Ranking-Aware Integration and Explorative Search of Distributed Bio-Data

Marco Masseroli, Matteo Picozzi, Giorgio Ghisalberti
marco.masseroli@polimi.it
In the Life Sciences:

- Numerous data, sparsely distributed in many heterogeneous sources
  - Many are ranked data (or partially ranked) of various types, representing different phenomena, e.g.:
    - Physical ordering, e.g. within a genome
    - Analytical order through algorithmically assigned scores, e.g. representing levels of sequence similarity
    - Experimentally measured values, such as gene expression levels
  - The ordering may represent a range of different notions, such as quantity, confidence, or location
BLAST search result for the sequence “Human asparagin synthetase mRNA”

UniProt search result for protein “5-hydroxytryptamine (serotonin) receptor 2A”

Gene expression data result from Array Express
Several integrated databanks, including:

- Entrez Gene, Ensembl
- Homologene
- IPI, UniProt/Swiss-Prot
- Gene Ontology, GOA
- BioCyc, KEGG, Reactome
- InterPro, Pfam
- OMIM, eVOC, …

Numerous integrated data, including:

- 8,085,152 genes of 8,410 organisms
- 31,347,655 proteins of 367,853 specie
- 33,252 Gene Ontology terms and 61,899 relations (is a, part of)
- 27,667 biochemical pathways
- 14,163 protein domains; 7,215 OMIM genetic disorders; …
Several Life Science questions:
- are complex
- to be answered require integration and comprehensive evaluation of different data
  - often distributed, many of which ranked

Answering complex questions requires integration of vertical search services to create multi-topic searches
- where the different topic searches either refine or augment previous search results

Bioinformatics data integration platforms exist
- Ordered data are poorly served or no supported at all by current data integration platforms
Motivating Life Science search examples

1. “Which genes encode proteins in different organisms with high sequence similarity to a protein X and have some biomedical features in common e.g. up/down significantly co-expressed in the same biological tissue or condition Y and involved in the biological function Z?”

2. “Which proteins of a given biochemical pathway are encoded by co-expressed genes and are likely to interact?”

3. “Which proteins in different organisms are most structurally and functionally similar to a given protein?”

4. “Which drugs treat diseases that are likely to be associated with a given genetic mutation?”

Information to answer such queries is available on the Internet, but no available software system is capable of computing the answer.
Common Aspects:

• Multi-topic queries (e.g. sequence similarity, gene expression)
• Ranking composition (e.g. similarity score, diff. expression p-value)
• The answers are on the Web

A knowledgeable user would do the query step-by-step:

• Search proteins similar to a given protein and get their ID
• Search genes that codify such proteins and get their symbol
• Search a gene expression DB and find the differential expression of such genes in the given biological condition / tissue
• Order results by best similarity and differential expression values

After hours of painful search the user might actually succeed!

• Can this be done better?
Search Computing (SeCo) is a 5 year project funded in November 2008 by the European Research Council (ERC) Advanced Grant program. It aims:

1. Develop the informatics framework required for computing multi-topic searches by combing single topic search results from search engines, which are often ranked, with other data and computational resources
   - directly supporting multi-topic ordered data
   - taking into account order when the results of several requests are combined
   - enabling exploration and expansion of search results

2. Apply SeCo technology in different fields, including Life Sciences => Bio-SeCo: Support answering complex bioinformatics queries
Bio-SeCo: SeCo technologies to answer Life Science questions

Life Science example query:

“Which genes encode proteins in different organisms with high sequence similarity to a protein X and have some biomedical features in common, e.g. up/down significantly co-expressed in the same biological tissue or condition Y and involved in a biological function Z?”

This multi-topic case study question can be decomposed into the following four single topic sub-queries, each of these sub-queries can be mapped to an available search service:
Bio-SeCo: SeCo technologies to answer Life Science questions

• “Which proteins in different organisms have high sequence similarity to a protein X?”
  → **BLAST**, a sequence similarity search program, in one of its many implementations, e.g. **WU-BLAST** (http://www.ebi.ac.uk/blast2/) or **NCBI-Blast** (http://blast.ncbi.nlm.nih.gov/Blast)

• “Which genes encode which proteins?”
  → **GPDW** (Genomic and Proteomic Data Warehouse), a query service to a database of genomic and proteomic data (GPDW_protein2gene)
Bio-SeCo: SeCo technologies to answer Life Science questions

- “Which genes are up/down significantly co-expressed in the same biological condition / tissue Y?”
  → **Array Express** Gene Expression Atlas, a search engine of gene expression data (http://www.ebi.ac.uk/gxa/)

- “Which genes are involved in a biological function Z?”
  → **GPDW** (Genomic and Proteomic Data Warehouse), a query service to a database of genomic and proteomic data (GPDW_gene2biologicalFunctionFeature)
Conceptualization of search services

- According to the Search Computing framework each search service has to be **modelled** in order to allow an **organic connection** with the other services.
- SeCo modelling is performed at 3 different levels: **Conceptual**, **Logical** and **Physical**

1. Conceptual Level
2. Logical Level
3. Physical Level
For modelling each service they are realized:

- a **Service Mart** (SM)
- one or more **Access Patterns** (AP)
- a **Service Interface** (SI)
**Bio-SeCo: Sequence alignment search**

### Service mart

`sequenceAlignmentSearch(sequenceAlignmentProgram, searchedDB, querySequence, querySequenceID, querySequenceIDName, foundSequenceSymbol, foundSequenceID, foundSequenceIDName, foundSequenceDescription, foundSequenceOrganism, bestAlignmentScore, bestAlignmentExpectation, bestAlignmentProbability, alignments(score, expectation, probability, matchQuerySequence, matchFoundSequence, matchPattern))`

### Ex. Access pattern

`sequenceAlignmentSearch_byID(sequenceAlignmentProgramI, searchedDBI, querySequenceIDI, querySequenceIDNameI, foundSequenceSymbolO, foundSequenceIDO, foundSequenceIDNameO, foundSequenceDescriptionO, foundSequenceOrganismO, bestAlignmentScoreR, bestAlignmentExpectationR, bestAlignmentProbabilityR)`
Bio-SeCo: WU-BLAST

Service interface

WU_BLAST_byID ("Washington University BLAST",
sequenceAlignmentSearch_byID,
http://www.ebi.ac.uk/Tools/webservices/wsd/WSWUBlast.wsdl)

Input example:
• seaquenchAlignmentProgram: BLASTP
• searchedDB: uniprotKB
• querySequenceID: O14543  • querySequenceIDName: uniprot

Output example:
• foundSequenceSymbol: SOCS3_MOUSE
• foundSequenceID: O35718  • foundSequenceIDName: uniprot
• foundSequenceOrganism: Mus musculus
• foundSequenceDescription: Suppressor of citokine signaling 3
• bestAlignmentScore: 990
• bestAlignmentExpectation: 2.99 e⁻⁹⁸  • bestAlignmentProbability: 2.99 e⁻⁹⁸
Bio-SeCo: Connection patterns

Their pair-wise coupling *connection patterns* useful for computing the answer to the considered case study question are as follows:

- **existsCodingGene_byProteinID** (sequenceAlignmentSearch, protein2gene):
  
  $$[(\text{sequenceAlignmentSearch.foundSequenceID} = \text{protein2gene.proteinID} \ \text{AND} \ \text{sequenceAlignmentSearch.foundSequenceIDName} = \text{protein2gene.proteinIDName})]$$

- **existsExpressedGene_byGeneID** (protein2gene, geneExpressionSearch):
  
  $$[(\text{protein2gene.geneIDName} = \text{"ensembl"} \ \text{AND} \ \text{geneExpressionSearch.queryEnsemblGeneID} = \text{protein2gene.geneID})]$$

- **existsBiologicalFunctionFeature-name_byGeneID** (protein2gene, biological_function_feature):
  
  $$[(\text{biological_function_feature.geneID} = \text{protein2gene.geneID} \ \text{AND} \ \text{biological_function_feature.geneIDName} = \text{protein2gene.geneIDName})]$$
Services registered in the framework are pair-wise related each other through connection patterns that define the available resource network.
Resource network exploration approach through query expansion
Query interface for multi-topic search
http://www.bioinformatics.dei.polimi.it/bio-seco/seco/
Query interface for multi-topic search

- **Biological Function Feature**: Find biological function features
- **Gene**: Find protein coding genes
- **Gene Expression Factor Search**: Find biological conditions or tissues where genes are expressed
- **Gene Expression Search**: Find genes expressed in a biological condition or tissue
- **Genetic disorder**: Find the genetic disorder by biomolecular entity
- **Pathway**: Find the pathway by biomolecular entity
- **Proteins**: Given a gene, find the proteins coded by it
- **Sequence Alignment Search**: Find protein sequences similar to a query sequence

© Marco Masseroli, PhD
“Which proteins in different organisms have high sequence similarity to the protein with UniProt ID: P26367?”

Using **BLAST**, a sequence similarity search program, in one of its implementations, e.g. **NCBI-BLAST**
Results of sequence alignment search on NCBI-BLAST

“Which proteins in different organisms have high sequence similarity to the protein with UniProt ID: P26367?”

Using **BLAST**, a sequence similarity search program, in one of its implementations, e.g. **NCBI-BLAST**

<table>
<thead>
<tr>
<th>Global Tuple Data</th>
<th>NCBI Blast Sequence Alignment Search by Protein ID</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Global Score</td>
</tr>
<tr>
<td></td>
<td>1.00000</td>
</tr>
<tr>
<td></td>
<td>1.00000</td>
</tr>
<tr>
<td></td>
<td>1.00000</td>
</tr>
<tr>
<td></td>
<td>1.00000</td>
</tr>
<tr>
<td></td>
<td>1.00000</td>
</tr>
<tr>
<td></td>
<td>1.00000</td>
</tr>
<tr>
<td></td>
<td>1.00000</td>
</tr>
<tr>
<td></td>
<td>0.50638</td>
</tr>
</tbody>
</table>
Results of protein2gene search on GPDW

"Which genes encode which proteins?"

Using a query service (GPDW_protein2gene) to our GPDW (Genomic and Proteomic Data Warehouse), e.g. for protein with UniProt ID: P26367
Results of gene expression search on Array Express

"Which genes are significantly up or down expressed in tumor?"

Using **Array Express Gene Expression Atlas**, a search engine of gene expression data (http://www.ebi.ac.uk/gxa/), e.g. for gene with Ensembl ID: **ENSG00000007372**
“Which genes are involved in a biological process?

Using a query service GPDW_gene2biologicalFunctionFeature) to our GPDW (Genomic and Proteomic Data Warehouse), e.g. for gene with Entrez Gene ID: 9021 and biological process regulation of metabolic process.

<table>
<thead>
<tr>
<th>Global Tuple Data</th>
<th>GPDW Biological Function Feature by Gene ID and Feature Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global Score</td>
<td>Tuple Score</td>
</tr>
<tr>
<td>1.00000</td>
<td>1.00000</td>
</tr>
<tr>
<td>1.00000</td>
<td>1.00000</td>
</tr>
<tr>
<td>1.00000</td>
<td>1.00000</td>
</tr>
<tr>
<td>1.00000</td>
<td>1.00000</td>
</tr>
</tbody>
</table>
| 1.00000           | 1.00000           | GO:0061246                   | go                                 | regulation of protein metabolic process | Any process that modulates the frequency, rate or extent of the chemical reactions and pathways involving a protein.
Combined search results

The submitted final global query included as input:

- The human *Paired box protein Pax-6 isoform a* protein (UniProt ID P26367) as amino acid sequence X
- *tumor* as pathological biological condition Y
- *regulation of programmed cell death* as biological process Z

Unpredictably, on October 8th 2012, Bio-SeCo discovered the human *PAX7* and *PAX2*, mouse *Pax8* and human *PAX8* genes, ranked by their global score of 0.90661, 0.90407, 0.90354 and 0.90289, respectively (with 1.0 as best score).

The global score is computed according to a score function as a combination of partial scores of intermediate ranked results, e.g. of ranked sequence alignment expectation and gene expression p-value.
### NCBI Blast Sequence Alignment Search by Protein ID

<table>
<thead>
<tr>
<th>Tuple Score</th>
<th>Found Sequence Symbol</th>
<th>Found Sequence ID</th>
<th>Best Alignment Expectation</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.25289</td>
<td>PAX7_HUMAN</td>
<td>P23759</td>
<td>1.35413E-76</td>
</tr>
<tr>
<td>0.23255</td>
<td>PAX2_HUMAN</td>
<td>Q02962</td>
<td>1.72295E-70</td>
</tr>
<tr>
<td>0.22831</td>
<td>PAX8_MOUSE</td>
<td>Q00288</td>
<td>3.22281E-69</td>
</tr>
<tr>
<td>0.22311</td>
<td>PAX8_HUMAN</td>
<td>Q06710</td>
<td>1.16475E-67</td>
</tr>
</tbody>
</table>

### GPDW Gene by Protein ID

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Gene ID Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENSG0000009709</td>
<td>ensembl</td>
</tr>
<tr>
<td>ENSG0000075891</td>
<td>ensembl</td>
</tr>
<tr>
<td>ENSMUSG0000025976</td>
<td>ensembl</td>
</tr>
<tr>
<td>ENSG0000125618</td>
<td>ensembl</td>
</tr>
<tr>
<td>18510</td>
<td>entrez_gene</td>
</tr>
</tbody>
</table>

### Array Express Gene Expression Search by Gene ID

<table>
<thead>
<tr>
<th>Tuple Score</th>
<th>Found Gene Symbol</th>
<th>Best Expression Pvalue</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00000</td>
<td>PAX7</td>
<td>1.0E-11</td>
</tr>
<tr>
<td>1.00000</td>
<td>PAX2</td>
<td>1.0E-11</td>
</tr>
<tr>
<td>1.00000</td>
<td>Pax8</td>
<td>1.0E-11</td>
</tr>
<tr>
<td>1.00000</td>
<td>PAX8</td>
<td>1.0E-11</td>
</tr>
<tr>
<td>0.19063</td>
<td>emb2444</td>
<td>0.0080</td>
</tr>
<tr>
<td>0.17146</td>
<td>AT5G18510</td>
<td>0.013</td>
</tr>
<tr>
<td>0.13016</td>
<td>AT3G18510</td>
<td>0.037</td>
</tr>
</tbody>
</table>

### GPDW Biological Function Feature by Gene Concept

<table>
<thead>
<tr>
<th>Biological Function Feature ID</th>
<th>Biological Function Feature Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0043067</td>
<td>regulation of programmed cell death</td>
</tr>
<tr>
<td>GO:0043067</td>
<td>regulation of programmed cell death</td>
</tr>
<tr>
<td>GO:0043067</td>
<td>regulation of programmed cell death</td>
</tr>
<tr>
<td>GO:0043067</td>
<td>regulation of programmed cell death</td>
</tr>
</tbody>
</table>
See **Bio-SeCo** online at

http://www.bioinformatics.dei.polimi.it/bio-seco/seco/

**Tomorrow DEMO**

Thank you for your attention!

Any question?