duplications in a genome
  - The „same“ gene in another organism
  - Mutation of a gene
  - Genes with a different start site

- A gene?
  - Pseudo genes (never transcribed, yet highly similar)
  - Non-coding genes
  - miRNA (25 bases!)

- Gene definitions change(d) over centuries, decades, and ... last years
Is Data Quality an Issue in BDB?

- Most important quality aspects: **Completeness and error-freeness**
- BDB have terrible problems in both aspects
  - Complete collections exist nowhere (maybe except PDB and GenBank)
  - All BDB have a severe level of all kinds of errors
  - Much copy-and-paste problems (predictions become reality)
- Recall: Most BDB are filled from (high-throughput) experiment
  - Experiments that are not perfect
  - Measurements that are highly **context-dependent**
  - Performing the same experiment again will produce different results
- Recall: **Things change** a lot over time
  - New techniques
  - New knowledge
Are Data Volumes huge?

- All of EMBL now has ~150 TB (zipped), ENSEMBL has ~1TB (MySQL dump), UniProt has ~5GB (zipped)
- Probably 90% of the 1300 DB’s in NAR have <1GB
- All secondary databases have “little” data
- Primary data explodes due to Next Generation Sequencing
Sequencing has become commodity

- Sequencing dozens of genomes/exomes feasible for any mid-size research project
- In 5 years: Hundreds of genomes
  - (Inter-)national projects: 100,000+ genomes
- Access to genomes is crucial: Bioinformatics goes medical
  - “Translational Bioinformatics”
Data Tsunami

Stein, L. D. (2010). *Genome Biol*
Is Reproducibility an Issue?
Is Reproducibility an Issue? Studies on reproducibility

  - 31/50 (62%) provide no information
    - no version of the tool + no parameters used + no exact genomic reference seq.
  - 7/50 (14%) provide all the necessary details
Is Reproducibility an Issue? 
Studies on reproducibility

  - 31/50 (62%) provide no information
    - no version of the tool + no parameters used + no exact genomic reference seq.
  - 7/50 (14%) provide all the necessary details

  - 10 papers in the top-50 IF journals → 500 papers (publishers)
    - 149 (30%) were not subject to any data availability policy
      (0% made their data available)
    - Of the remaining 351 papers
      - 208 papers (59%) did not adhere to the data availability instructions
      - 143 make a statement of willingness to share
      - 47 papers (9%) deposited full primary raw data online
Impacts of irreproducibility...

Raise standards for preclinical cancer research

C. Glenn Begley and Lee M. Ellis propose how methods, publications and incentives must change if patients are to benefit.

Efforts over the past decade to characterize the genetic alterations in human cancers have led to a better understanding of molecular drivers of this complex set of diseases. Although we in the cancer field hoped that this would lead to more effective drugs, historically, our ability to translate preclinical oncology results has the highest failure rate compared with other therapeutic areas. Given the high cost in oncology, it is understandable that barriers to clinical development may be lower than for other disease areas, and a large number of drugs with suboptimal preclinical validation will fail. Researchers must reassess their approach to translating discovery research into gene clinical success and impact.

Many landmark findings in preclinical oncology research are not reproducible, in part because of inadequate cell lines and animal models.

47/53 “landmark” publications could not be replicated

[Begley, Ellis Nature, 483, 2012]

Must try harder

Too many sloppy mistakes are creeping into scientific papers at the data — and at themselves.

Error prone

Biologists must realize the pitfalls massive amounts of data.

If a job is worth doing, it is worth doing twice

Researchers and funding agencies need to put a premium on ensuring that results are reproducible, argues Jonathan F. Russell.

The case for open computer programs

Six red flags for suspect work

C. Glenn Begley explains how to recognize the preclinical papers in which the data won’t stand up.

Know when your numbers are significant

Sarah Cohen-Boulakia, Université Paris Sud
Impacts of irreproducibility (cont.)

- Attacks on authors, editors, reviewers, publishers, funders...

→ Nature checklist
→ Science requirements for data and code availability

http://www.nature.com/nature/focus/reproducibility/index.html
Wrap-up

Integration more necessary than ever in the Life Sciences

- Biological data sources
  - Increasingly numerous, heterogeneous, distributed,...

- Provenance is needed to understand and interpret data, ranking techniques has to be developed

- Breadth of scientific questions increases
- Reproducibility is a major issue
  - Scientific workflows
- Data sources contains errors
- Need standardization
  - Ontologies
This Tutorial

- Part I – Data Integration for the Life Sciences
  - Biological data & biological databases
  - Some Myths, some Truths

- Part II – Integration -- Presence

- Part III – Current Trends and Conclusions
Global schema (relational) = set of tables *integrated*
Local Schema = set of entries (XML, relational...)
Mapping between the local and global schemas to be designed
Wrappers transform the format of the source data sets into the global schema of the DWH. Syntactic integration.

The data warehouse can contain a collection of (redundant) tables or curated data (semantic integration).

(+): Fine (semantics) integration is possible
(-): Updating the warehouse is the major issue.

The Presence

XML + Python + MySQL

- Or better

  XML +
  (Perl | Java | Python) +
  (MySQL | Oracle | PostGreSql)

- Big role of open source libraries and frameworks
- Ontologies are common practice
The Presence

- Architecture
  - Portals are used a lot but do not perform *tight* integration
  - Federated systems are mostly dead
    - Despite frequent papers stating the opposite
    - Survival in some niches: DAS, some mash-ups (no queries)
  - “*Data Warehouses*” approaches everywhere

- Semantic integration
  - No schema matching, little query rewriting
  - *Performed manually* (in custom-written wrappers)

- Several systems up-and-running integrating *dozens of sources*
  - Freshness in the presence of data cleansing remains a hard problem
- Generic relational schema for representing sequences and features
- Standard storage layer for BioPerl, BioPython, BioJava
- Ready-made parsers from Genbank, UniProt, NCBI Taxonomy, ...
BioWarehouse [LPW+06]

- Follows common ETL design
- Unified schema defined manually
  - Leads to semantic differences within tables
  - No cleansing or de-duplication
  - Mappings are programmed in the „loader“
- Loader for 14 sources
- Full provenance information
- Versioned data
- Ships with JAVA lib and GUI

[LPW+06]
“GMOD is the Generic Model Organism Database project, a collection of open source software tools for creating and managing genome-scale biological databases”

- Developed by app. 20 organizations
- Ships with schema (Chado), genome browser, annotation pipeline, exchange middleware, web-app development tool, ... InterMine

- Essentially everything that many small/midsize genome projects need
- Of course: Integrating several GMOD databases is fairly simple
BioMart actually is capable of accessing distributed data sources
Source schemas must comply to BioMart layout and naming conventions
Links and schemas have to be declared and configured in the middleware
No semantic integration, no query optimization / rewriting
BioMart Portal: >100 databases
Full provenance information
- You query a source, not a relation
Highly successful
Comparision of 4 solutions [TB13]

- 11 queries, several environments, profiles, gold standards, benchmark...
- **InterMine**
  - (+) excellent results and flexibility,
  - (-) demanding in terms of development effort
  → for labs with IT resources.
- **PathwayTools**
  - (-) little customization
  - (+) easy-to-use, accurate
- **BioMart**
  - (-) not highly generic/expressive
  - (+) tight integration, unified and customizable interface; configured with minimal efforts.
- **BioXRT**
  - (-) not supported anymore
... and many more ...

- All following the „DWH“-approach

- GUS [DCB+01]
- IMG [MKP+05]
- ArrayExpress [SPLO05]
- Atlas [SHX+05]
- Biozon [BY06]
- GeWare [RKL07]
- GenoQuery [LLF08]
- ...
Distributed Annotation System

[JSB+08]

- Federated system serving a single type of information
  - Genomic annotation
- **DAS server** receives query (genomic coordinates) and broadcasts to all **DAS providers**
- Results are chained and reported
- No semantic integration, no annotation types, simple XML format, very simple protocol, ambiguous query semantic
- Highly successful
Wrap-Up

- Probably >95% of integration projects use materialization
- Successful systems implemented by domain scientists, with little participation of DR
- Very little semantic integration, very little query optimization, very little data fusion, very little schema matching / schema integration
- Full provenance information can/should be recorded
This Tutorial

- Part I – Data Integration for the Life Sciences
  - Biological data & biological databases
  - Some Myths, some Truths

- Part II – Integration -- Presence

- Part III – Current Trends and Conclusions
Trend 1

Analysis is integration and integration is analysis
Integration Classical View (recall)
Classical View - DWH

Queries

Data Warehouse

RDBMS Files Web Service Apps Web Source

Sarah Cohen-Boulakia, Université Paris Sud
Classical View - Expanded

Data integration and analysis workflow

Workflow input ports:
- microarray_pathways
- qiime_pathways
- common_pathways
- regex
- split_pathway_ids
- kegg_pathways
- merge_pathway_desc
- remove_null_values

Workflow output ports:
- intersecting_pathways

Diagram:
- RDBMS
- Files
- Web Service
- Apps
- Web Source

Data Warehouse

Sarah Cohen-Boulakia, Université Paris Sud
The True Architecture in Many Projects

Data Warehouse

Workflow input ports
- microarray_pathways
- qtl_pathways
- common_pathways
- regex
- split_pathway_ids
- kegg_pathways
- merge_pathway_desc
- remove_null_values

Workflow output ports
- intersecting_pathways

Sarah Cohen-Boulakia, Université Paris Sud
The Trend
Life Science Research Food Chain

- Experiments
- Data Collection
- Data Analysis
- New Hypothesis
- Results
With DI Workflows

Experiments

Data Collection

Other Data

New Hypothesis

Results
**Scientific Workflow Management System**

- SWFS = WFS for scientific tasks
  - “Data analysis pipeline”
- Complex pipelines are broken into tasks and their connection
- Data flow driven
- Tasks can be executed locally or distributed
- SWFS manages scheduling, process control, logging, recovery, reproducibility, ...
- Equipped with graphical workflow designer
- Several systems available (Taverna, Galaxy, Kepler, ...)

---

Sarah Cohen-Boulakia, Université Paris Sud
Trend 2

Data quality depends on provenance
## Criteria for Relevance

<table>
<thead>
<tr>
<th>Category</th>
<th>Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>User provided</td>
<td>• Assessment of quality of used data sources</td>
</tr>
<tr>
<td></td>
<td>• Assessment of quality of links</td>
</tr>
<tr>
<td></td>
<td>• Currentness, completeness, trust, ...</td>
</tr>
<tr>
<td>Query dependent</td>
<td>• Number of paths allowing to obtain a data item</td>
</tr>
<tr>
<td></td>
<td>• Length of paths</td>
</tr>
<tr>
<td>Domain specific</td>
<td>• Similarity of linked sequences</td>
</tr>
<tr>
<td></td>
<td>• Quality of matching leading to a link</td>
</tr>
<tr>
<td></td>
<td>• ...</td>
</tr>
<tr>
<td>Graph intrinsic</td>
<td>• Topology of the data graph</td>
</tr>
<tr>
<td>Technical issues</td>
<td>• Execution time (joins, distributed query optimization)</td>
</tr>
<tr>
<td></td>
<td>• Budget-based optimization</td>
</tr>
<tr>
<td></td>
<td>• Best-effort optimization</td>
</tr>
</tbody>
</table>
Which source is better?
Which link is better?
Example

Which path is better?
Example

Which objects are reached by more paths?
Life Science Research Food Chain

Experiments

Data Collection

Data Analysis

New Hypothesis

Results
Integration + Ranking

Experiments

Data Collection

Priorize

integrate

Other Data

Rank by Relevance

User

Cutoff

Analysis

Results

New Hypothesis

Prioritize

Data Collection

integrate

Rank by Relevance

Cutoff

Analysis

Results

New Hypothesis

Prioritize

Data Collection

integrate

Rank by Relevance

Cutoff

Analysis

Results

New Hypothesis
Trend 3

Semantic integration can be performed using ontologies (and Web semantics approaches)
Classical View

Queries

Integration Layer, Global Schema

RDBMS
Files
Web Service
Apps
Web Source

Sarah Cohen-Boulakia, Université Paris Sud
Life Science Research Food Chain

Experiments

Data Collection

Data Analysis

New Hypothesis

Results
... using Semantic Web Techniques

Experiments

Data Collection

Other Data

RDF + ontologies

Data Analysis (with OWL)

New Hypothesis

Results

Sarah Cohen-Boulakia, Université Paris Sud
Conclusions

- Data Integration in the Life Science (DILS) is more important than ever
- Portals perform syntactic integration and are frequently used
- Data warehouses are designed in several places. It remains the most frequently used in the Life Science community
- Faced with the increasing number of
  - data,
  - sources,
  - analytic tools,
  - and the increasing complexity of analysis pipelines...
- challenges are numerous...
Conclusions (cont.)

- The complexity of the questions to be answered has increased a lot
  - Integration requires analysis and analysis requires integration
  - **Scientific workflows**

- The diversity of the sources has increased a lot
  - Inclusion of **quality** as a first-class citizen
  - **Ranking** of integrated search results

- The number of sources to be used has increased a lot
  - **Scalability** of integration in number of sources
  - One major goal of the **Semantic Web**, development of **ontologies**
Data and Software Carpentry

- Initiatives worth looking at

- ELIXIR European project (Infrastructure for bioinformatics)
  - Software and data carpentry (coordinator for the French Node)
  - Contact-me 😊: cohen@lri.fr