NETTAB 2011 workshop

Clinical Bioinformatics

October 12-14, 2011, Pavia, Italy

POLITECNICO DI MILANO

Dipartimento di Elettronica e Informazione





(Bio) Search Computing

Application in the Life Sciences

Marco Masseroli

marco.masseroli@polimi.it

Search Computing

Search computing (SeCo) is a new approach (and a platform that implement it) for the integration of search engines and their results and of other data and computational resources

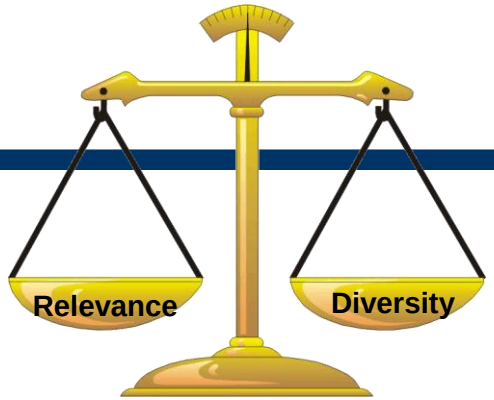
- Provides direct support for multi-domain ordered data
- Reflects the fact that search engines produce ranked outputs
- Order is taken into account when the results of several requests are combined

Motivating daily life search examples:

- "Where can I attend a high rated concert in a place close to an inexpensive hotel and to a good restaurant?"
- "Who is the best doctor who can cure insomnia in a close-by hospital?"
- "Where can I attend an interesting scientific conference in my field in a location reachable with a cheap flight and at the same time relax on a beautiful beach nearby?"

This information is available on Internet, but no software system is capable of computing the answer

Ranking and diversification



Ranking is a first class citizen in SeCo:

- Results of search services are ranked
- Global rankings can be defined over result combinations
- Top-k algorithms are being studied for granting optimal extraction of results that best fit the request (relevance maximization)

However: relevance is not the only parameter that is useful for user satisfaction.

Result diversification for search might be needed for:

- Clustering similar results
- Ambiguity of query terms and word sense diversity
- For a specific word sense, diversity of available information sources

Liquid Query

POLITECNICO DI MILANO

" A new paradigm allowing users to **formulate** and get **responses** to **multi-domain** queries through an **exploratory information seeking** approach, based upon **structured** information sources exposed as software services…"

- **Composite** answers obtained by **aggregating** search results from various domains
- **Highlight** the contribution of each search service
- **Join** of results based on the structural information afforded by the search service interfaces
- **Refine** the user query
- **Re-shape** the result list

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

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 (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

Data and search service scenario in the Life Sciences

Further investigation needed regarding:

- The complex nature of Life Science data
- The frequency of unavailable or missing values
- The diversity of ordering types
- How combining those orderings

Such challenging Life Science applications may represent a good test bed for advanced search computing applications, with valuable spinoffs for other scenarios

Data and search service scenario in the Life Sciences

Ordered data are poorly served by current data integration platforms

Search computing may:

- Provide support for ordering as a first class citizen in integration platforms in the Life Sciences
- Increase the complexity of Life Science questions that integration tools can support directly

Data and search service scenario in the Life Sciences – Bio-SeCo

Aim: Support **answering complex bioinformatics queries** that regard:

- Different types of distributed data
- Also ranked data

Thus, their answer require:

- Integration of data (ranked and not)
- i.e. integration of the services to access these data

What do we have to answer such queries?

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

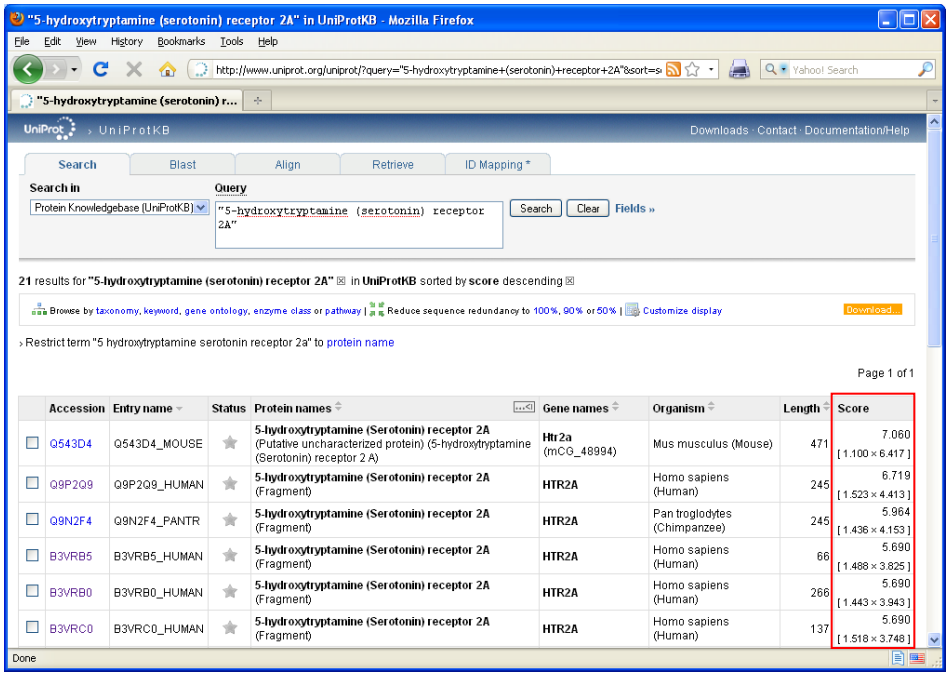
- BLAST (http://www.ebi.ac.uk/Tools/blast2/index.html) to search for nucleotide and amino acid sequences in gene and protein databanks
- ArrayExpress (http://www.ebi.ac.uk/microarray-as/ae/) to query gene expression data
- UniProt (http://www.uniprot.org/) to search in the UniProt databank for proteins related with search keyword(s)
- PubMed (http://www.ncbi.nlm.nih.gov/pubmed/) to search for scientific publications
- Entrez Gene (http://www.ncbi.nlm.nih.gov/gene/) to search in the Entrez Gene databank for genes related with search keyword(s)
- …

Life Sciences computational and data access web services

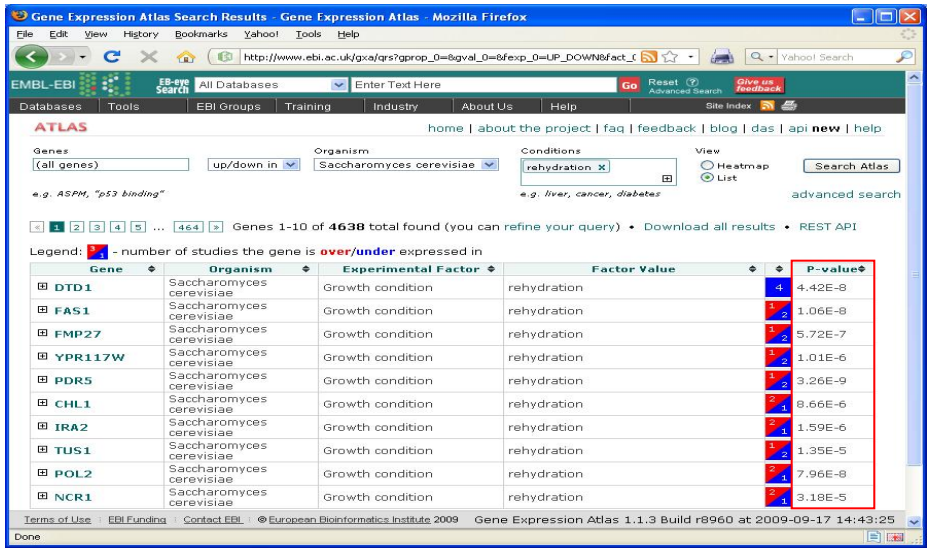
POLITECNICO DI MILANO

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

Alignment	DB:ID	Source	Length	Score	Identity%	Positives%	EO
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 6506607.	1992	9960	100	100	0.
6	EM_PAT:EA062820	Sequence 645 from patent US 7171311.	1992	9960	100	100	0.
7	EM_PAT:EA248485	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:CS063065	Sequence 49 from Patent EP1522594.	1994	9895	99	99	0.
13	EM_PAT:CS080846	Sequence 49 from Patent WO2005040414.	1994	9895	99	99	0.
14	EM_PAT:DD387278	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"

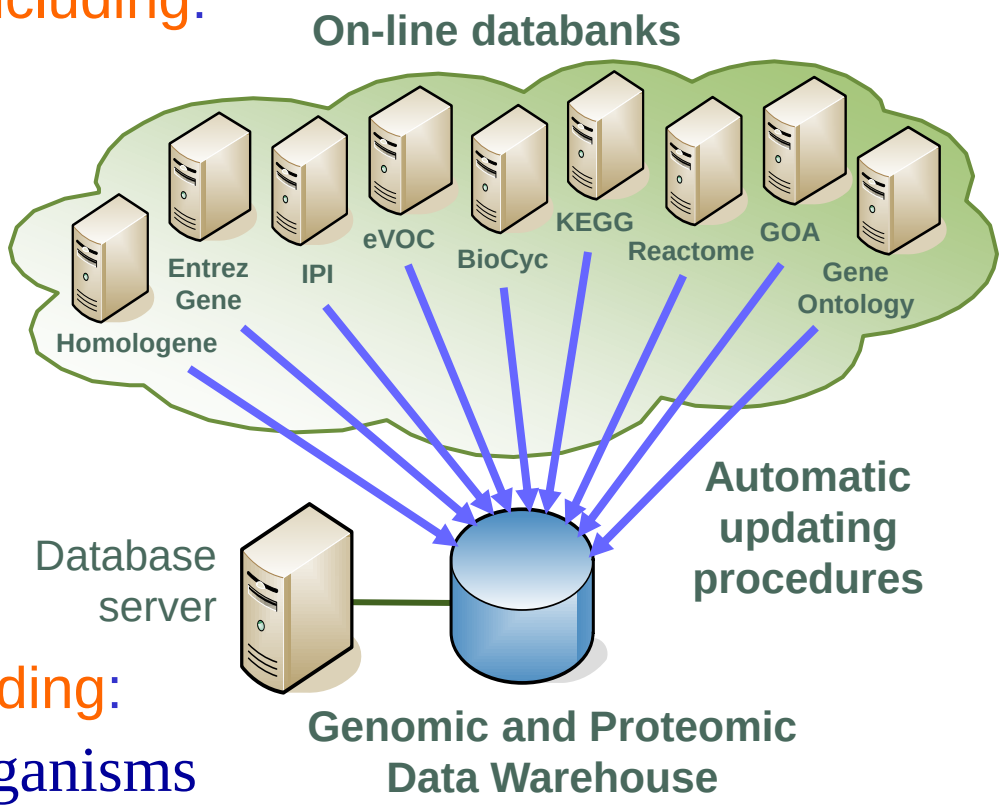


Gene expression data result from Array Express

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:

- 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; …

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-topic searches

- where the different topic 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) is a 5 year project funded in November 2008 by the European Research Council (ERC) Advanced Grant program

It aims:

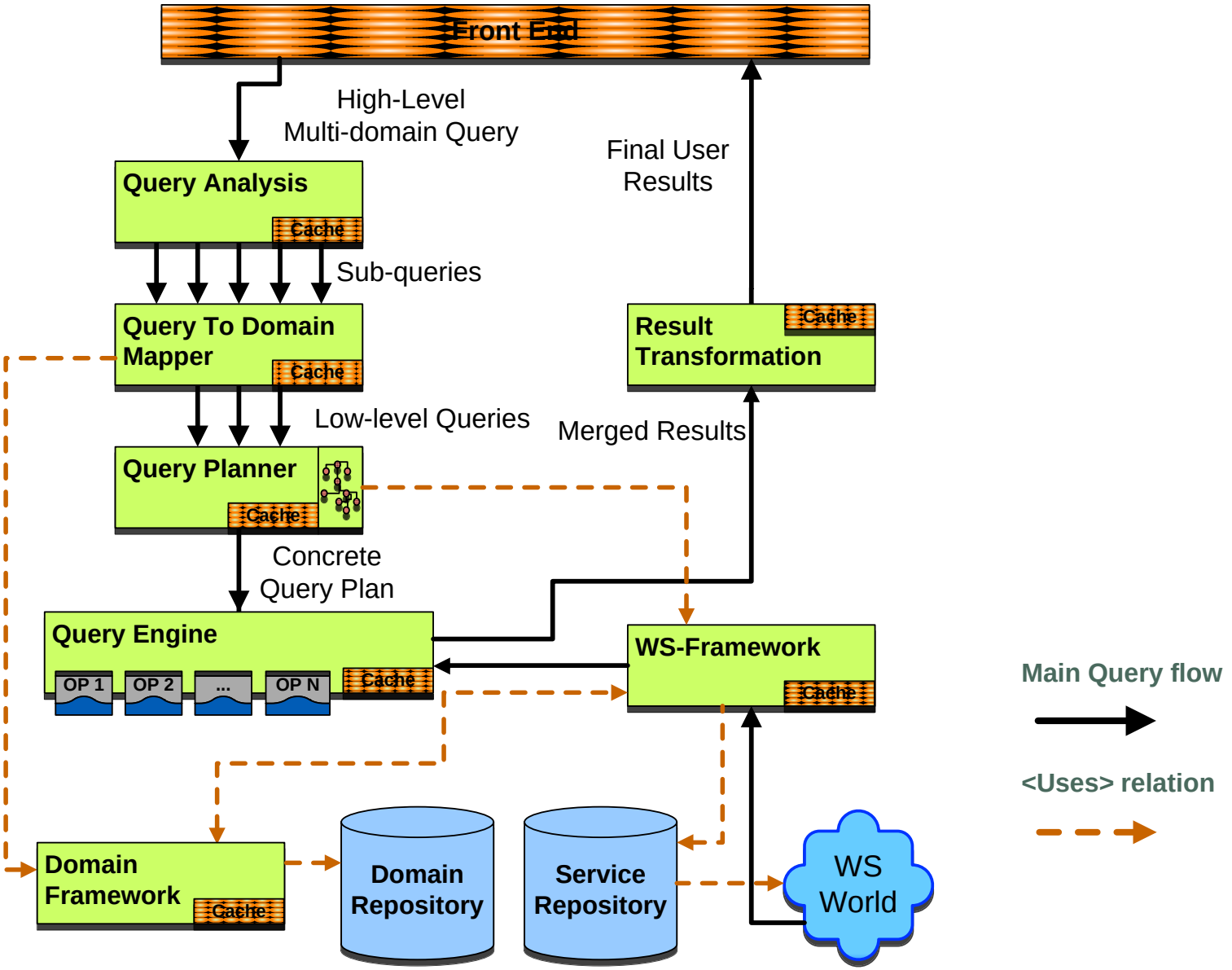
1. Develop the informatics framework required for computing multi-domain searches by combing single domain search results from search engines, which are often ranked, with other data and computational resources
   - directly supporting multi-domain ordered data
   - taking into account order when the 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: an ERC funded project

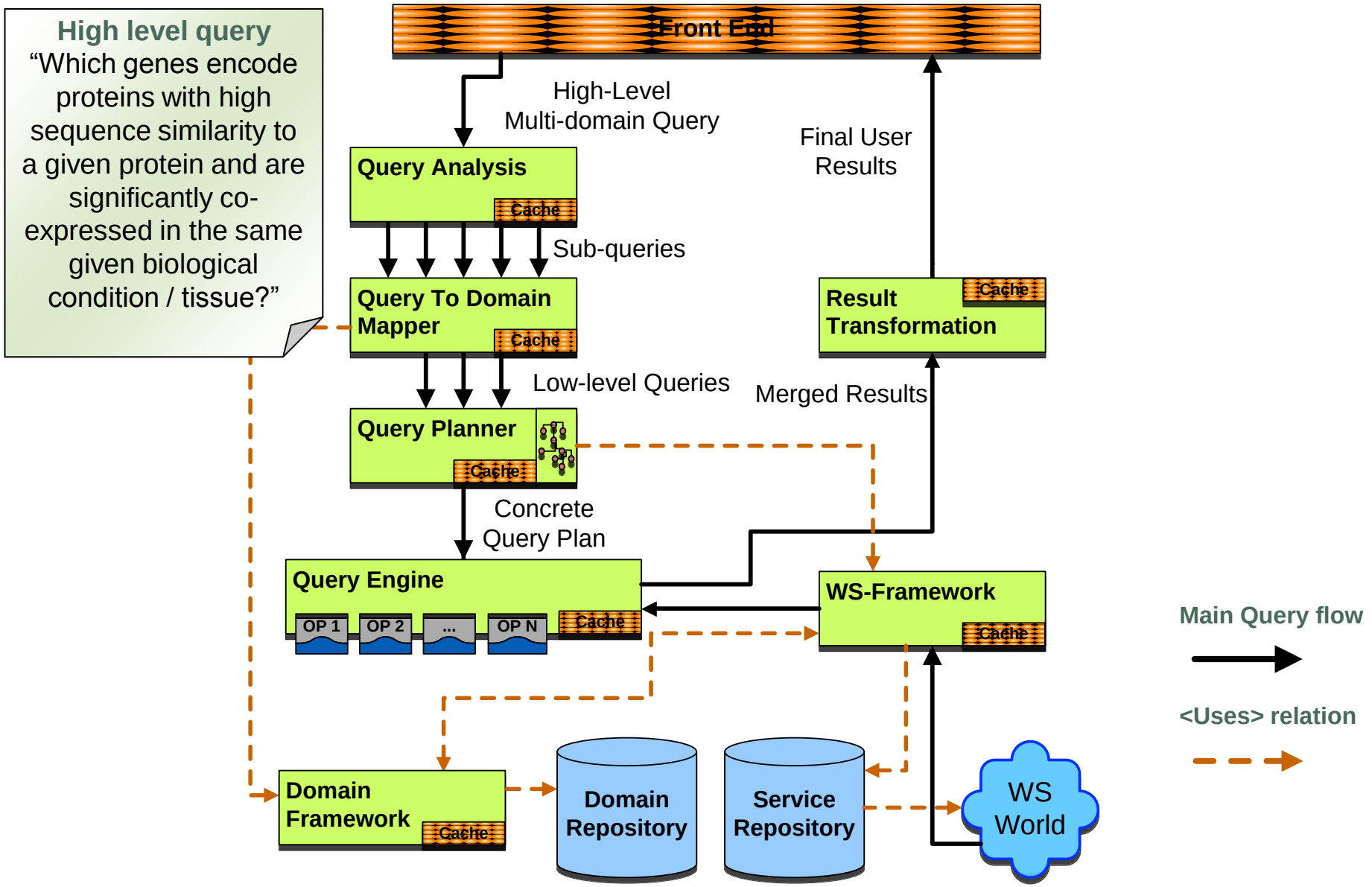
Results of first two year project and two SeCo workshop are collected in two Springer books (http://www.search-computing.it/book.html)

- Ceri S, Brambilla M, editors. *Search Computing - Challenges and Directions*. Heidelberg, D: Springer; 2010. p. 291-306. (Lecture Notes in Computer Science; vol 5950)
- Ceri S, Brambilla M, editors. *New Trends in Search Computing*. Heidelberg, D: Springer; 2011. p. 203-214. (Lecture Notes in Computer Science; vol 6585)

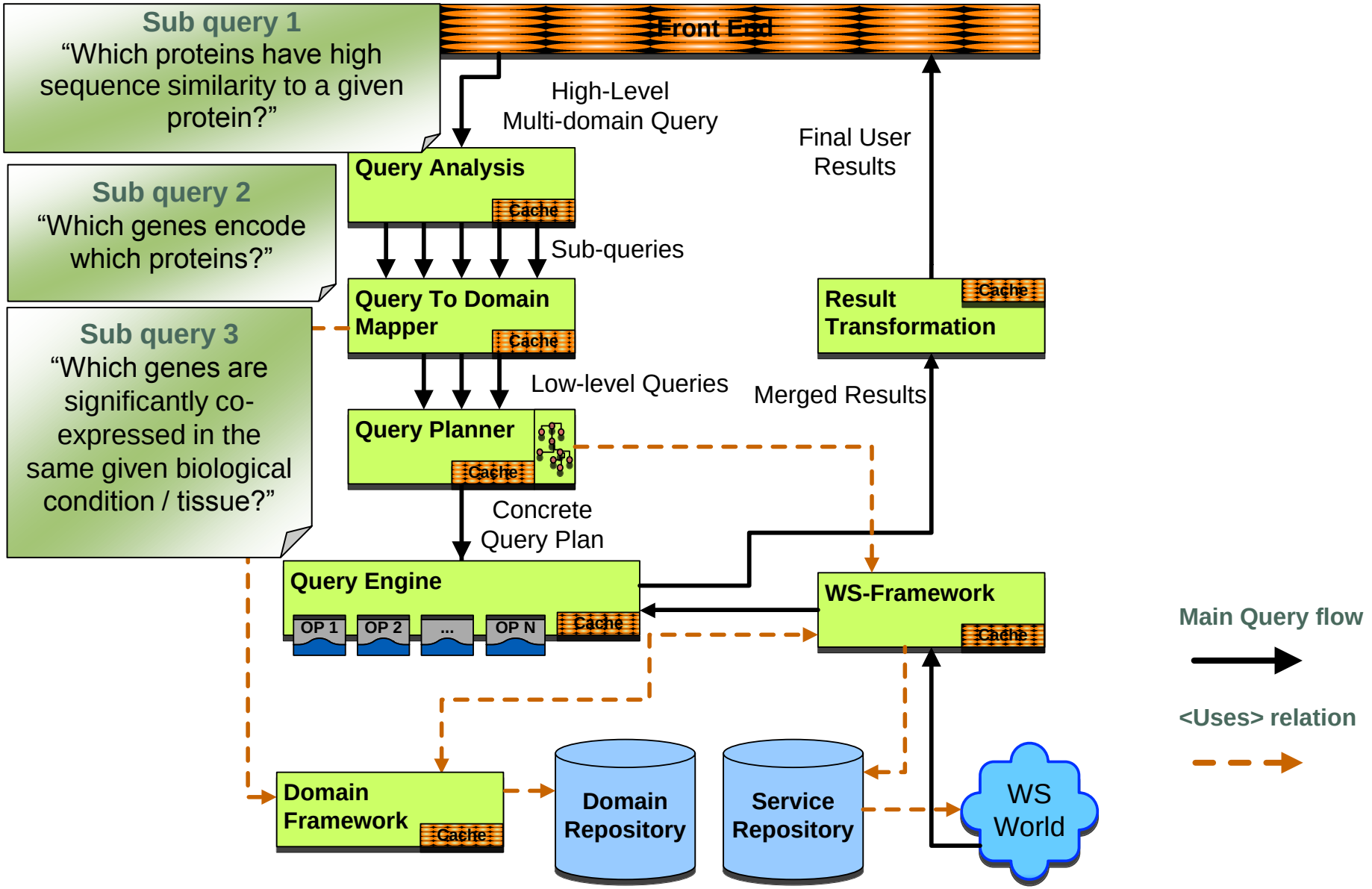
Search Computing framework



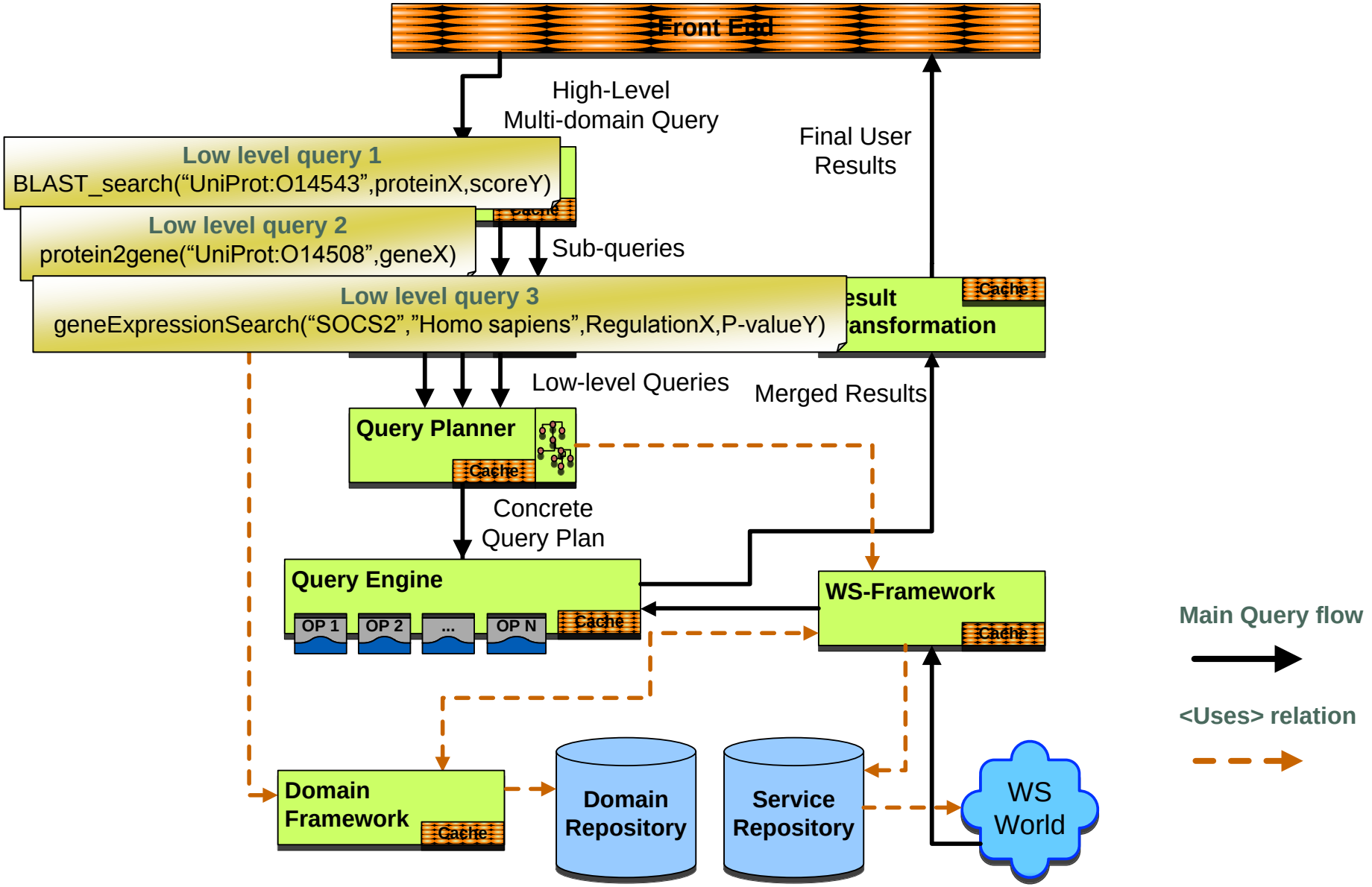
Search Computing framework



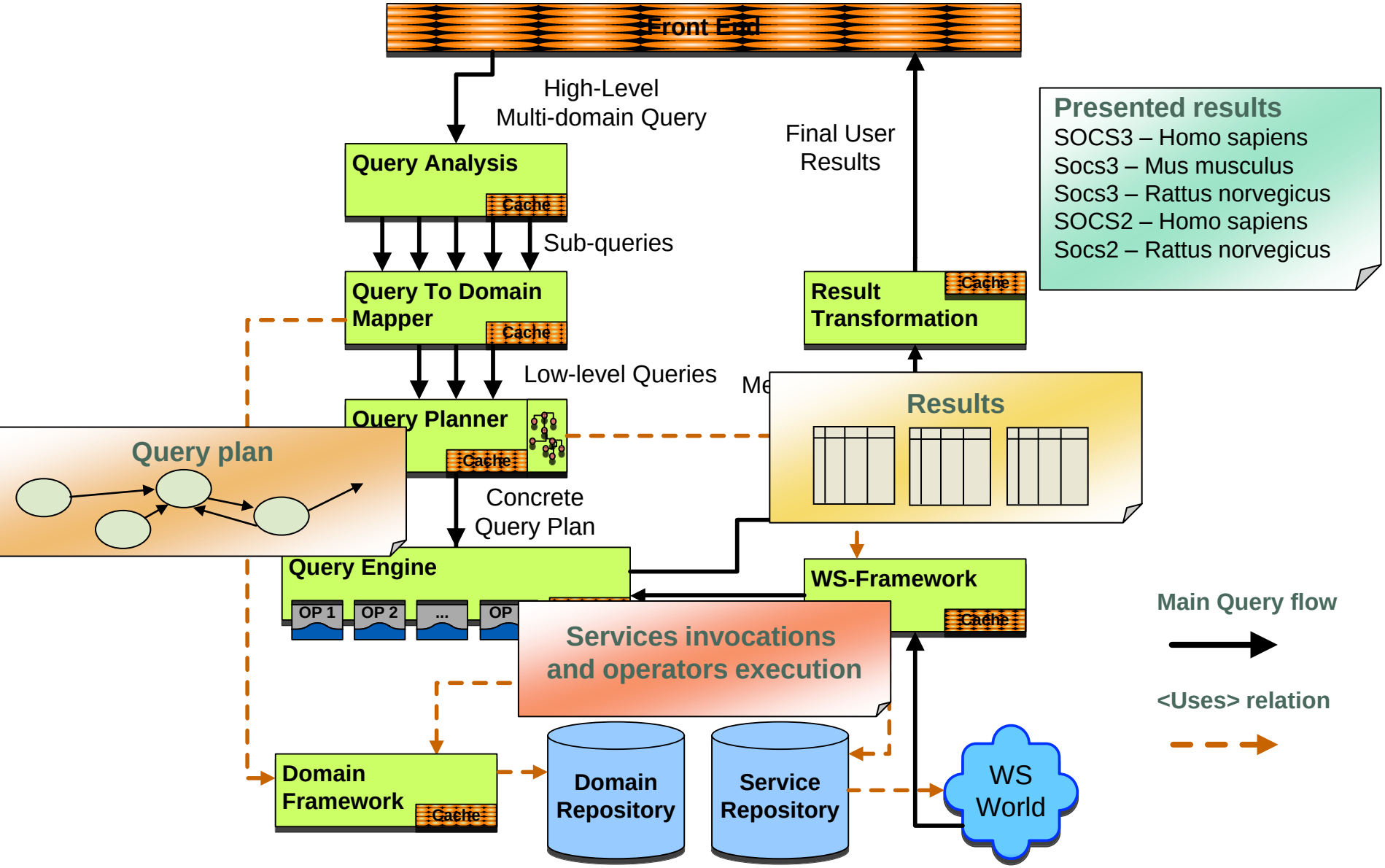
Search Computing framework



Search Computing framework



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[R], …)

Corresponding SM attributes

Auxiliary attributes (i.e. query attributes)

Physical Level: Service interfaces

Selector

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

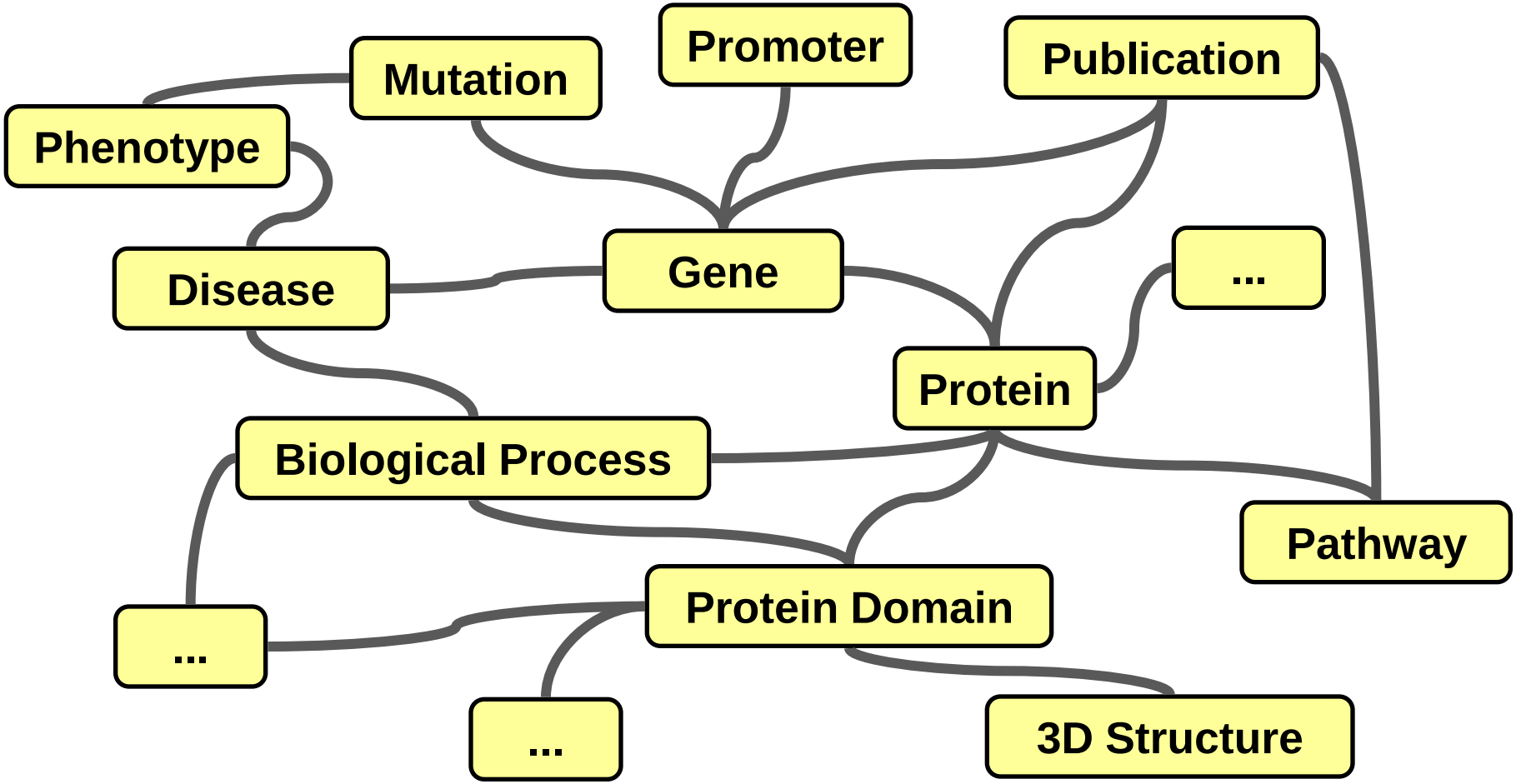
Selector attributes

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



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/)

Bio-SeCo: SeCo technologies for Life Sciences

- According to the Search Computing framework each search service can be modelled with:
  - a **Service Mart** (SM)
  - one or more **Access Patterns** (AP)
  - a **Service Interface** (SI)

WU-BLAST
- 1 Service Mart
- 2 Access Patterns
- 1 Service Interface

GPDW_Gene2Protein
- 1 Service Mart
- 1 Access Patterns
- 1 Service Interface

ArrayExpress
- 1 Service Mart
- 2 Access Patterns
- 1 Service Interface

Bio-SeCo: Sequence alignment search

Alignment	DB:ID	Source	Length	Score	Identity%	Positives%	E()
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 6506607.	1992	9960	100	100	0.
6	EM_PAT:EA062820	Sequence 645 from patent US 7171311.	1992	9960	100	100	0.
7	EM_PAT:EA248485	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:CS063065	Sequence 49 from Patent EP1522594.	1994	9895	99	99	0.
13	EM_PAT:CS080846	Sequence 49 from Patent WO2005040414.	1994	9895	99	99	0.
14	EM_PAT:DD387278	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.

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

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})

POLITECNICO DI MILANO

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/)

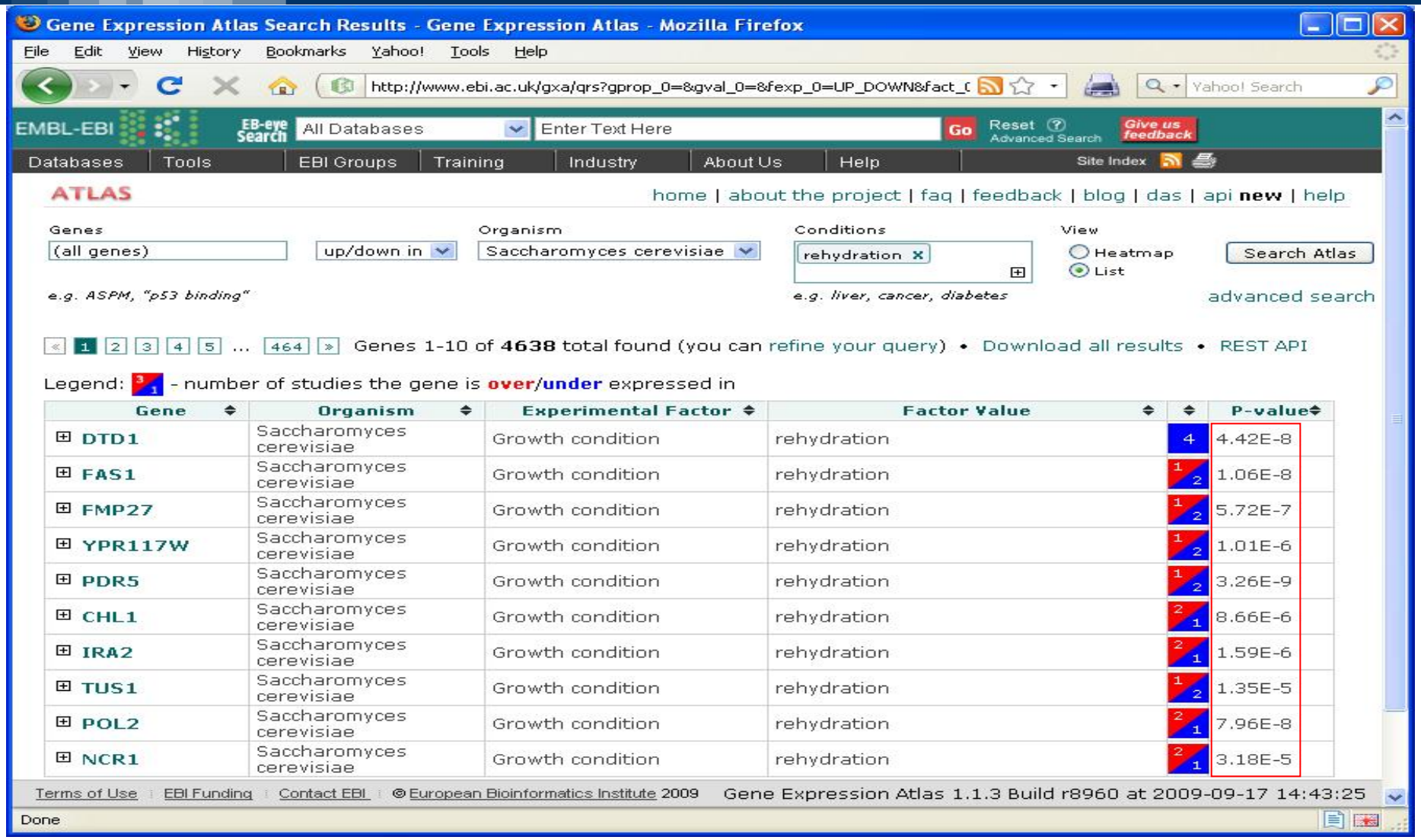
Input example:

- proteinID: O35718
- proteinIDName: uniprot

Output example:

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

Bio-SeCo: Gene expression search



Gene expression data result from *Array Express*

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})

POLITECNICO DI MILANO

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)]

Bio-SeCo: Service workflow

POLITECNICO DI MILANO

WU - BLAST

		Ranking	pvalue
Protein ID	P26367	Sequence Align Program	blastp
Protein ID Name	uniprot	Searched DB	uniprotkb
		Align Sort Criterion	pvalue

Exists Coding Gene
(on Protein ID Name, Protein ID)

GPDW_gene2protein

Exists Expressed Gene
(on Gene Symbol, Organism)

Array Express

		Ranking	pvalue
Factor term	Disease state	Regulatiuon	up
Factor Value	tumor		

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)?"

Query Parameters

Protein ID name	uniprot
Protein ID	O14543
Gene expression regulation	updown
Biological tissue or condition	brain

Visualization Options

Visualization Type	Table View

Search reset

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

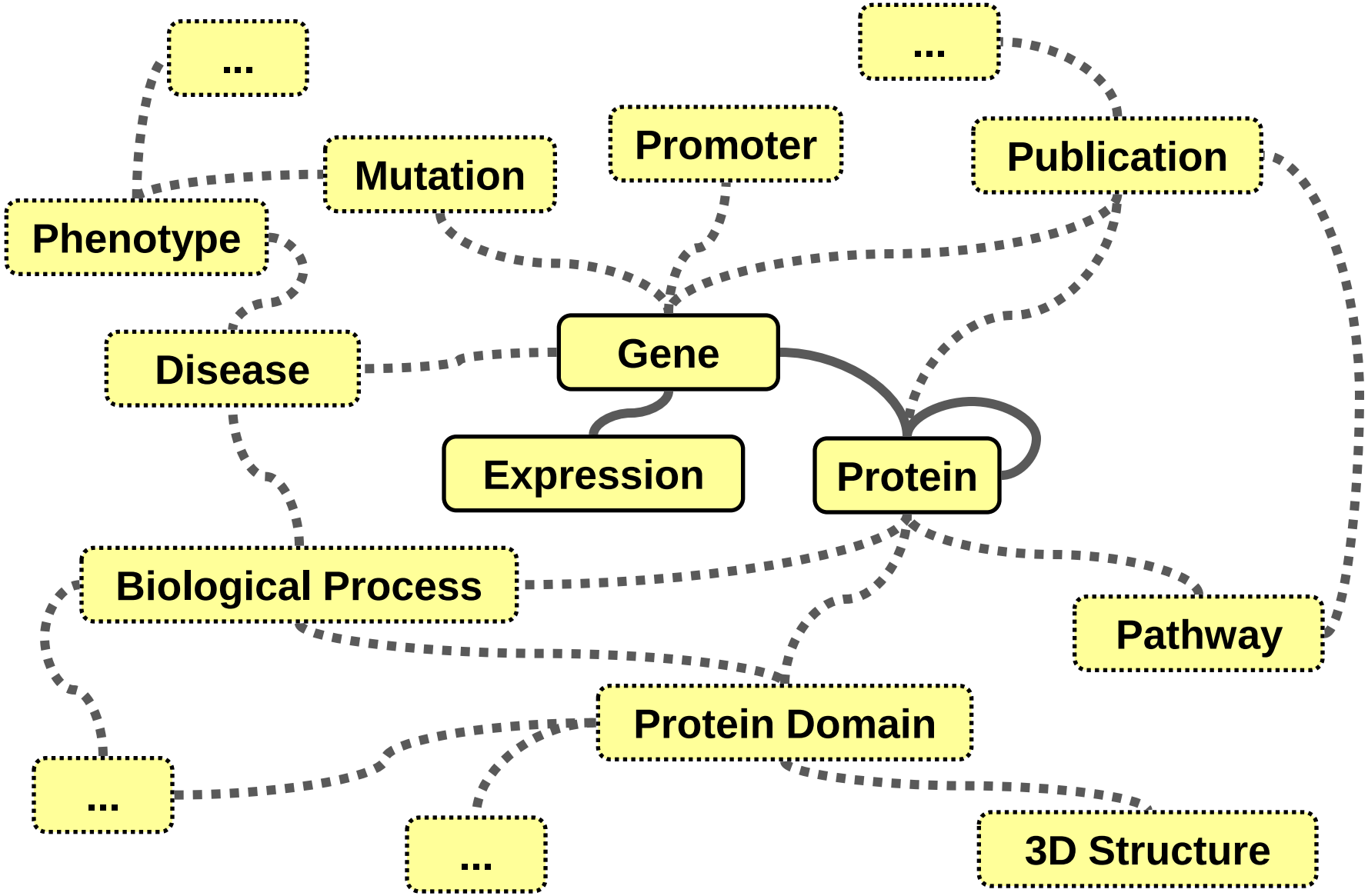
Combined search results

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/

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 ...

Search Computing

The Search Computing Project

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

Menu

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

SeCo on LinkedIn

my LinkedIn profile

SeCo on Twitter

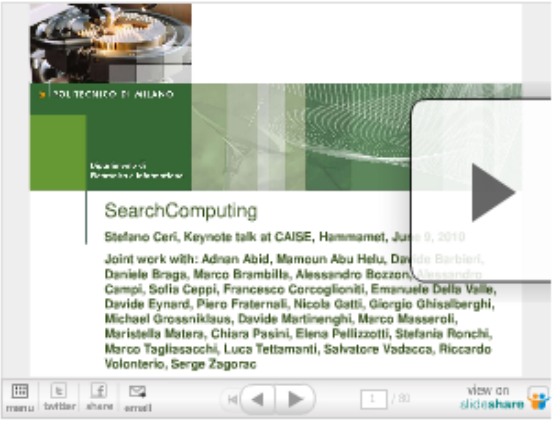
Search Computing

Search Computing (Seco) is a project funded by the European Research Council (ERC), responding to the 2008 Call for "IDEAS Advanced Grants", a program dedicated to the support of investigation-driven frontier research.

The Challenge | The Book | The Demonstrators | The Workshop | Materials | The Blog

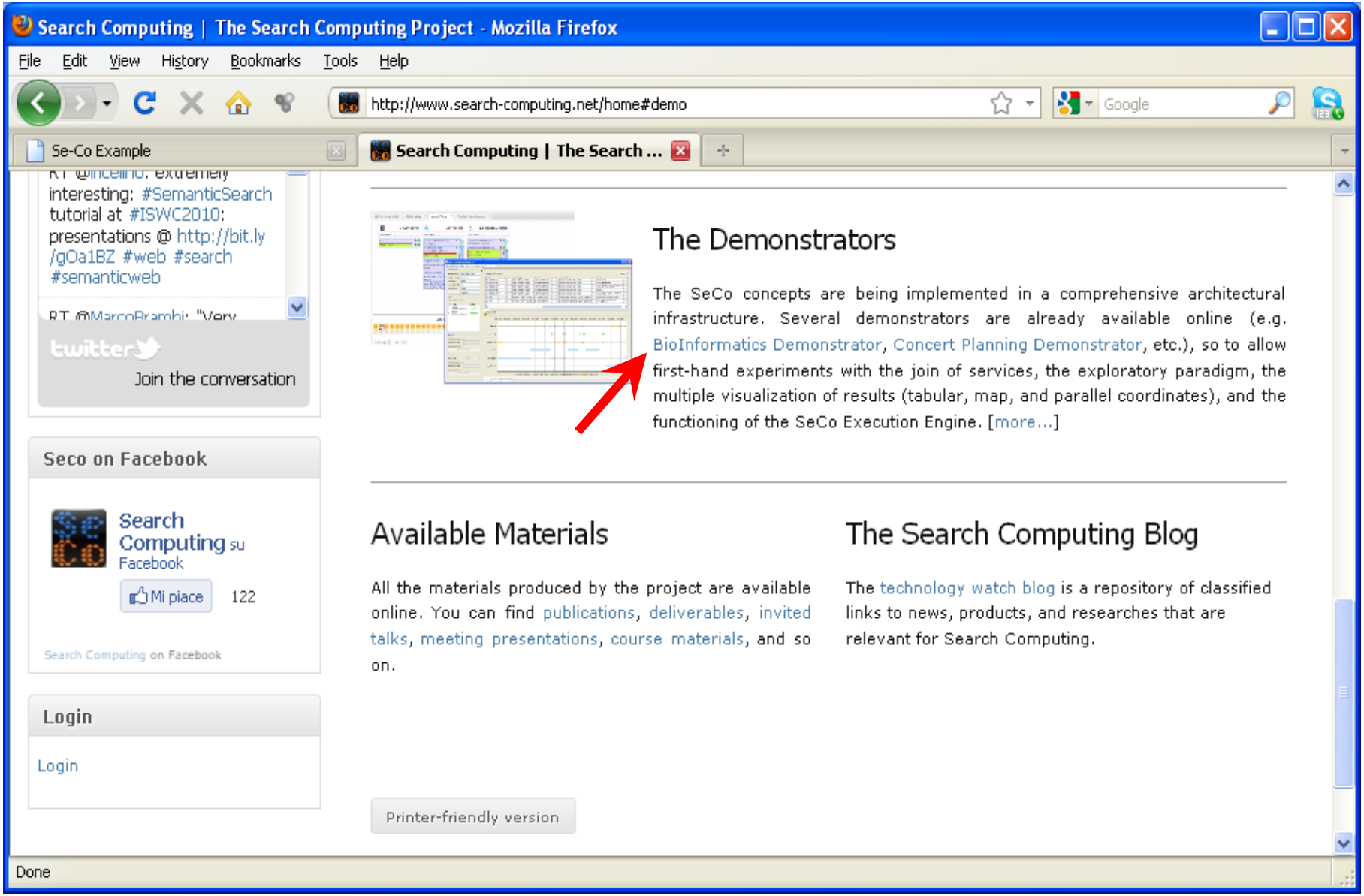
The Challenge

Search computing focuses on building the answers to complex search queries like "Where can I attend an interesting conference in my field close to a sunny beach?" by interacting with a constellation of cooperating search services, using ranking and joining of results as the dominant factors for service composition. By leveraging the peculiar features of search services, the project devises query approaches, execution plans, plan optimization techniques, query configuration tools, and exploratory user interfaces. [more...].



Done

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



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 | ...

Search Computing

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.

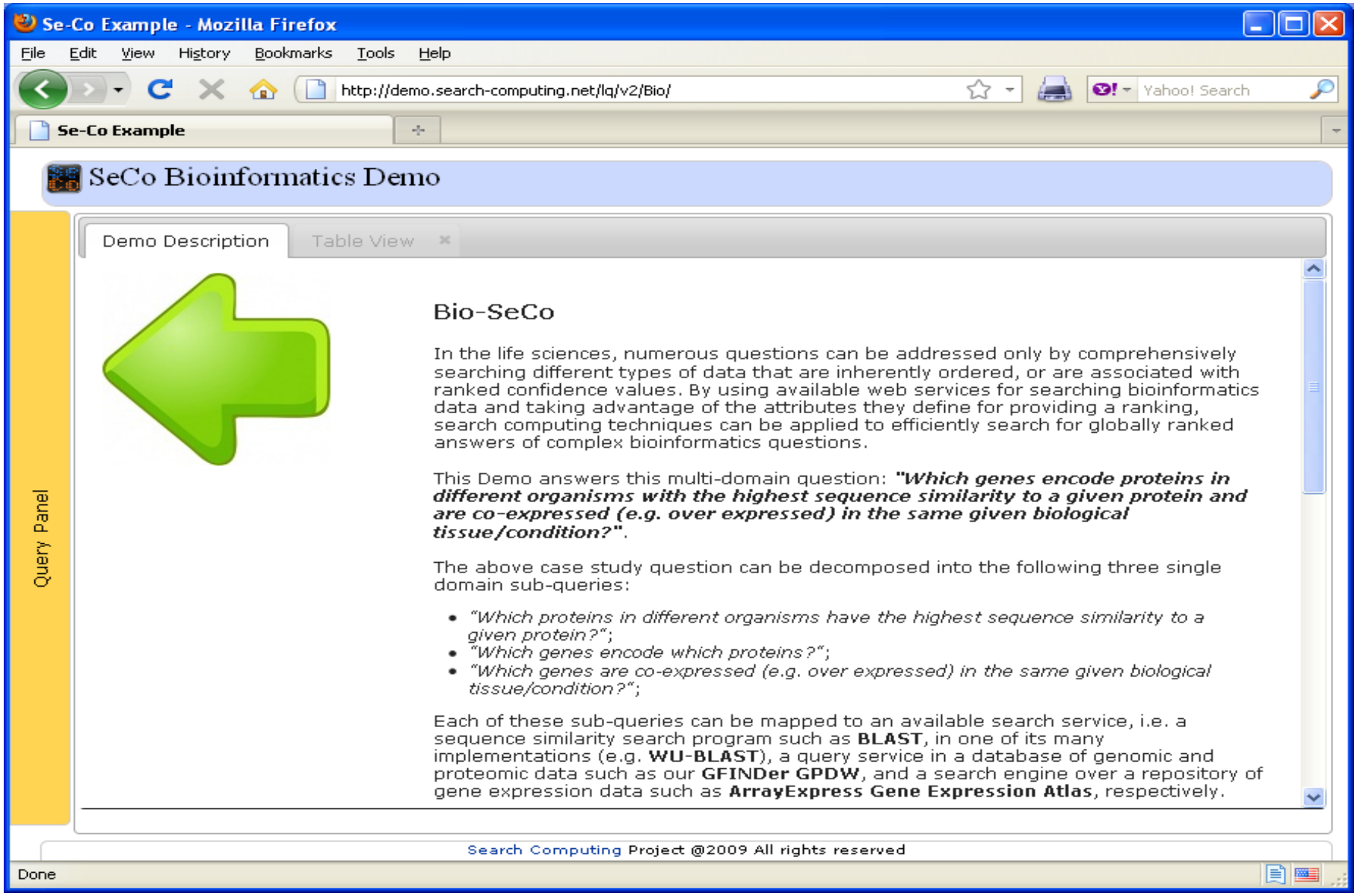
Search computing provides a platform for expressing requests over multiple search services, such that the results of the integrated requests take account of the rankings of individual search results.

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.

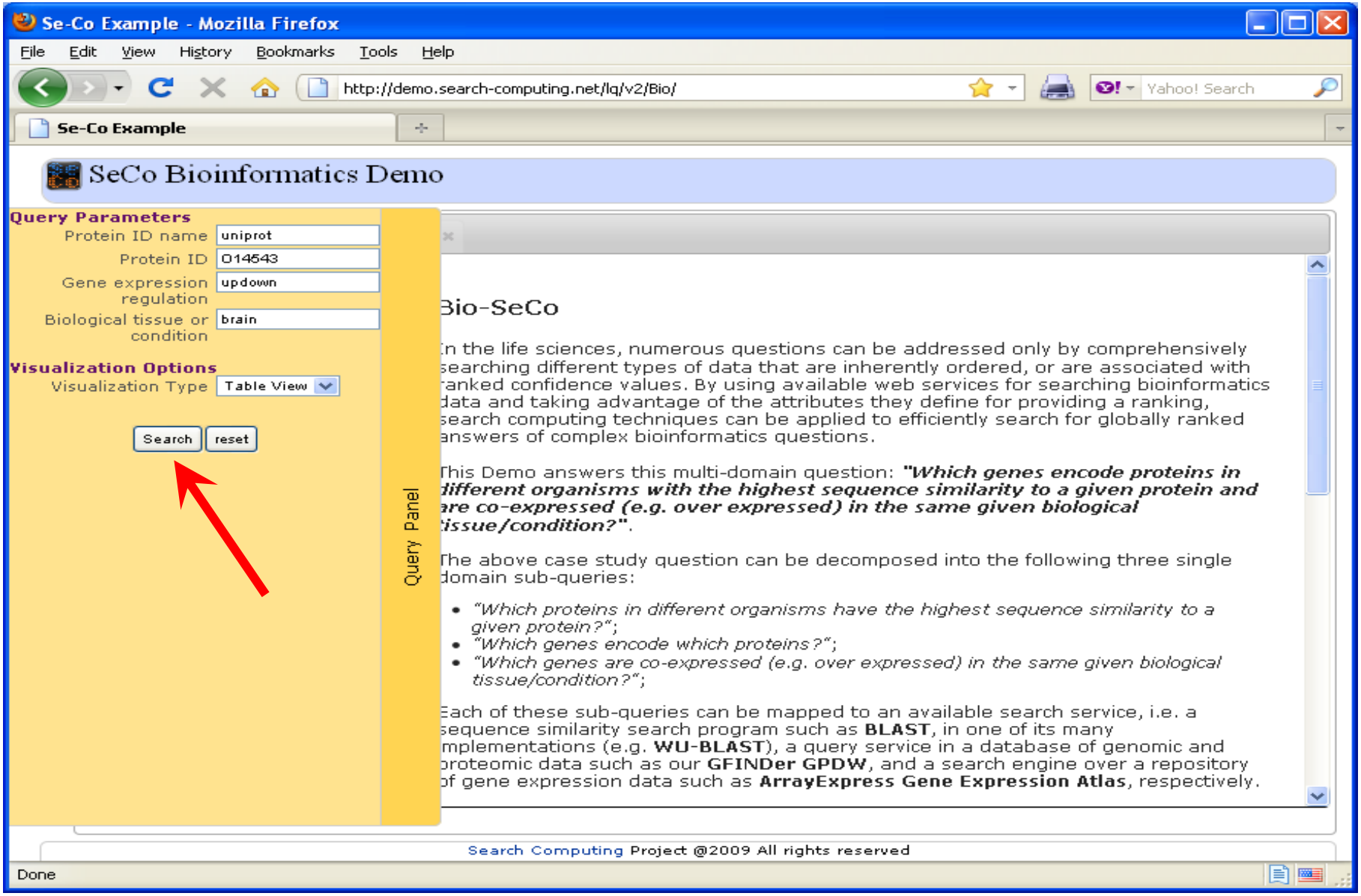
This demo explores the application of a search computing platform in a bioinformatics use case, with a view to identifying the extent to which the existing platform for multi-domain search provides useful facilities for representing and integrating bioinformatics search services.

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/

POLITECNICO DI MILANO

Se-Co Example - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://demo.search-computing.net/lq/v2/Bio/

Se-Co Example

SeCo Bioinformatics Demo

Query Panel

Demo Description | Table View

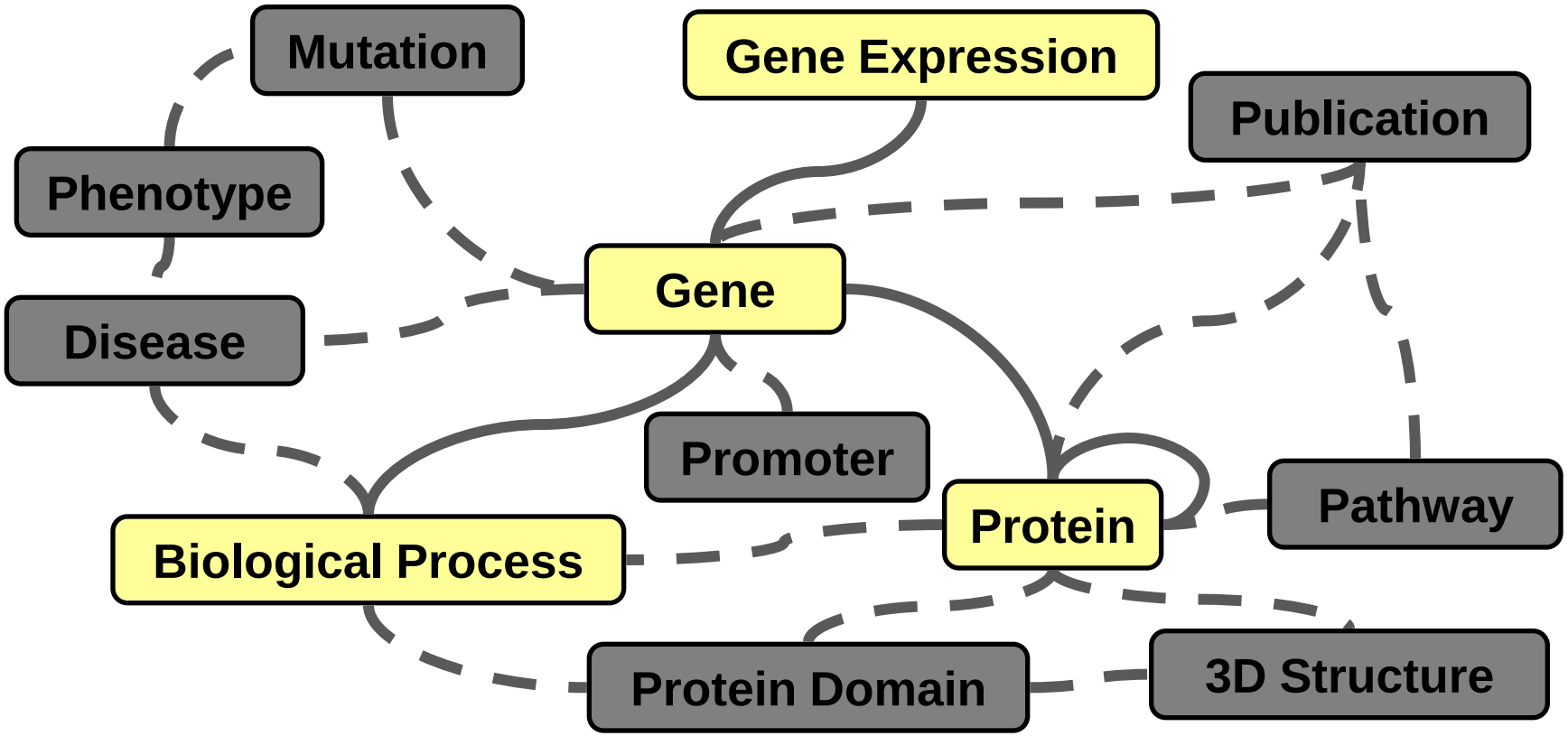
The server retrieved 42 combinations... Get More... Show Columns Group Results

Showing 1 to 23 of 23 entries (filtered from 42 total entries) First Previous 1 Next Last Search: brain

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

Transferring data from demo.search-computing.net...

Resource network of Life Sciences web services



Extended query

- Which are the genes (if they exist) that encode proteins in different organisms with high sequence similarity to an amino acid sequence X and have some biomedical features in common (e.g. significantly co-expressed in the same biological tissue or condition Y and involved in a biological process Z)?

- A new registered service
  - Gene2BiologicalFunction

Bio-SeCo: Biolgical function feature

Service mart

biological_function_feature(geneID, geneIDName, biologicalFunctionFeatureID, biologicalFunctionFeatureIDName, biologicalFunctionFeatureProvenance, biologicalFunctionFeatureName, biologicalFunctionFeatureDefinition)

Ex. Access pattern

biological_function_feature-name_byGeneID(geneID^{I}, geneIDName^{I}, biologicalFunctionFeatureIDName^{I}, biologicalFunctionFeatureName^{I}, biologicalFunctionFeatureFeatureName^{I}, biologicalFunctionFeatureID^{O}, biologicalFunctionFeatureIDName^{O}, biologicalFunctionFeatureName^{O}, biologicalFunctionFeatureDefinition^{O})

Service interface

GPDW_biological_function_feature-name_byGeneID("Genomic and Proteomic Data Warehouse", biological_function_feature-name_byGeneID, http://www.bioinformatics.polimi.it/GFINDer/)

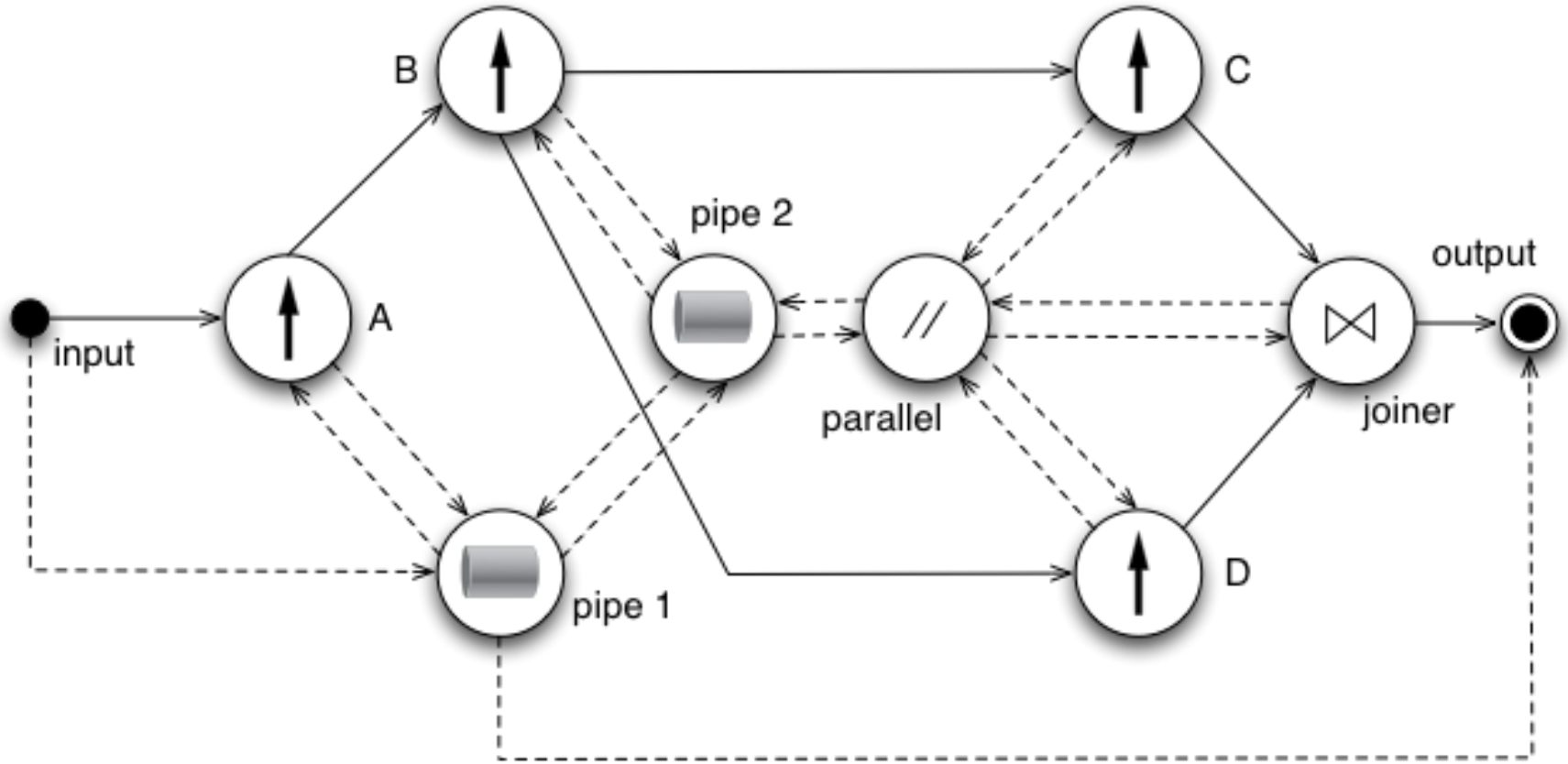
Input example:

- geneID: 5080
- geneIDName: entrez_gene

Output example:

- biologicalFunctionFeatureID: GO:0019222
- biologicalFunctionFeatureIDName: go
- biologicalFunctionFeatureName: "regulation of metabolic process"
- biologicalFunctionFeatureDefinition: "Any process that modulates the frequency, rate or extent of the chemical reactions and pathways within a cell or an organism"

Query plans



A: SequenceAlignment search service

B: Protein2gene service

C: GeneExpression search service

D: Gene2BiologicalFunction service

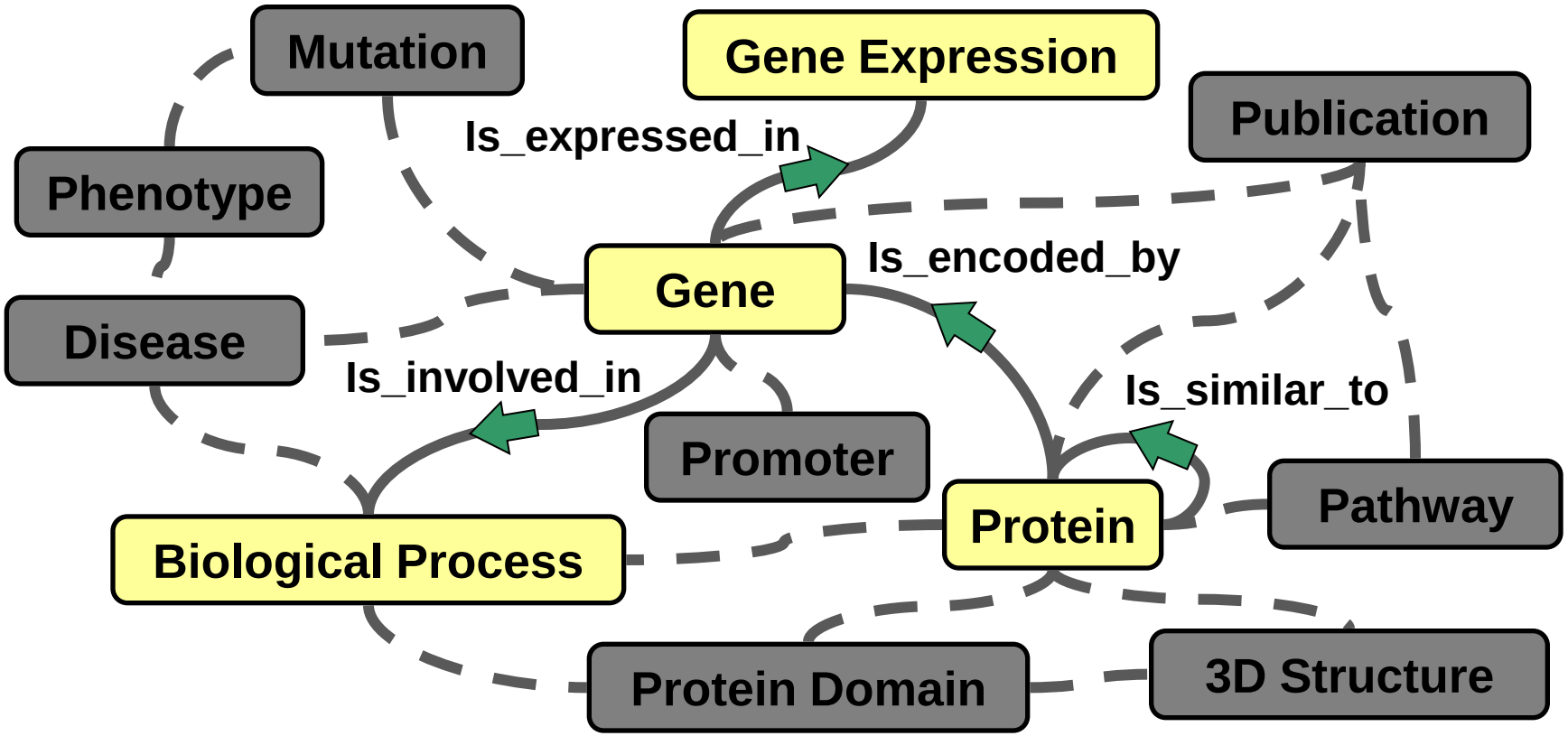
Extended query results

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

Unpredictably, on July 10th 2011, our system discovered only the human *PAX7* gene, with global rank 5.62 E^{-165}, as encoding the *Paired box protein Pax-7* with expectation 1.78 E^{-66} of sequence similarity to the input human protein *Paired box protein Pax-6 isoform a* and with *p-value* 1.0 E^{-100} of expression in tumor

Resource network exploration approach



Thank you for your attention!

Any question?