International Conference on Computational Science, ICCS 2011
Singapore, 1-3 June, 2011

POLITECNICO DI MILANO

Dipartimento di Elettronica e Informazione

Bio Search Computing: Bioinformatics web service integration for data-driven answering of complex Life Science questions

Marco Masseroli, Giorgio Ghisalberti, Stefano Ceri

marco.masseroli@polimi.it

Data and search service scenario in the Life Sciences

In the Life Sciences:

- Numerous data, sparsely distributed in many heterogeneous sources
  - Many are ranked data of various types, representing different phenomena, e.g.:
    - physical ordering, e.g. within a genome
    - algorithmically assigned scores, e.g. representing levels of sequence similarity
    - experimentally measured values, such as gene expression levels

Data and search service scenario in the Life Sciences

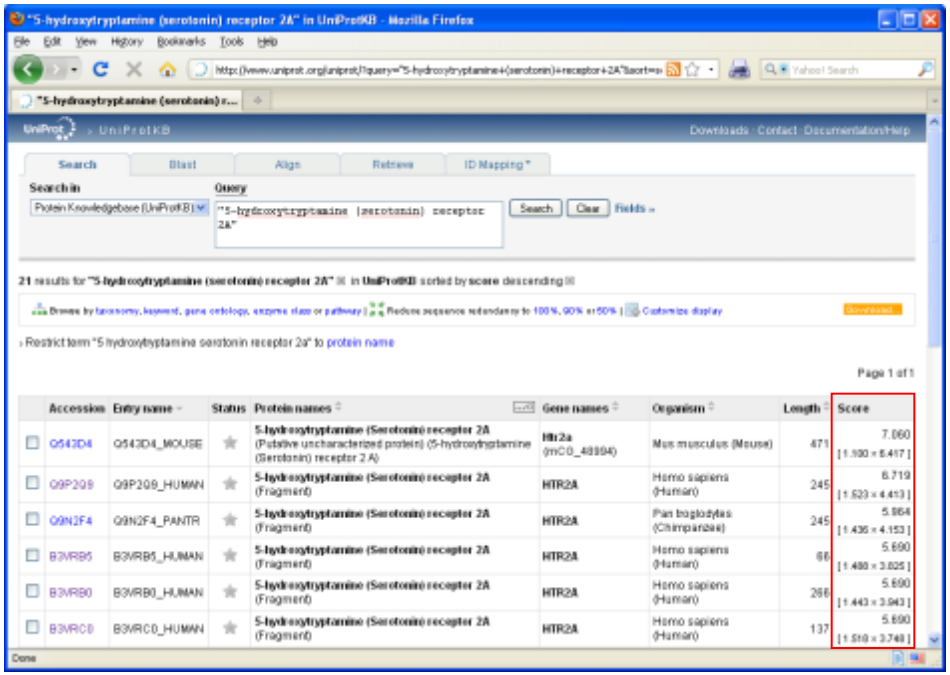
- Many individual and vertical search services:
  - Give rapid and selective access to data from potentially huge repositories
  - Provide results (often ranked) of user defined searches within a data repository
  - Seek individual items that meet the criteria specified in a request,
    - whereas in practice information relevant to a requirement may be spread over several resources
  - Are ineffective to answer a request that involves combining results from more than one search engine

Life Sciences computational and data access web services

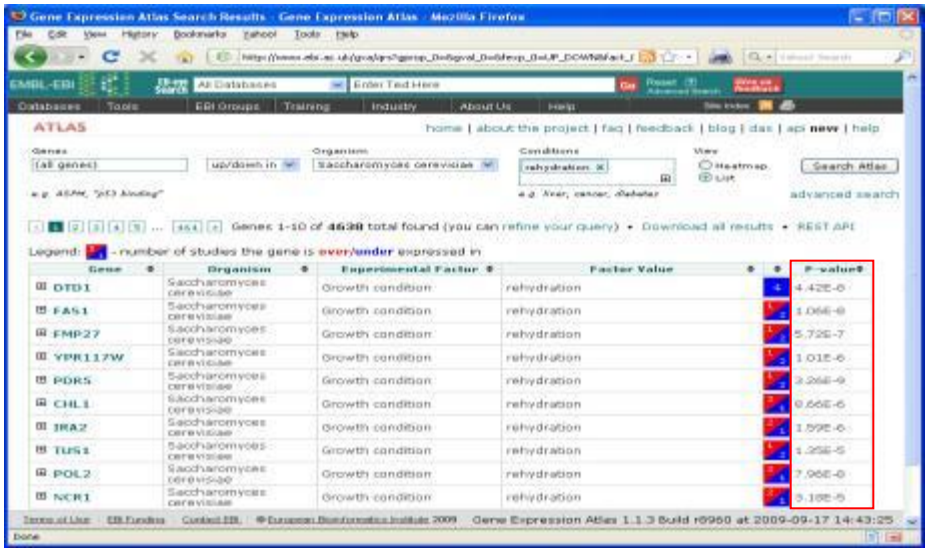
POLITECNICO DI MILANO

BLAST search result for the sequence "Human asparagine synthetase mRNA"

Alignment	DB:ID	Source	Length	Score	Identity%	Positives%	E0
1	EM_PAT:DD130059	Diagnosis and Prognosis of Breast Cancer Patients.	1992	9960	100	100	0.
2	EM_PAT:DD208683	Expression Profile of Prostate Cancer.	1992	9960	100	100	0.
3	EM_PAT:DD415310	Diagnosis and Prognosis of Breast Cancer Patients.	1992	9960	100	100	0.
4	EM_PAT:GM974767	Sequence 120 from Patent EP2003213.	1992	9960	100	100	0.
5	EM_PAT:AR274918	Sequence 55 from patent US 6606607.	1992	9960	100	100	0.
6	EM_PAT:EA062020	Sequence 645 from patent US 7171311.	1992	9960	100	100	0.
7	EM_PAT:EA248495	Sequence 120 from patent US 7229774.	1992	9960	100	100	0.
8	EM_PAT:EA427947	Sequence 120 from patent US 7332290.	1992	9960	100	100	0.
9	EM_PAT:GP320972	Sequence 645 from patent US 7514209.	1992	9960	100	100	0.
10	EM_HUM:M27396	Human asparagine synthetase mRNA, complete cds.	1992	9960	100	100	0.
11	EM_PAT:CQ875273	Sequence 16 from Patent WO2004076613.	1994	9895	99	99	0.
12	EM_PAT:CS353065	Sequence 49 from Patent EP1522594.	1994	9895	99	99	0.
13	EM_PAT:CS380845	Sequence 49 from Patent WO2005040414.	1994	9895	99	99	0.
14	EM_PAT:DD367278	COMPOSITIONS AND METHODS FOR THE DIAGNOSIS AND TREATMENT OF TUMOR.	1994	9895	99	99	0.
15	EM_PAT:DL464877	COMPOSITIONS, KITS, AND METHODS FOR IDENTIFICATION, ASSESSMENT, PREVENTION, AND THERAPY OF CANCER.	1994	9895	99	99	0.
16	EM_PAT:FB671589	Sequence 49 from Patent EP1892306.	1994	9895	99	99	0.



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

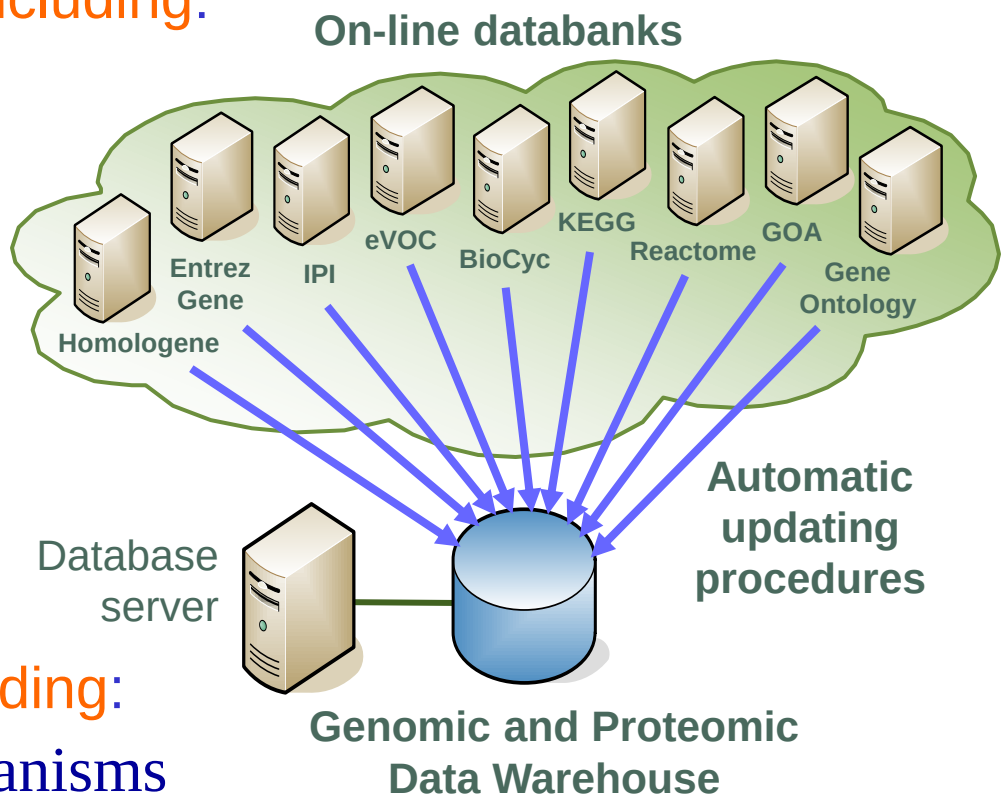


Gene expression data result from Array Express

GPDW: Genomic and Proteomic Data Warehouse

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:

- 6,202,155 genes of 6,276 organisms
- 19,821,588 proteins of 197,257 specie
- 27,836 Gene Ontology terms and 45,644 relations (*is a*, *part of*)
- 8,525 biochemical pathways
- 8,980 protein domains; 2,392 OMIM genetic disorders; …

Life Science questions and their answering

- 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-domain searches

- where the different domain searches either refine or augment previous search results

Bioinformatics data integration platforms exist

- No support for ranked data

Motivating Life Science search examples

1. "Which genes encode proteins in different organisms with high sequence similarity to a given protein and are significantly co-expressed (e.g. up expressed) in the same given biological condition / tissue (e.g. in tumor / brain)?"
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 software system is capable of computing the answer

Motivating Life Science search examples

Common Aspects:

- Multi-domain 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!

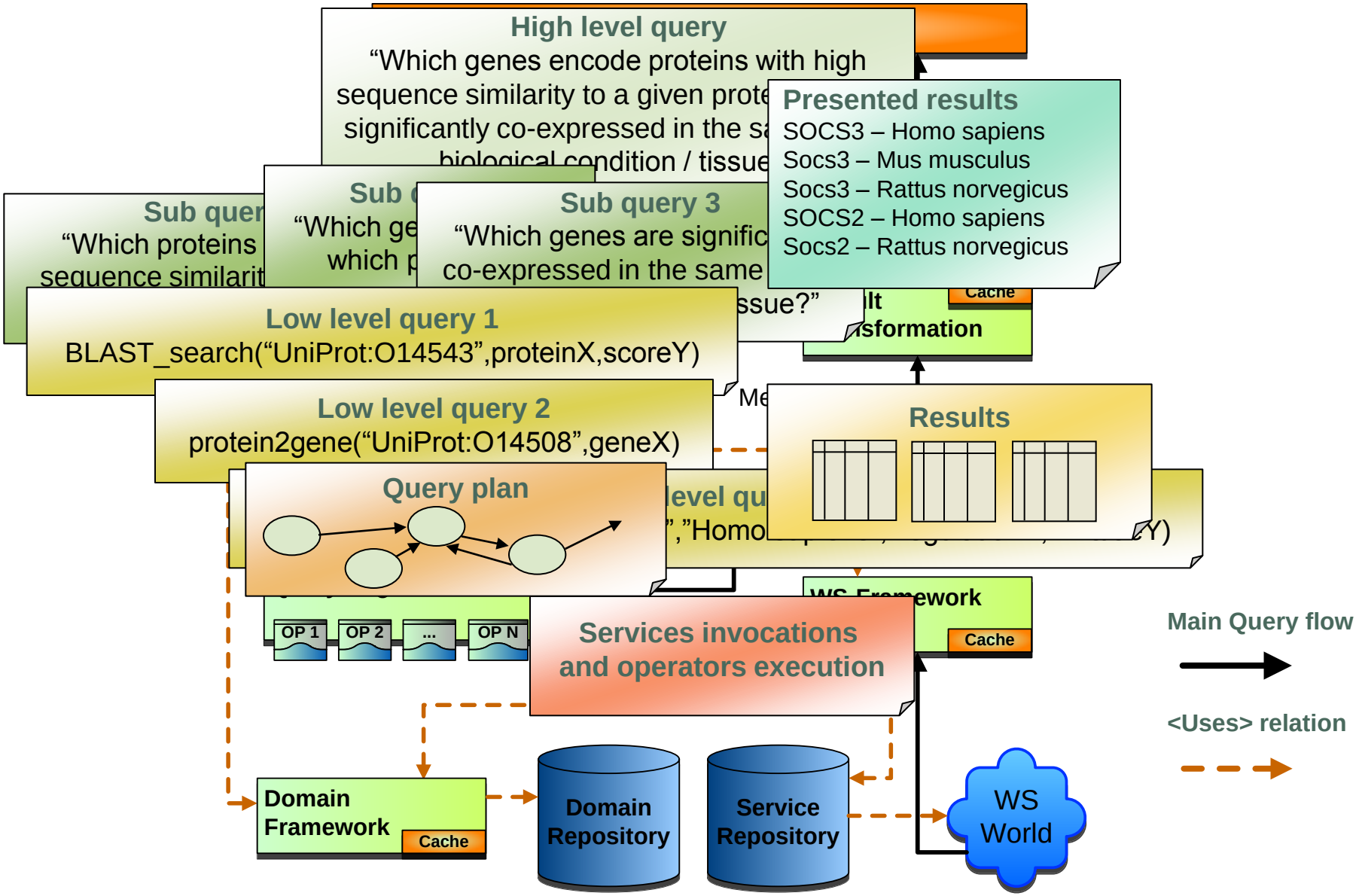
- Can this be done better?

Search Computing: an ERC funded project

Search Computing (SeCo) project (2.5 MEuro funding in 5 years) aims:

1. Develop the informatics framework required for computing multi-domain searches by combing single domain search results from search engines, which are often ranked, with other data and computational resources
   - directly supporting multi-domain 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

Search Computing framework



Registration and conceptualization of search services

Three levels of conceptualization of services and associations

Conceptual level: Service marts

SequenceAlignmentSearch(QuerySequence, FoundSequence, FoundSequenceLenght, Score, …)

Logical level: Access patterns

BLAST_search(Query_Sequence[I], Found_Sequence[O], Found_Sequence_lenght[O], Score[O][R], …)

Corresponding SM attributes

Auxiliary attributes (i.e. query attributes)

Physical Level: Service interfaces

Selector

WU-BLAST: Query_Sequence | Found_Sequence | Length | Score | % identity | …

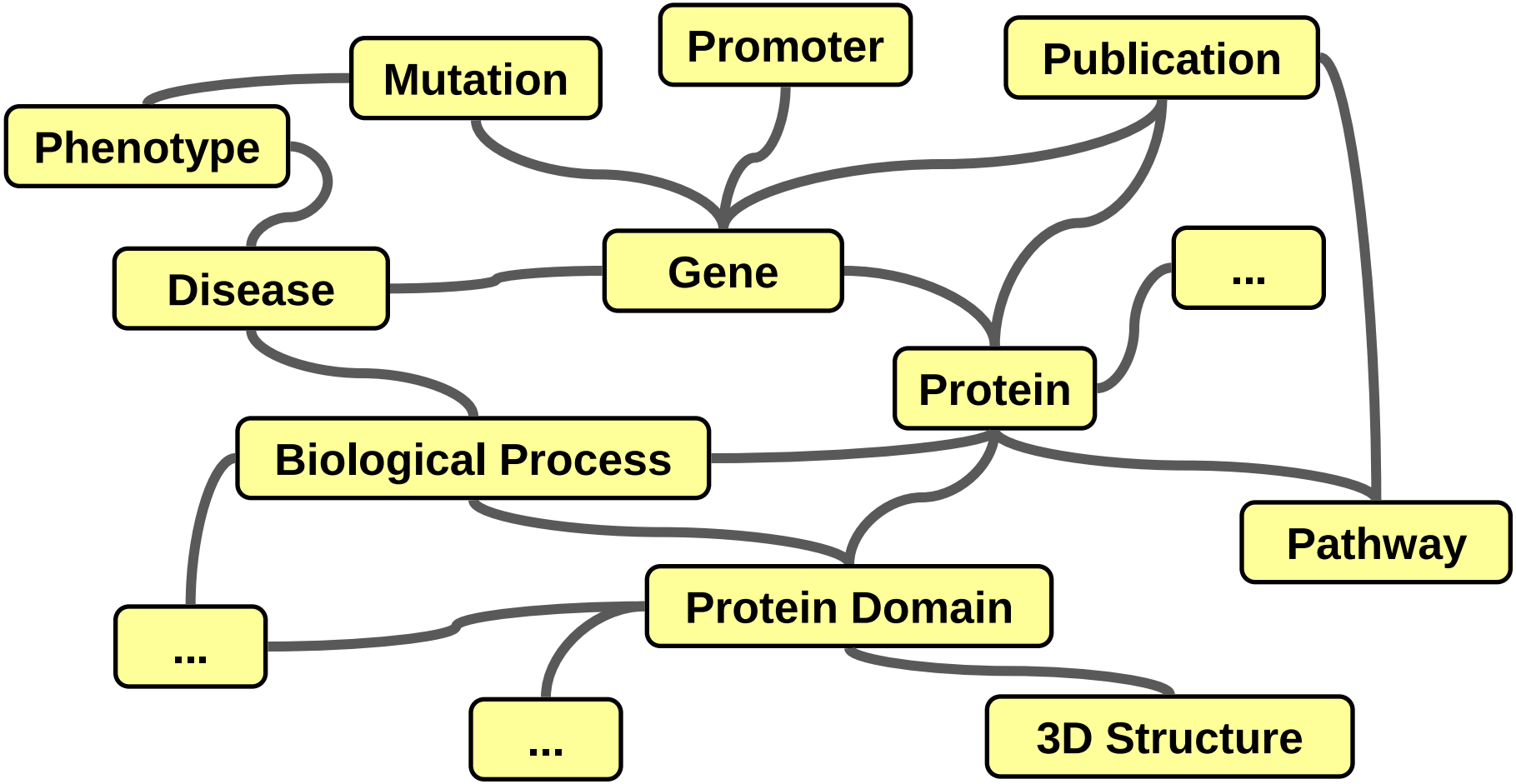
Selector attributes

Corresponding SM attributes

Auxiliary attributes (i.e. query attributes)

Resource network of Life Sciences web services

Services registered in the framework are pair-wise related each other through connection patterns that define the available resource network



Bio-SeCo: SeCo technologies for Life Sciences

Life Science example query:

1. "Which genes encode proteins in different organisms with high sequence similarity to a given protein and are significantly co-expressed (e.g. up expressed) in the same given biological condition / tissue (e.g. in tumor / brain)?"

This multi-domain case study question can be decomposed into the following three single domain sub-queries, each of these sub-queries can be mapped to an available search service:

- "Which proteins in different organisms have high sequence similarity to a given protein?"

  → **BLAST**, a sequence similarity search program, in one of its many implementations, e.g. **WU-BLAST** (http://www.ebi.ac.uk/blast2/)

Bio-SeCo: SeCo technologies for Life Sciences

- "Which genes encode which proteins?"

  → **GPDW** (Genomic and Proteomic Data Warehouse), a query service to a database of genomic and proteomic data (**GPDW_protein2gene**)

- "Which genes are significantly co-expressed (e.g. up expressed) in the same given biological condition / tissue (e.g. in tumor / brain)?"

  → **Array Express Gene Expression Atlas**, a search engine of gene expression data (http://www.ebi.ac.uk/gxa/)

In Search Computing each search service need to be modelled with:

- a Service mart, one or more Access patterns, a Service interface

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(sequenceAlignmentProgram^{I}, searchedDB^{I}, querySequenceID^{I}, querySequenceIDName^{I}, foundSequenceSymbol^{O}, foundSequenceID^{O}, foundSequenceIDName^{O}, foundSequenceDescription^{O}, foundSequenceOrganism^{O}, bestAlignmentScore^{R}, bestAlignmentExpectation^{R}, bestAlignmentProbability^{R})

Service interface

WU_BLAST_byID("Washington University BLAST",
sequenceAlignmentSearch_byID,
http://www.ebi.ac.uk/Tools/webservices/wsdl/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: protein2gene and GPDW

Service mart

protein2gene(proteinID, proteinIDName, proteinSymbol, organism, geneID, geneIDName, geneSymbol)

Ex. Access pattern

protein2gene_byID(proteinID^{I}, proteinIDName^{I}, geneID^{O}, geneIDName^{O}, geneSymbol^{O}, organism^{O})

Service interface

GPDW_byID("Genomic and Proteomic Data Warehouse", protein2gene_byID, http://www.bioinformatics.polimi.it/GPDW/)

Input example:

- proteinID: O35718
- proteinIDName: uniprot

Output example:

- geneID: 12702
- geneIDName: entrez_gene
- geneSymbol: Socs3
- organism: Mus musculus

Bio-SeCo: Gene expression search

Service mart

geneExpressionSearch(queryProperty, queryPropertyValue, queryEnsemblGeneID, queryOrganism, queryRegulation, queryFactorValue, foundGeneSymbol, expressionFactorValue, expressionRegulation, experimentNumber, bestExperimentPvalue)

Ex. Access pattern

geneExpressionSearch_byGeneProperty(queryProperty^{I}, queryPropertyValue^{I}, queryOrganism^{I}, queryRegulation^{I}, queryFactorValue^{I}, foundGeneSymbol^{O}, expressionFactorValue^{O}, expressionRegulation^{O}, experimentNumber^{R}, bestExperimentPvalue^{R})

Service interface

Array_Express_byGeneProperty("Array Express Gene Expression Atlas", geneExpressionSearch_byGeneProperty, http://www.ebi.ac.uk/gxa/api?gene<queryProperty>Is=<queryPropertyValue>&species=<queryOrganism>&format=xml&indent)

Input example:

- queryProperty: Gene
- queryOrganism: Mus musculus
- queryFactorValue: brain
- queryPropertyValue: Socs3
- queryRegulation: updown

Output example:

- foundGeneSymbol: SOCS3
- expressionRegulation: UP
- bestExperimentPvalue: 1.12 e^{-23}
- expressionFactorValue: brain
- experimentNumber: 24

Bio-SeCo: Connection patterns

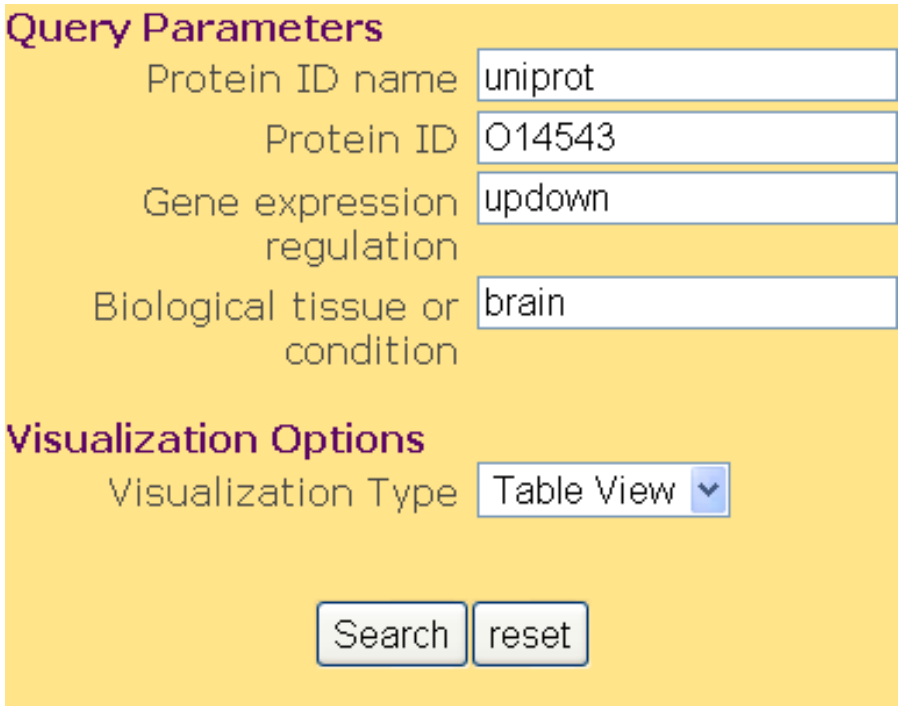
Their pair-wise coupling ***connection patterns*** useful for computing the answer to the considered case study question are as follows:

existsCodingGene_byProteinID(sequenceAlignmentSearch, protein2gene):
[(sequenceAlignmentSearch.foundSequenceID = protein2gene.proteinID
AND sequenceAlignmentSearch.foundSequenceIDName = protein2gene.proteinIDName)]

existsExpressedGene_byGeneSymbol(protein2gene, geneExpressionSearch):
[("Gene" = geneExpressionSearch.queryProperty
AND protein2gene.geneSymbol = geneExpressionSearch.queryPropertyValue
AND protein2gene.organism = geneExpressionSearch.queryOrganism)]

Query interface for multi-domain search

"Which genes encode proteins in different organisms with high sequence similarity to a given protein (e.g. with UniProt ID: O14543) and are significantly co-expressed (e.g. up or down expressed) in the same given biological condition / tissue (e.g. in brain)?"



Results of sequence alignment search on WU-BLAST

"Which proteins in different organisms have high sequence similarity to the protein with UniProt ID: O14543?"

Using **BLAST**, a sequence similarity search program, in one of its implementations, e.g. **WU-BLAST** (http://www.ebi.ac.uk/blast2/)

Sequence Alignment			
Protein ID	Protein Name	Protein Symbol	Expectation
O14543	Suppressor of cytokine signaling 3	SOCS3_HUMAN	2.5999999999999996e-99
Q6FI39	SOCS3 protein	Q6FI39_HUMAN	2.5999999999999996e-99
O35718	Suppressor of cytokine signaling 3	SOCS3_MOUSE	2.9999999999999993e-98
B1AQL6	Suppressor of cytokine signaling 3	B1AQL6_MOUSE	2.9999999999999993e-98
O88583	Suppressor of cytokine signaling 3	SOCS3_RAT	2.0999999999999999e-97
A9JRX2	Socs8 protein	A9JRX2_DANRE	3.6e-21
O88582	Suppressor of cytokine signaling 2	SOCS2_RAT	2.5e-20
O14508	Suppressor of cytokine signaling 2	SOCS2_HUMAN	3.1e-20

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)

Gene Protein Association

Protein ID	Gene Symbol	Organism
O14543	SOCS3	Homo sapiens
Q6FI39	SOCS3	Homo sapiens
O35718	Socs3	Mus musculus
B1AQL6	Socs3	Mus musculus
O88583	Socs3	Rattus norvegicus
A9JRX2	socs8	Danio rerio
O88582	Socs2	Rattus norvegicus
O14508	SOCS2	Homo sapiens

Results of gene expression search on Array Express

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

Using **Array Express** **Gene Expression Atlas**, a search engine of gene expression data (http://www.ebi.ac.uk/gxa/)

Gene Expression					
Gene Symbol	Organism	Factor	Regulation	Experiment Number	P-value
Socs3	Mus musculus	brain	UP	24	1.1218185040451748e-23
Socs3	Mus musculus	brain	UP	24	1.1218185040451748e-23
Socs3	Rattus norvegicus	brain	DOWN	6	5.427190918894098e-10
SOCS3	Homo sapiens	brain	UP	11	2.5128574776545065e-9
SOCS3	Homo sapiens	brain	UP	11	2.5128574776545065e-9
SOCS2	Homo sapiens	brain	DOWN	12	2.9868274520339355e-9
Socs2	Rattus norvegicus	brain	DOWN	5	0.005287489853799343
socs8	Danio rerio	brain	DOWN	1	0.0186142735183239

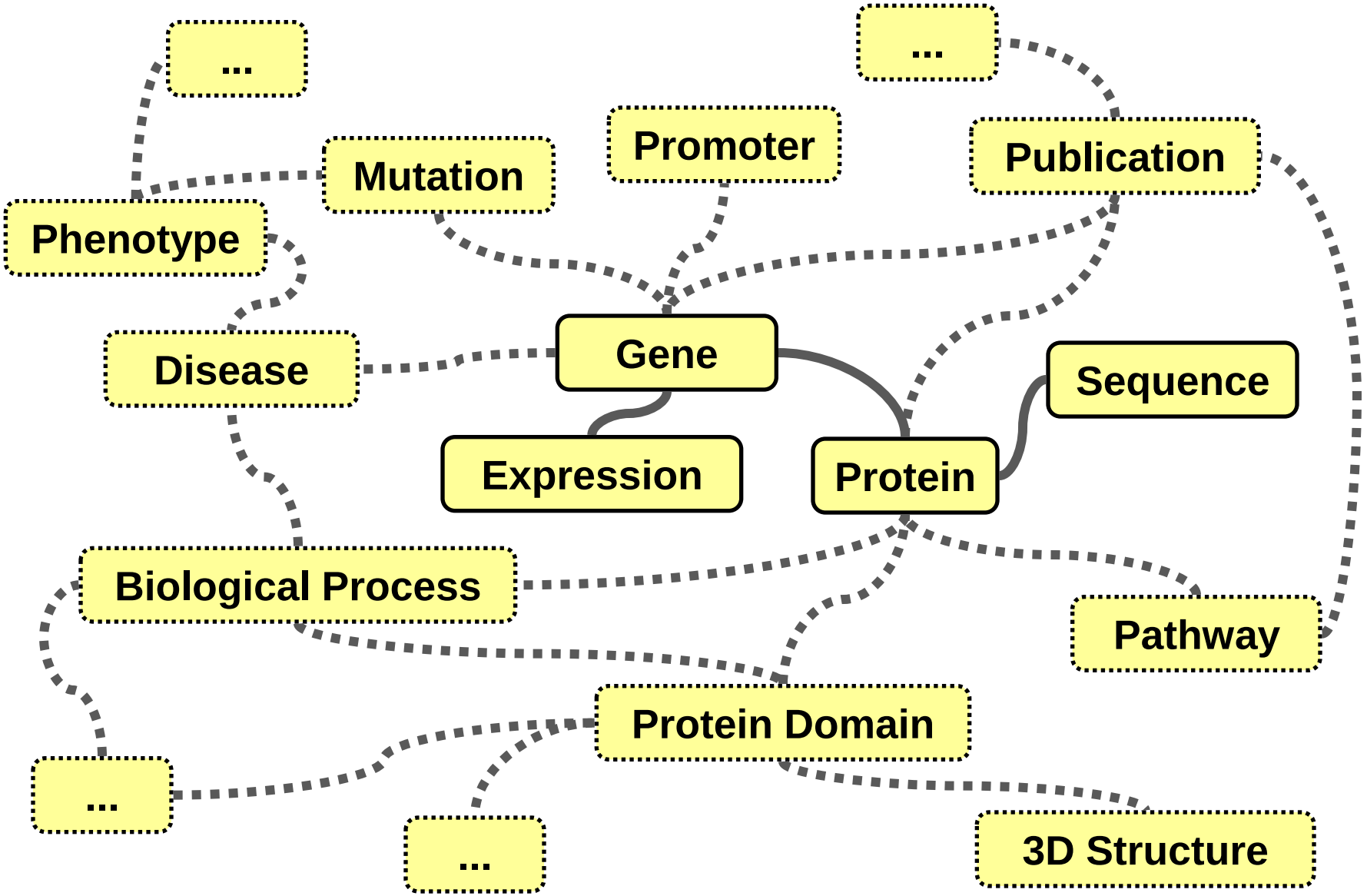
Combined search results

Combination	Sequence Alignment			
Rank	Protein ID	Protein Name	Protein Symbol	Expectation
3.365e-121	O35718	Suppressor of cytokine signaling 3	SOCS3_MOUSE	2.9999999999999993e-98
3.365e-121	B1AQL6	Suppressor of cytokine signaling 3	B1AQL6_MOUSE	2.9999999999999993e-98
6.533e-108	O14543	Suppressor of cytokine signaling 3	SOCS3_HUMAN	2.5999999999999996e-99
6.533e-108	Q6FI39	SOCS3 protein	Q6FI39_HUMAN	2.5999999999999996e-99
1.140e-106	O88583	Suppressor of cytokine signaling 3	SOCS3_RAT	2.0999999999999999e-97
9.259e-29	O14508	Suppressor of cytokine signaling 2	SOCS2_HUMAN	3.1e-20
6.701e-23	A9JRX2	Socs8 protein	A9JRX2_DANRE	3.6e-21
1.322e-22	O88582	Suppressor of cytokine signaling 2	SOCS2_RAT	2.5e-20

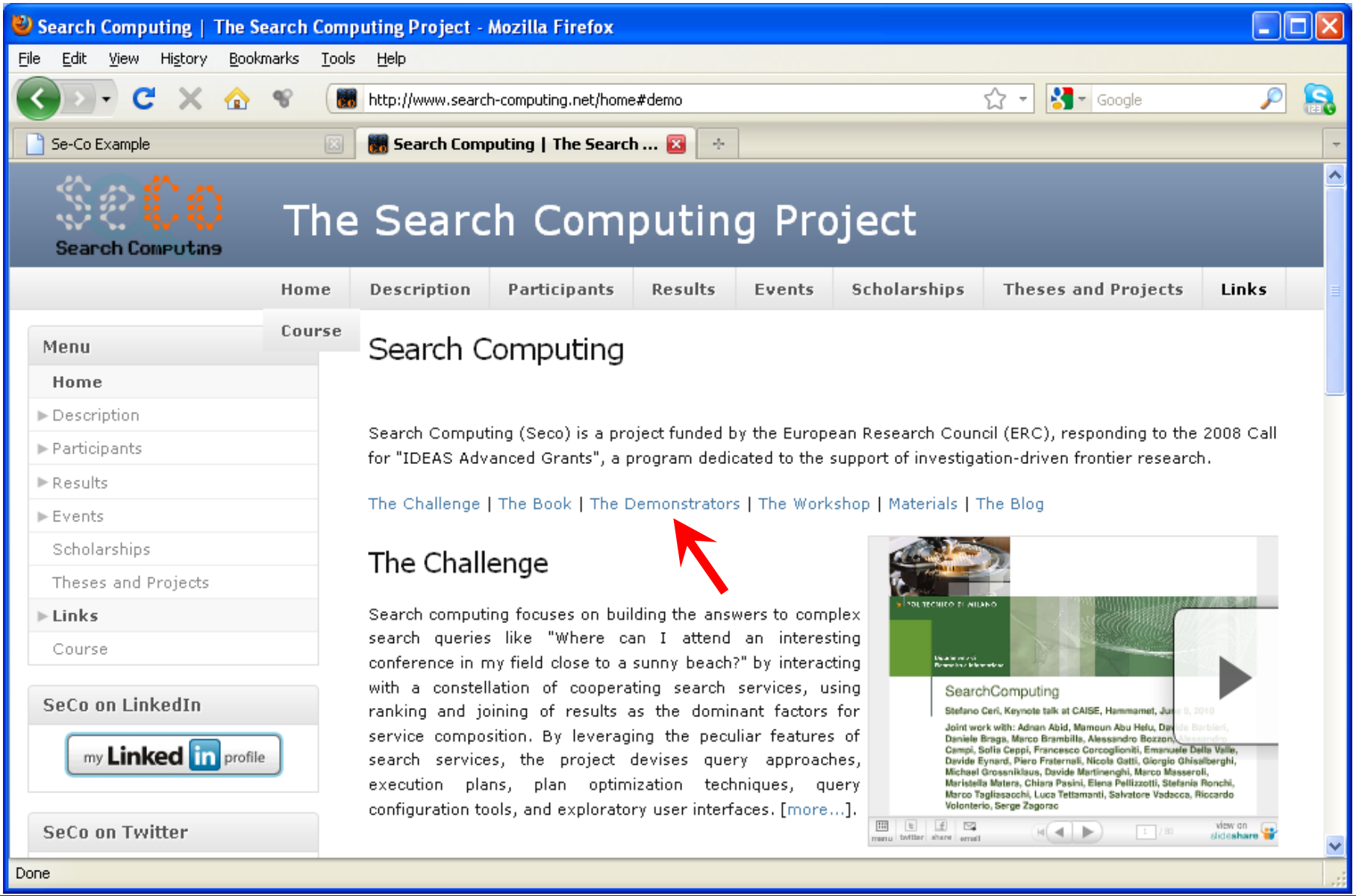
Gene Protein Association		Gene Expression			
Gene Symbol	Organism	Factor	Regulation	Experiment Number	P-value
Socs3	Mus musculus	brain	UP	24	1.1218185040451748e-23
Socs3	Mus musculus	brain	UP	24	1.1218185040451748e-23
SOCS3	Homo sapiens	brain	UP	11	2.5128574776545065e-9
SOCS3	Homo sapiens	brain	UP	11	2.5128574776545065e-9
Socs3	Rattus norvegicus	brain	DOWN	6	5.427190918894098e-10
SOCS2	Homo sapiens	brain	DOWN	12	2.9868274520339355e-9
socs8	Danio rerio	brain	DOWN	1	0.0186142735183239
Socs2	Rattus norvegicus	brain	DOWN	5	0.005287489853799343

*Combination.Rank = sequenceAlignment.Expectation * geneExpression.P-value*

Query expansion on the resource network



Bio-SeCo demo on http://www.search-computing.org/



Bio-SeCo demo on http://www.search-computing.org/

POLITECNICO DI MILANO

Search Computing | The Search Computing Project - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://www.search-computing.net/home#demo

Se-Co Example | Search Computing | The Search ...

interesting: #SemanticSearch tutorial at #ISWC2010: presentations @ http://bit.ly/gOa1BZ #web #search #semanticweb

twitter Join the conversation

Seco on Facebook

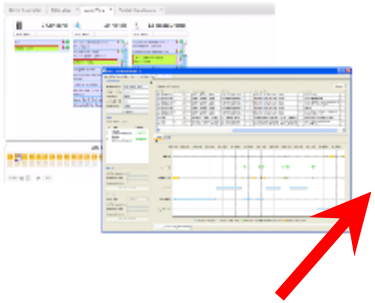
Search Computing su Facebook

Mi piace 122

Search Computing on Facebook

Login

Login



The Demonstrators

The SeCo concepts are being implemented in a comprehensive architectural infrastructure. Several demonstrators are already available online (e.g. BioInformatics Demonstrator, Concert Planning Demonstrator, etc.), so to allow first-hand experiments with the join of services, the exploratory paradigm, the multiple visualization of results (tabular, map, and parallel coordinates), and the functioning of the SeCo Execution Engine. [more...]

Available Materials

All the materials produced by the project are available online. You can find publications, deliverables, invited talks, meeting presentations, course materials, and so on.

The Search Computing Blog

The technology watch blog is a repository of classified links to news, products, and researches that are relevant for Search Computing.

Printer-friendly version

Done

Bio-SeCo demo on http://www.search-computing.org/

POLITECNICO DI MILANO

Bioinformatics Demonstration | The Search Computing Project - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://www.search-computing.net/UIDemoBio

Se-Co Example | Bioinformatics Demonstration | ...

The Search Computing Project

Home | Description | Participants | Results | Events | Scholarships | Theses and Projects | Links | Course

Menu

- Home
- Description
- Participants
- Results
  - Publications
  - Related Publications
  - First Book
  - Second Book
  - Demonstrations
  - Deliverables
  - Technology Watch Blog
  - Course
  - Readings
- Events
- Scholarships
- Theses and Projects
- Links
- Course

SeCo on LinkedIn

Home

Bioinformatics Demonstration

The demonstrator is available **HERE**.

Web search tools have become ubiquitous, with both generic and domain-specific search services providing users with rapid and selective access to data from potentially huge repositories. However, individual search tools are often ineffective for use in applications in which the answer to a request involves combining results from more than one search engine. In particular, web search services typically seek individual documents that meet the criteria specified in a request, whereas in practice information relevant to a requirement may be spread over several resources.

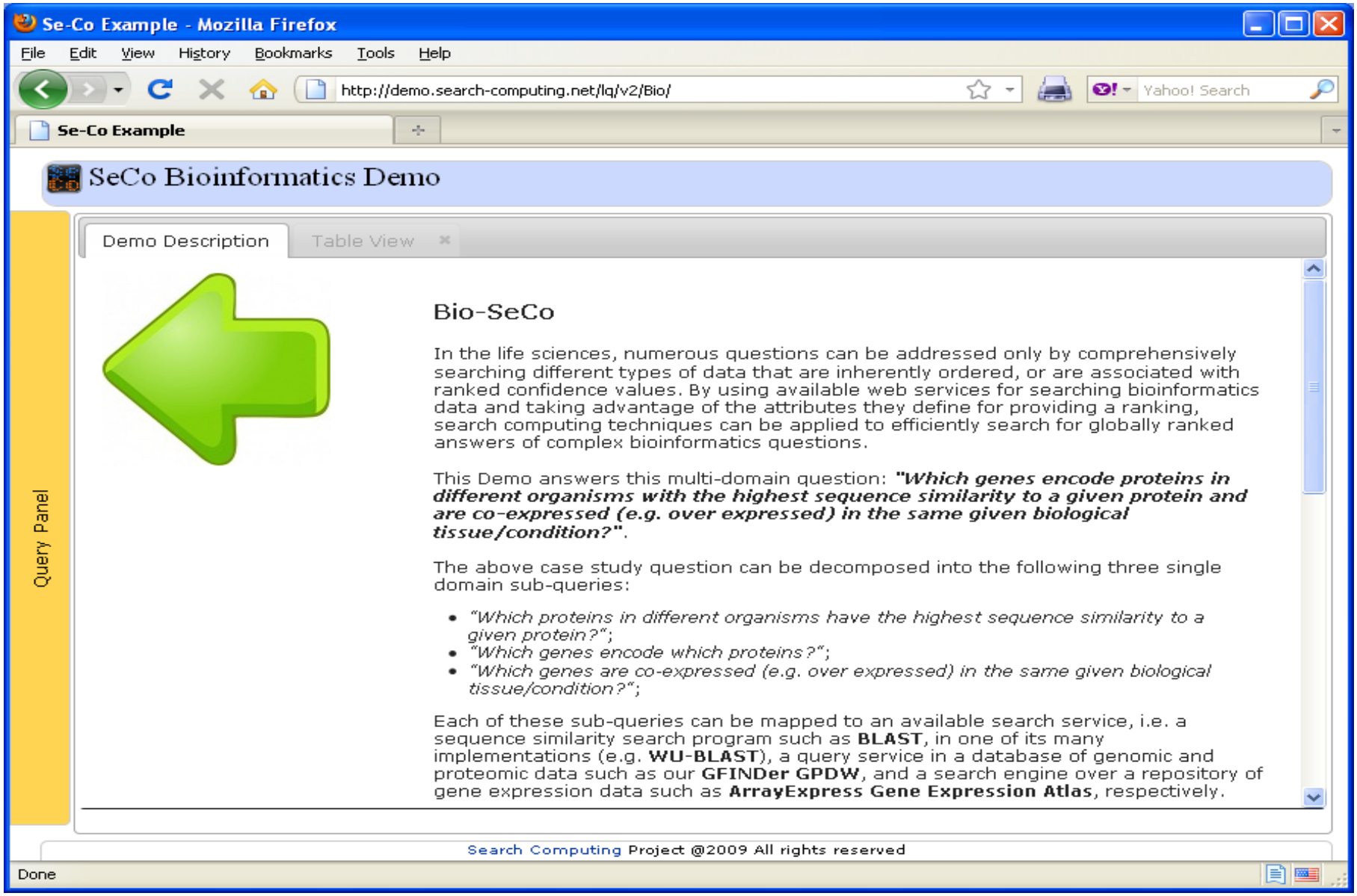
Search computing provides a platform for expressing requests over multiple search services, such that the results of the integrated requests take account of the rankings of individual search results.

In the life sciences, many resources provide vertical search capabilities, in that they are focused on a single domain. In practice, many life science services provide ranked data as results, where the ranking may reflect a property of an algorithm (e.g. a similarity score) or of an experimental result (e.g. an expression level). Furthermore, it is often essential to combine multiple vertical search services to create multi-domain searches, where the different domain searches either refine or augment previous results.

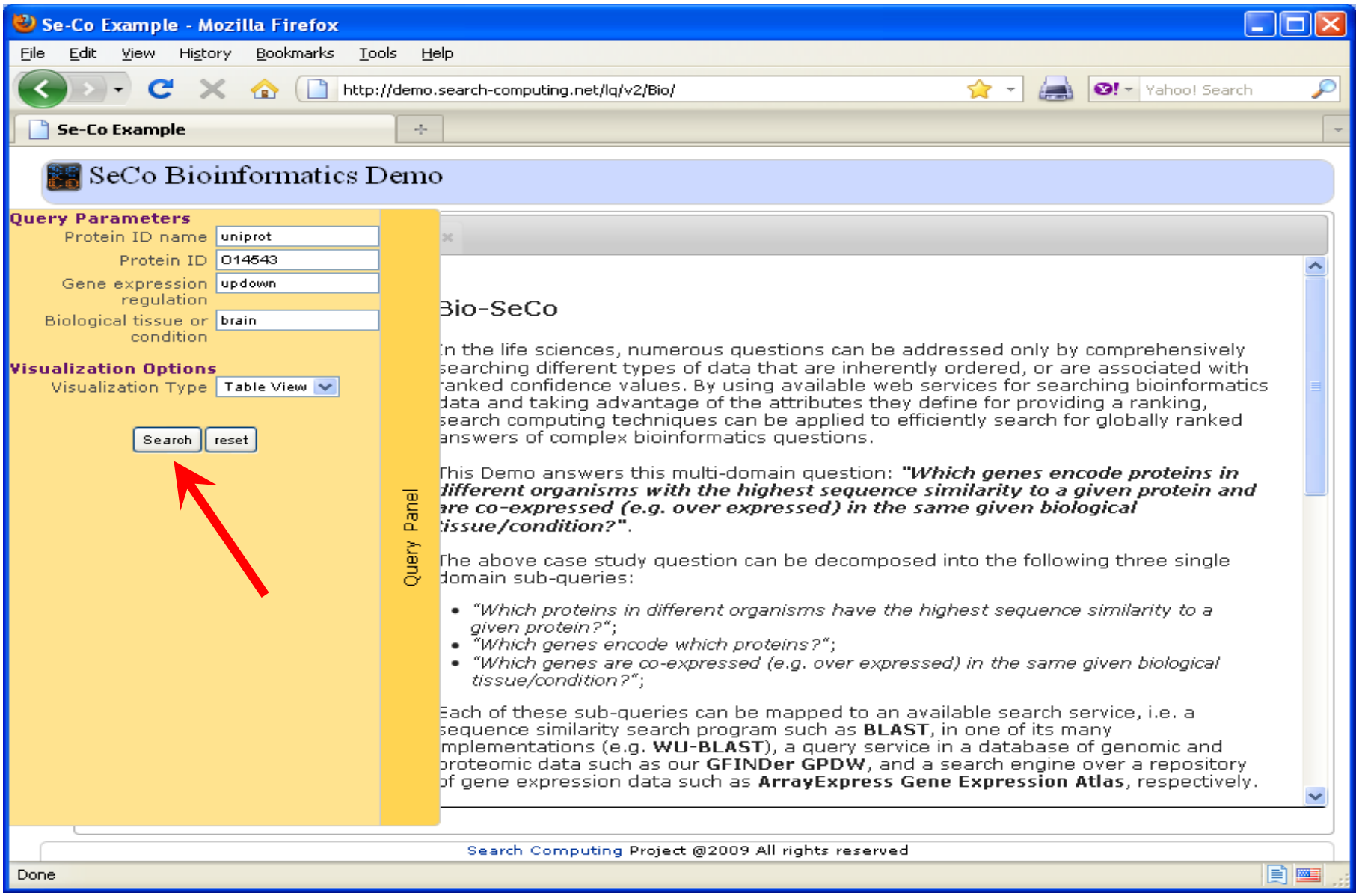
This demo explores the application of a search computing platform in a bioinformatics use case, with a view to identifying the extent to which the existing platform for multi-domain search provides useful facilities for representing and integrating bioinformatics search services.

Done

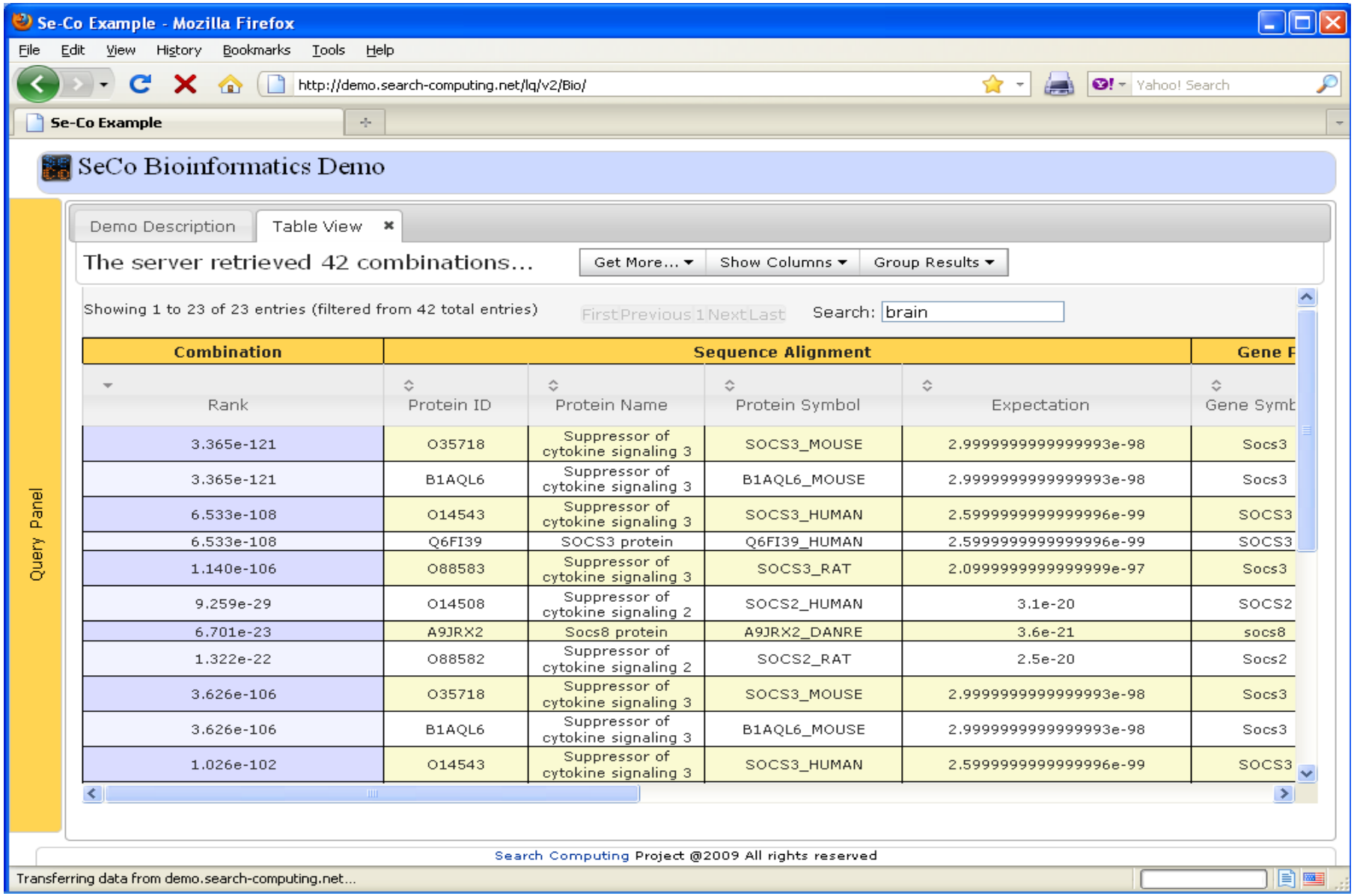
Bio-SeCo demo on http://www.search-computing.org/



Bio-SeCo demo on http://www.search-computing.org/



Bio-SeCo demo on http://www.search-computing.org/



Combination	Sequence Alignment				Gene F
Rank	Protein ID	Protein Name	Protein Symbol	Expectation	Gene Symb
3.365e-121	O35718	Suppressor of cytokine signaling 3	SOCS3_MOUSE	2.9999999999999993e-98	Socs3
3.365e-121	B1AQL6	Suppressor of cytokine signaling 3	B1AQL6_MOUSE	2.9999999999999993e-98	Socs3
6.533e-108	O14543	Suppressor of cytokine signaling 3	SOCS3_HUMAN	2.5999999999999996e-99	SOCS3
6.533e-108	Q6FI39	SOCS3 protein	Q6FI39_HUMAN	2.5999999999999996e-99	SOCS3
1.140e-106	O88583	Suppressor of cytokine signaling 3	SOCS3_RAT	2.0999999999999999e-97	Socs3
9.259e-29	O14508	Suppressor of cytokine signaling 2	SOCS2_HUMAN	3.1e-20	SOCS2
6.701e-23	A9JRX2	Socs8 protein	A9JRX2_DANRE	3.6e-21	socs8
1.322e-22	O88582	Suppressor of cytokine signaling 2	SOCS2_RAT	2.5e-20	Socs2
3.626e-106	O35718	Suppressor of cytokine signaling 3	SOCS3_MOUSE	2.9999999999999993e-98	Socs3
3.626e-106	B1AQL6	Suppressor of cytokine signaling 3	B1AQL6_MOUSE	2.9999999999999993e-98	Socs3
1.026e-102	O14543	Suppressor of cytokine signaling 3	SOCS3_HUMAN	2.5999999999999996e-99	SOCS3

Thank you for your attention!

Any question?