NETTAB 2011 workshop
Clinical Bioinformatics
October 12-14, 2011, Pavia, Italy

POLITECNICO DI MILANO
Dipartimento di Elettronica e Informazione

Bio Search Computing: Exploration and Global Ranking of Distributed BioMedical-Molecular Data

Marco Masseroli, Giorgio Ghisalberti, Stefano Ceri
marco.masseroli@polimi.it

Motivating Life Science search examples

POLITECNICO DI MILANO

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

POLITECNICO DI MILANO

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

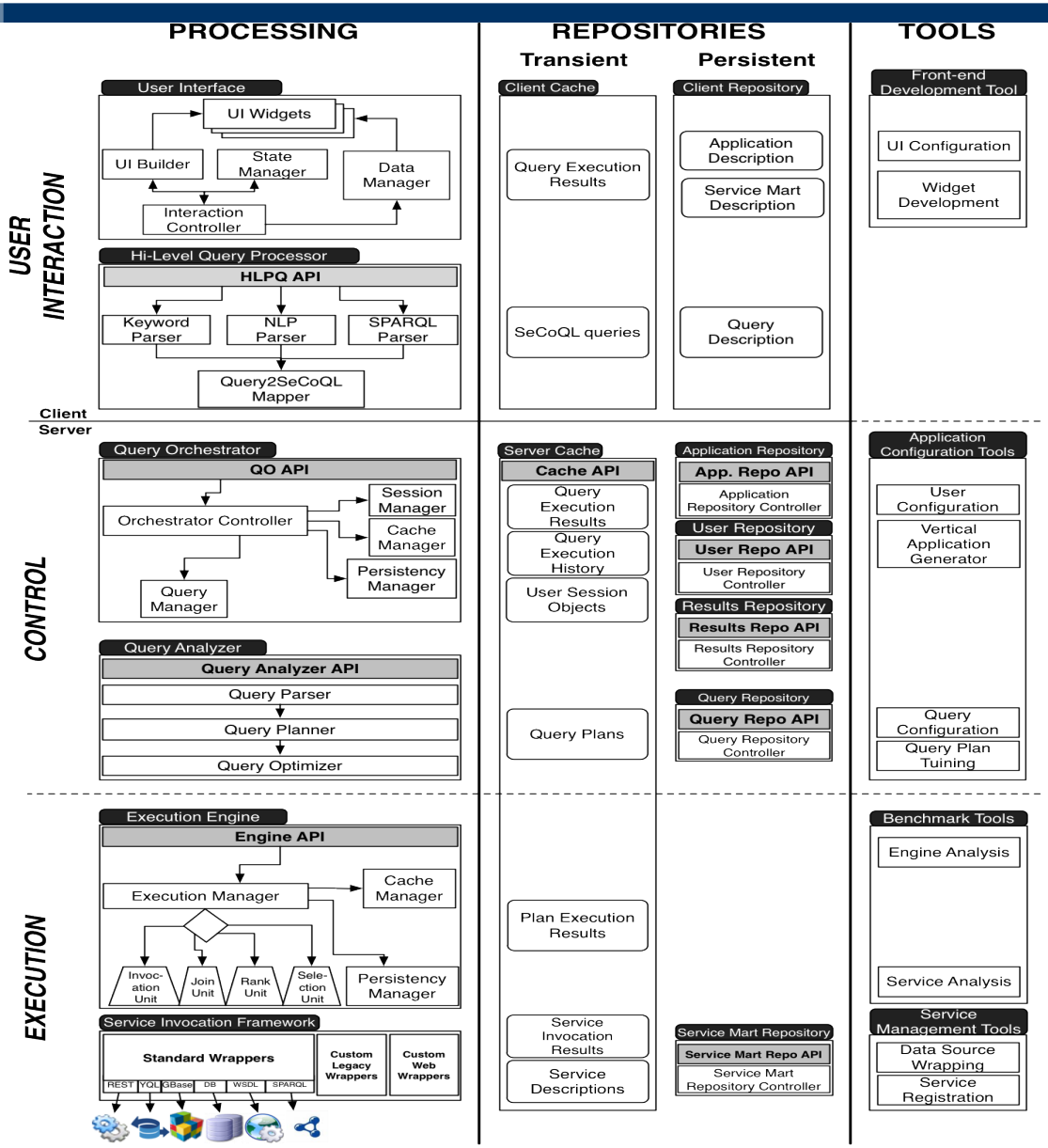
Search Computing (SeCo) is a 5 year project funded in November 2008 by the European Research Council (ERC) Advanced Grant program

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

1. Apply SeCo technology in different fields, including Life Sciences

Search Computing framework



Registration and conceptualization of search services

POLITECNICO DI MILANO

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)

Bio-SeCo: Sequence alignment search

POLITECNICO DI MILANO

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:

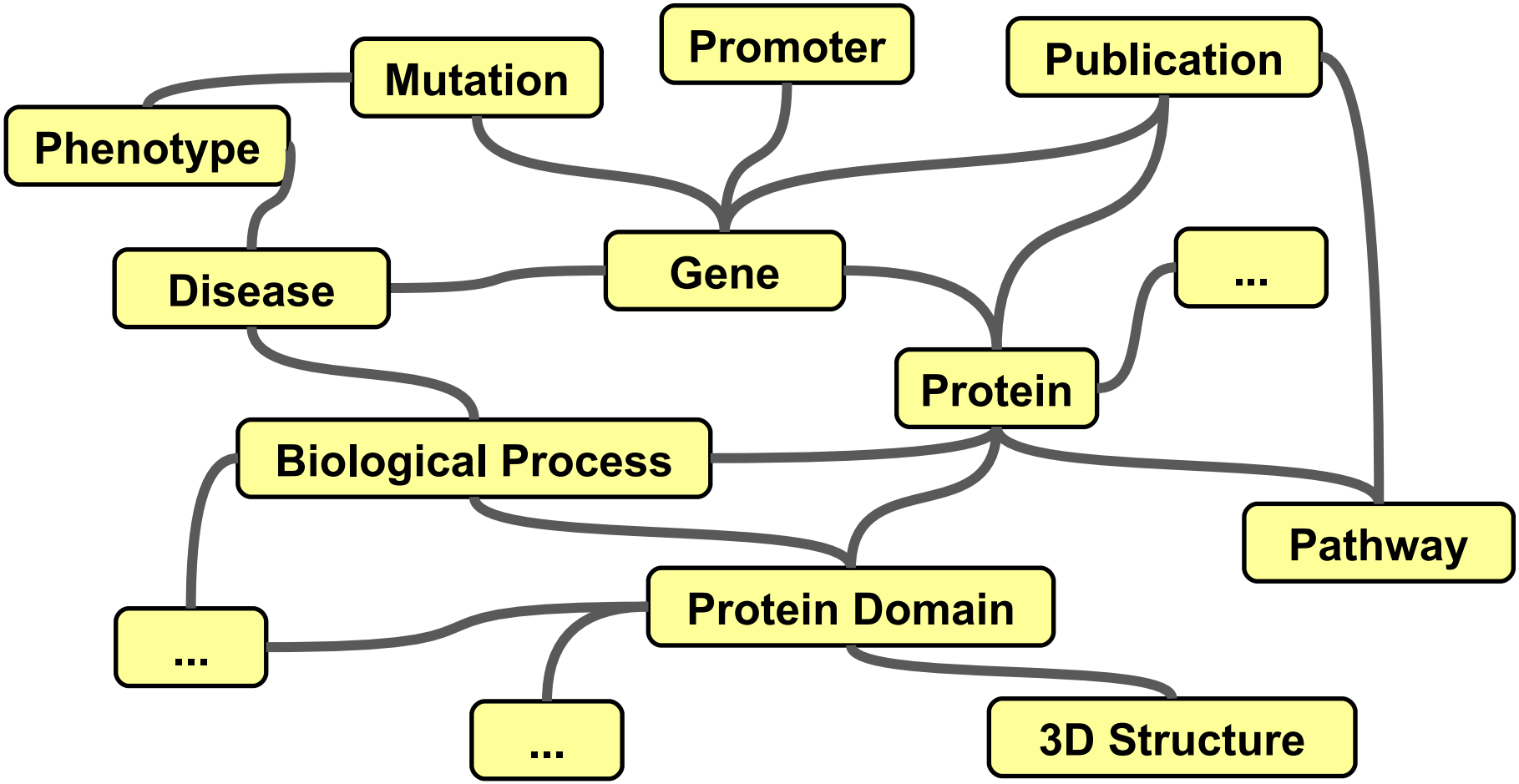
- seaquenchAlignmentProgram: BLASTP
- searchedDB: uniprotKB
- querySequenceID: O14543 •querySequenceIDName: uniprot

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^{-}

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: Connection patterns

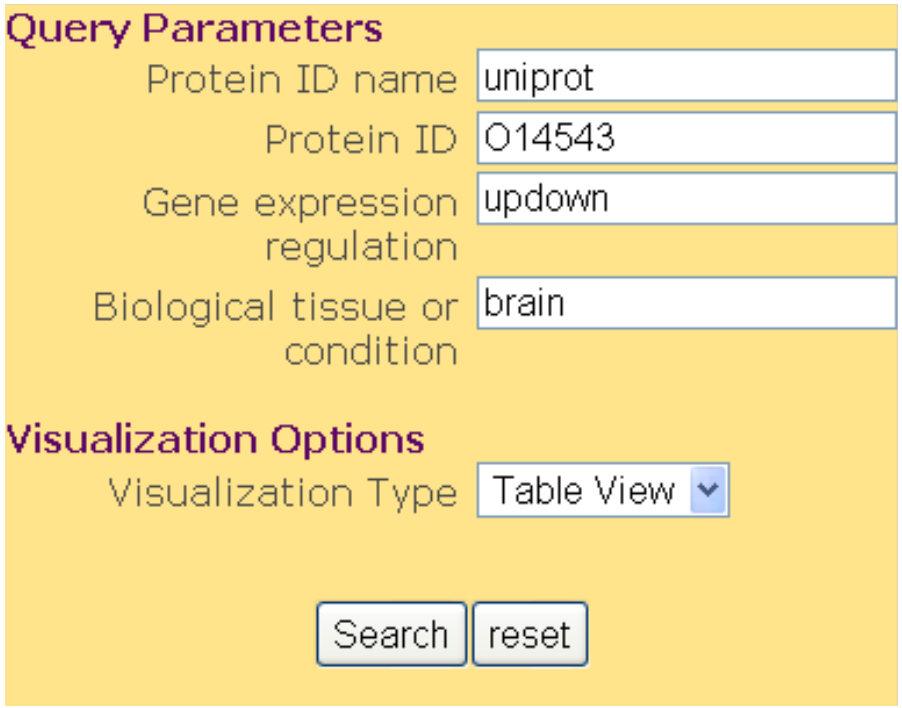
Their pair-wise coupling ***connection patterns*** useful for computing the answer to the first considered case study question are as follows:

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

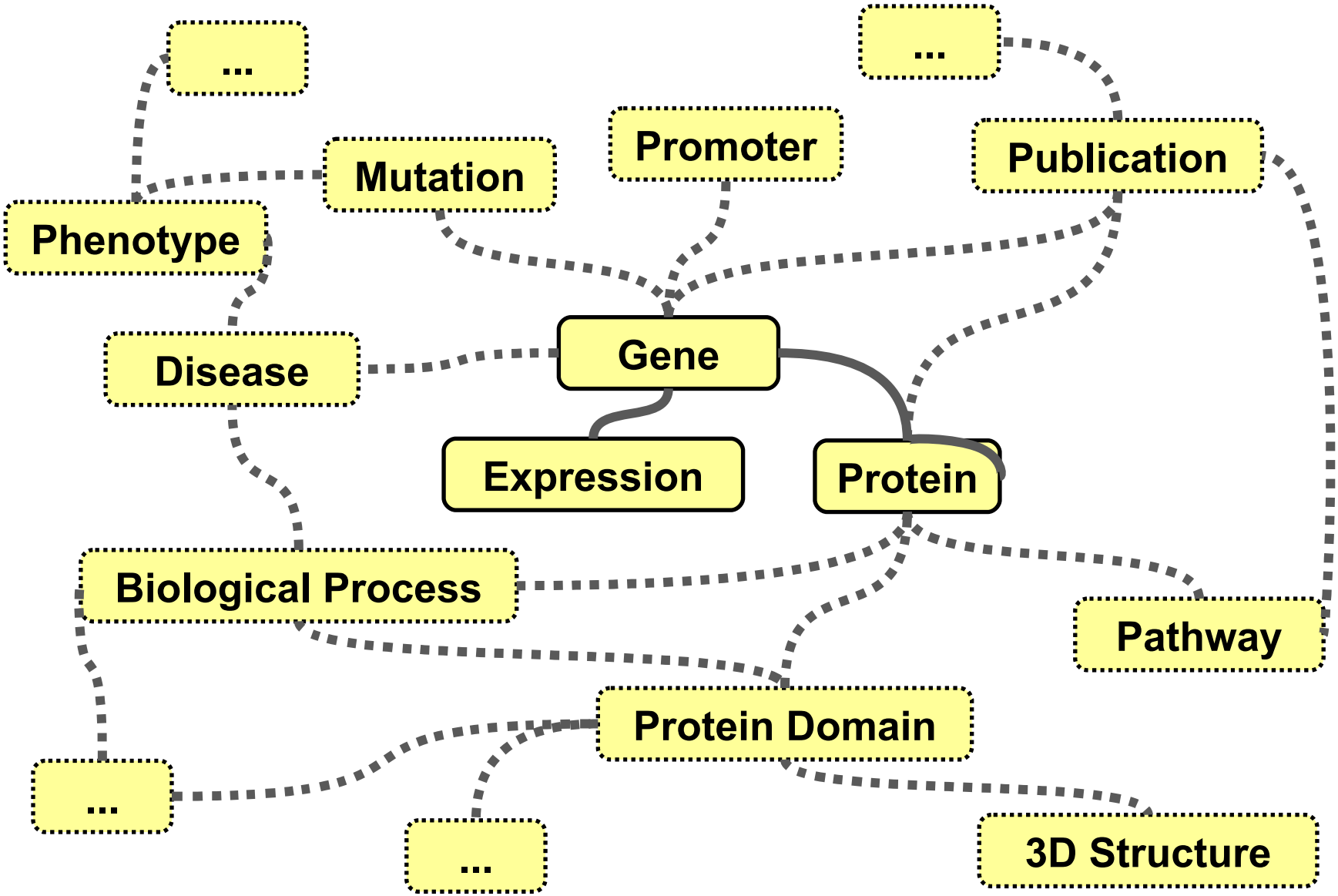
POLITECNICO DI MILANO

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

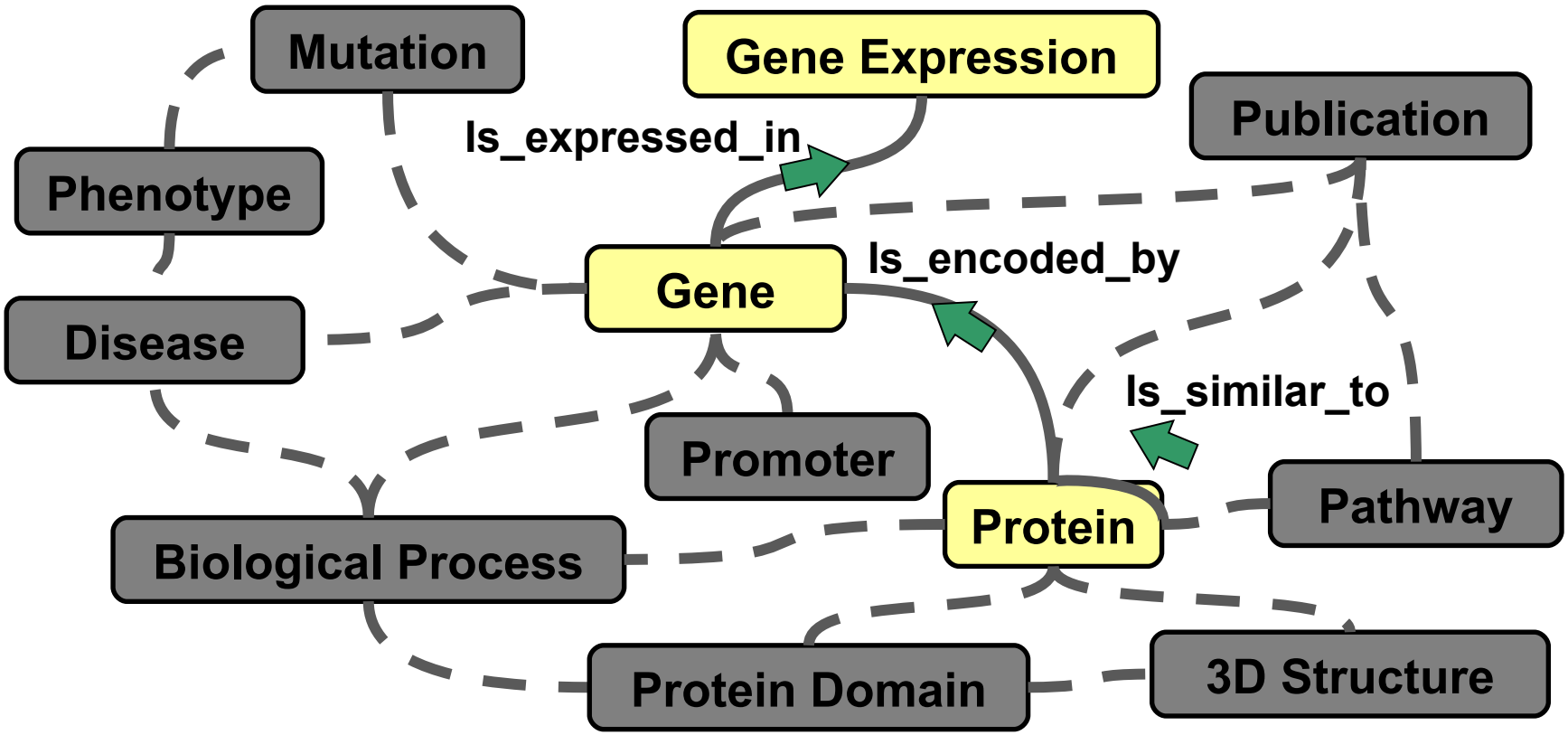
Gene Protein Association		Gene Expression			
Gene Symbol	Organism	Factor	Regulation	Experiment Number	P-value
Socs3	Mus musculus	brain	UP	24	1.1218185040451748e-23
Socs3	Mus musculus	brain	UP	24	1.1218185040451748e-23
SOCS3	Homo sapiens	brain	UP	11	2.5128574776545065e-9
SOCS3	Homo sapiens	brain	UP	11	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



Resource network exploration approach



Thank you for your attention!

Any question?