Integrated Access to Biological Data

Budva, 11th of June 2006
Ainhoa Llorente
Index

- Introduction and objectives
- Biological Data Repositories
- Biological ontologies
- Ontology merging and mapping
- Database annotation
- Example of an ontology merging/mapping
- Conclusions
- Questions and answers
I. Introduction and objectives

- **Starting point:**
  - Human genome sequencing has given rise to a great number of biological data repositories.

- **Problem addressed:**
  - Quantity and heterogeneity

- **Our aim:**
  - To provide an unified access point to diverse biological data repositories.

- **Our challenge:**
  - To change the existing vision of ontologies in biology:
    - **Up to now:** As mere guides for data structure.
    - **From now on:** As integrated modelling of the biological data by combining or associating ontologies.

To handle biological data repositories by means of semantic and artificial intelligence technologies.
II. Biological Data Repositories (i)

Most important categories

- Nucleotides Sequences → DNA
- Amino acid Sequences → Proteins
- Gene expression
- Scientific literature
- Corporate databases
- Health cards
II. Biological Data Repositories (ii)

http://www.math.tau.ac.il/~rshamir/algmb/00/scribe00/html/lec05/node3.html
II. Biological Data Repositories (iii)

THE MOST IMPORTANT DNA DATABASES

- DDBJ/EMBL/GenBank
  International Nucleotide Sequence Database
  DDBJ: DNA Data Bank of Japan
  CNB-DOOB: Center for Information Biology and DNA Data Bank of Japan
  NIG: National Institute of Genetics

- EMBL: European Molecular Biology Laboratory
- EBI: European Bioinformatics Institute

- NCBI: National Center for Biotechnology Information
- NLM: National Library of Medicine

- IAM: International Advisory Meeting
- ICMI: International Collaborative Meeting
II. Biological Data Repositories (iv)

- **The most important Protein Databases**
  - PIR: Protein Information Resource. ([http://pir.georgetown.edu/](http://pir.georgetown.edu/)).
  - PDB: Protein Data Bank ([http://www.rcsb.org/pdb/](http://www.rcsb.org/pdb/)).

- **Gene Expression**
  - GDX
  - ExpressDB ([http://arep.med.harvard.edu/ExpressDB](http://arep.med.harvard.edu/ExpressDB)).

- **Scientific Literature**
  - MEDLINE
  - PubMed
  - UpToDate
II. Biological Data Repositories (v)

- Corporate Databases
- Health Cards:
  - **Future Use:**
    - To facilitate the medical care attention
    - To increase patients mobility comfort
    - Administrative tasks
    - Emergency health cards
    - Specific care records
    - Patients general medical records
    - To match genetic patient data with biological databases
  - **Challenges:**
    - To unify the data to store
    - Unification of the media
    - Unification of medical nomenclature
III. Biological ontologies

- Gene Ontology
- The Microarray Gene Expression Data (MGED)
- UMLS
IV. Ontology Merging and Mapping

- Ontology Merging:

- Ontology Mapping:
## V. Database annotation

### Gene ontology annotations

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Amy1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>amylase 1, salivary</td>
</tr>
<tr>
<td>ID</td>
<td>MGI:BB019</td>
</tr>
</tbody>
</table>

#### Synonyms
Amy-1

#### Genetic Map
- Chromosome 3
- 50.0 cm
- Detailed Genetic Map (1 cm)
- Mapping data (50)

#### Sequence Map
- 11369237-113904173 bp
- strand
- (from Ensembl annotation of NCBI Build 33)

#### Mammalian orthology
- Human
- Cattle
- Rat
- (Mammalian Orthology)

#### Comparative Map
- Mouse
- Human
- Rattus

#### Sequences

<table>
<thead>
<tr>
<th>Sequence Type</th>
<th>Ensembl Name</th>
<th>Length</th>
<th>Strain/Species</th>
<th>Flank</th>
</tr>
</thead>
<tbody>
<tr>
<td>genomic</td>
<td>ENSMUSG000000004540</td>
<td>21756</td>
<td>C57BL/6J</td>
<td>10 Kb</td>
</tr>
<tr>
<td>transcript</td>
<td>NM_007446</td>
<td>1772</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>polypeptide</td>
<td>PO0668</td>
<td>511</td>
<td>Not Applicable</td>
<td></td>
</tr>
</tbody>
</table>

For the selected sequences, download in FASTA format: [Go](#)
VI. Example of ontology merging/mapping

Protege Architecture with OWL Plugin
VII. Conclusions

- The **application** of semantic web technologies to the biological domain is **rather limited** because:
  - the semantic web technologies
  - the tools needed to implement them are still under development.

- **Future advances could be applied to solve the immediate scientific needs:**
  - Data aggregation and interoperability.
  - Unique entry point for data and processes.
  - Agreement in terminology.
  - Syntax and semantics related to biological data.
  - Semantic data annotation to turn human-understandable data into machine-understandable data.
  - Inference languages to extract and generate knowledge from aggregated data.
VIII. Questions and answers
More information:
allorente@robotiker.es