



NETTAB 2012

Integrated Bio-Search

November 14-16, 2012, Como, Italy



POLITECNICO
DI MILANO

Dipartimento di
Elettronica e Informazione



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



Life Sciences computational and data access web services



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 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	8895	99	99	0.
12	EM_PAT:CS063065	Sequence 49 from Patent EP1522594.	1994	8895	99	99	0.
13	EM_PAT:CS080846	Sequence 49 from Patent WO2005040414.	1994	8895	99	99	0.
14	EM_PAT:DD387278	COMPOSITIONS AND METHODS FOR THE DIAGNOSIS AND TREATMENT OF TUMOR.	1994	8895	99	99	0.
15	EM_PAT:DL464877	COMPOSITIONS, KITS, AND METHODS FOR IDENTIFICATION, ASSESSMENT, PREVENTION, AND THERAPY OF CANCER.	1994	8895	99	99	0.
16	EM_PAT:FB671589	Sequence 49 from Patent EP1892306.	1994	8895	99	99	0.

Gene Expression Atlas Search Results - Gene Expression Atlas - Mozilla Firefox

EMBL-EBI 18-se Search All Databases Enter Text Here Go Reset Advanced Search Give us feedback

Databases Tools EBI Groups Training Industry About Us Help Site Index

ATLAS home | about the project | faq | feedback | blog | das | api new | help

Genes (all genes) up/down in Organism Saccharomyces cerevisiae Conditions rehydration X View Heatmap List Search Atlas

e.g. ASPM, "p53 binding" e.g. liver, cancer, diabetes advanced search

Legend: 1 2 3 4 5 ... 464 Genes 1-10 of 4638 total found (you can refine your query) • Download all results • REST API

Legend: 1 2 3 4 5 ... 464 Genes 1-10 of 4638 total found (you can refine your query) • Download all results • REST API

Gene	Organism	Experimental Factor	Factor Value	P-value
DTD1	Saccharomyces cerevisiae	Growth condition	rehydration	4.42E-8
FAS1	Saccharomyces cerevisiae	Growth condition	rehydration	1.06E-8
FMP27	Saccharomyces cerevisiae	Growth condition	rehydration	5.72E-7
YPR117W	Saccharomyces cerevisiae	Growth condition	rehydration	1.01E-6
PDR5	Saccharomyces cerevisiae	Growth condition	rehydration	3.26E-9
CHL1	Saccharomyces cerevisiae	Growth condition	rehydration	8.66E-6
IRA2	Saccharomyces cerevisiae	Growth condition	rehydration	1.59E-6
TUS1	Saccharomyces cerevisiae	Growth condition	rehydration	1.35E-5
POL2	Saccharomyces cerevisiae	Growth condition	rehydration	7.96E-8
NCR1	Saccharomyces cerevisiae	Growth condition	rehydration	3.18E-5

Terms of Use EBI Funding Contact EBI European Bioinformatics Institute 2009 Gene Expression Atlas 1.1.3 Build r8960 at 2009-09-17 14:43:25

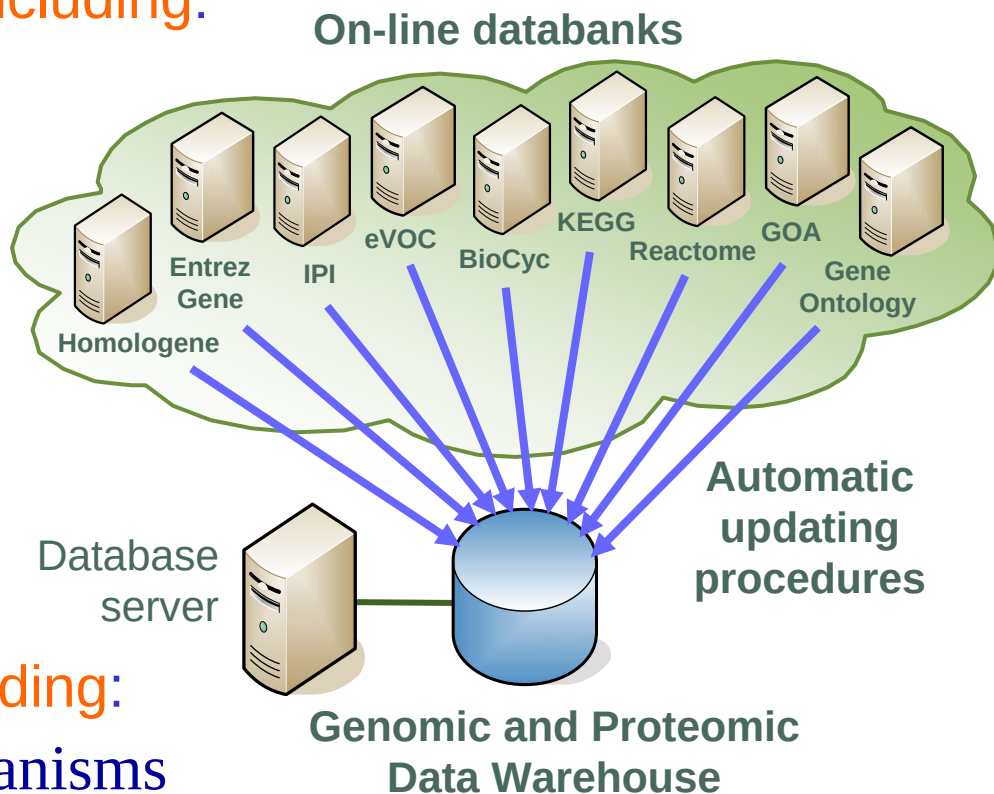
UniProt search result for protein “5-hydroxytryptamine (serotonin) receptor 2A”

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



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 combining 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



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:

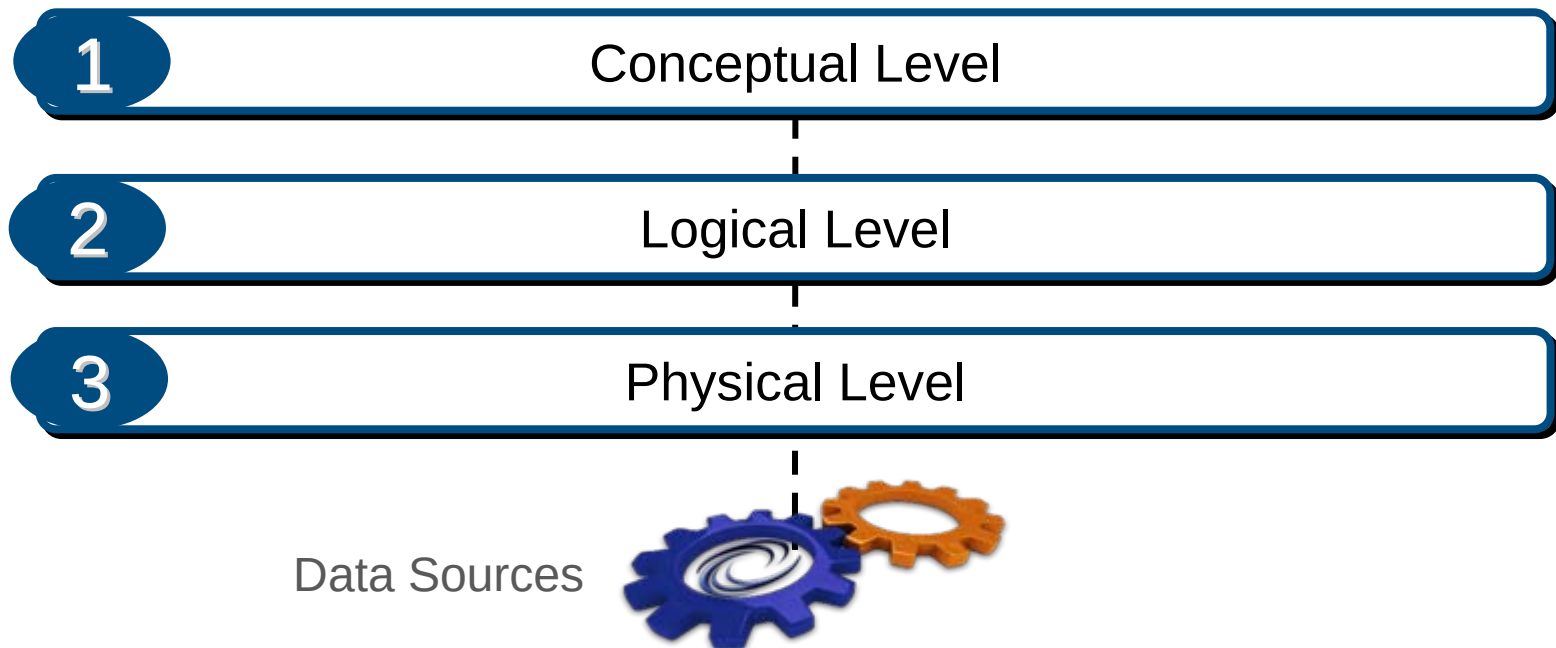


- “Which proteins in different organisms have high sequence similarity to a protein *X* ?”
 - BLAST, a sequence similarity search program, in one of its many implementations, e.g. **WU-BLAST** (<http://www.ebi.ac.uk/blast2/>) or **NCBI-Blast** (<http://blast.ncbi.nlm.nih.gov/Blast>)
- “Which genes encode which proteins ?”
 - GPDW (Genomic and Proteomic Data Warehouse), a query service to a database of genomic and proteomic data (**GPDW_protein2gene**)



- “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**)

- 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**



- For modelling each **service** they are realized:
 - 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

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
- foundSequenceOrganism: Mus musculus
- foundSequenceDescription: Suppressor of citokine signaling 3
- bestAlignmentScore: 990
- bestAlignmentExpectation: 2.99×10^{-98}
- foundSequenceIDName: uniprot
- bestAlignmentProbability: 2.99×10^{-98}



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

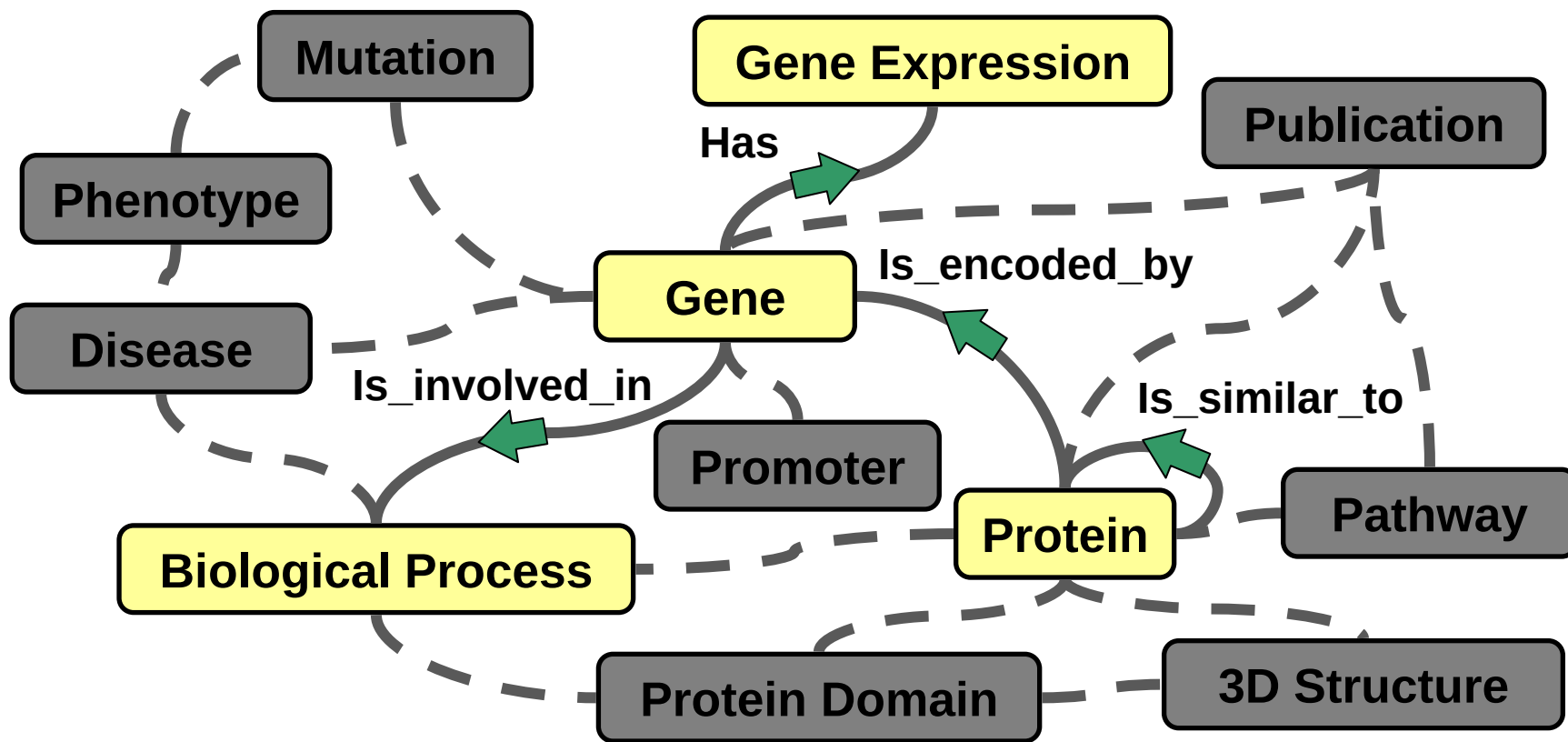
existsCodingGene_byProteinID(sequenceAlignmentSearch, protein2gene):
[(sequenceAlignmentSearch.foundSequenceID = protein2gene.proteinID
AND sequenceAlignmentSearch.foundSequenceIDName =
protein2gene.proteinIDName)]

existsExpressedGene_byGeneID(protein2gene, geneExpressionSearch):
[(protein2gene.geneIDName = “ensembl”
AND geneExpressionSearch.queryEnsemblGeneID = protein2gene.geneID)]

existsBiologicalFunctionFeature-name_byGeneID(protein2gene,
biological_function_feature): [(biological_function_feature.geneID =
protein2gene.geneID
AND biological_function_feature.geneIDName = protein2gene.geneIDName)]

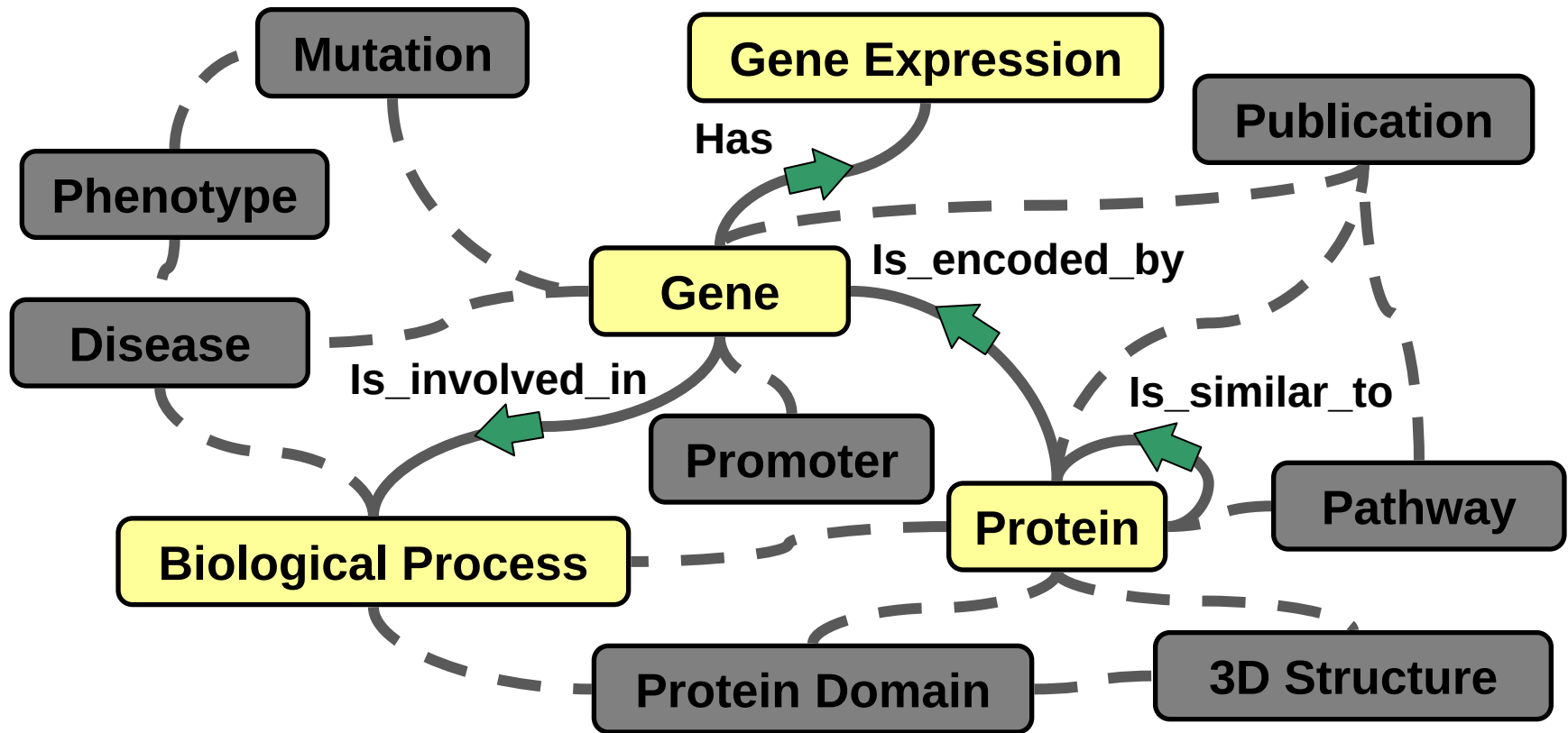


Services registered in the framework are pair-wise related each other through **connection patterns** that define the available **resource network**





Resource network exploration approach through query expansion





Query interface for multi-topic search

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

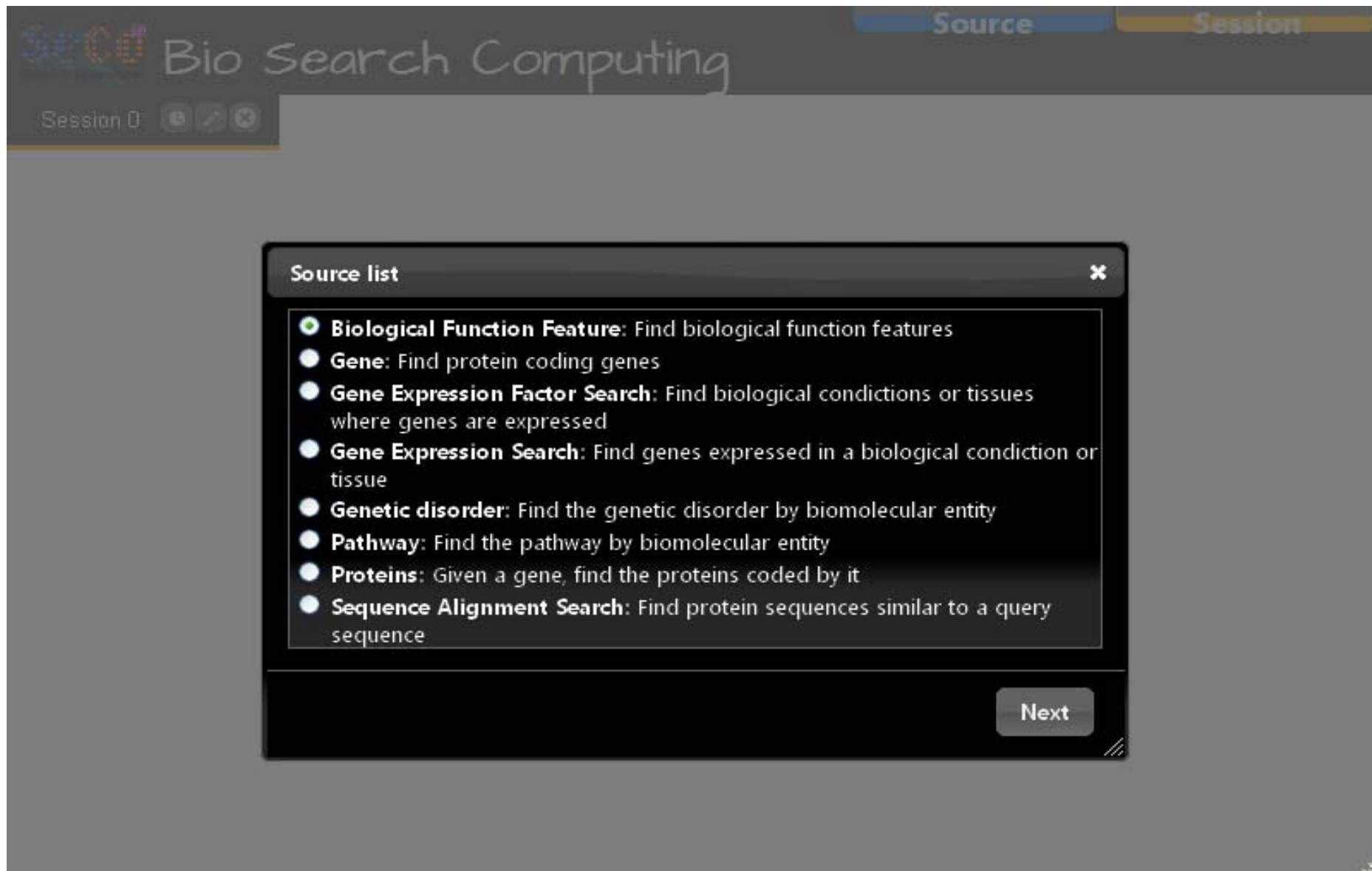


POLITECNICO
DI MILANO



Bio Search Computing

New Query Session

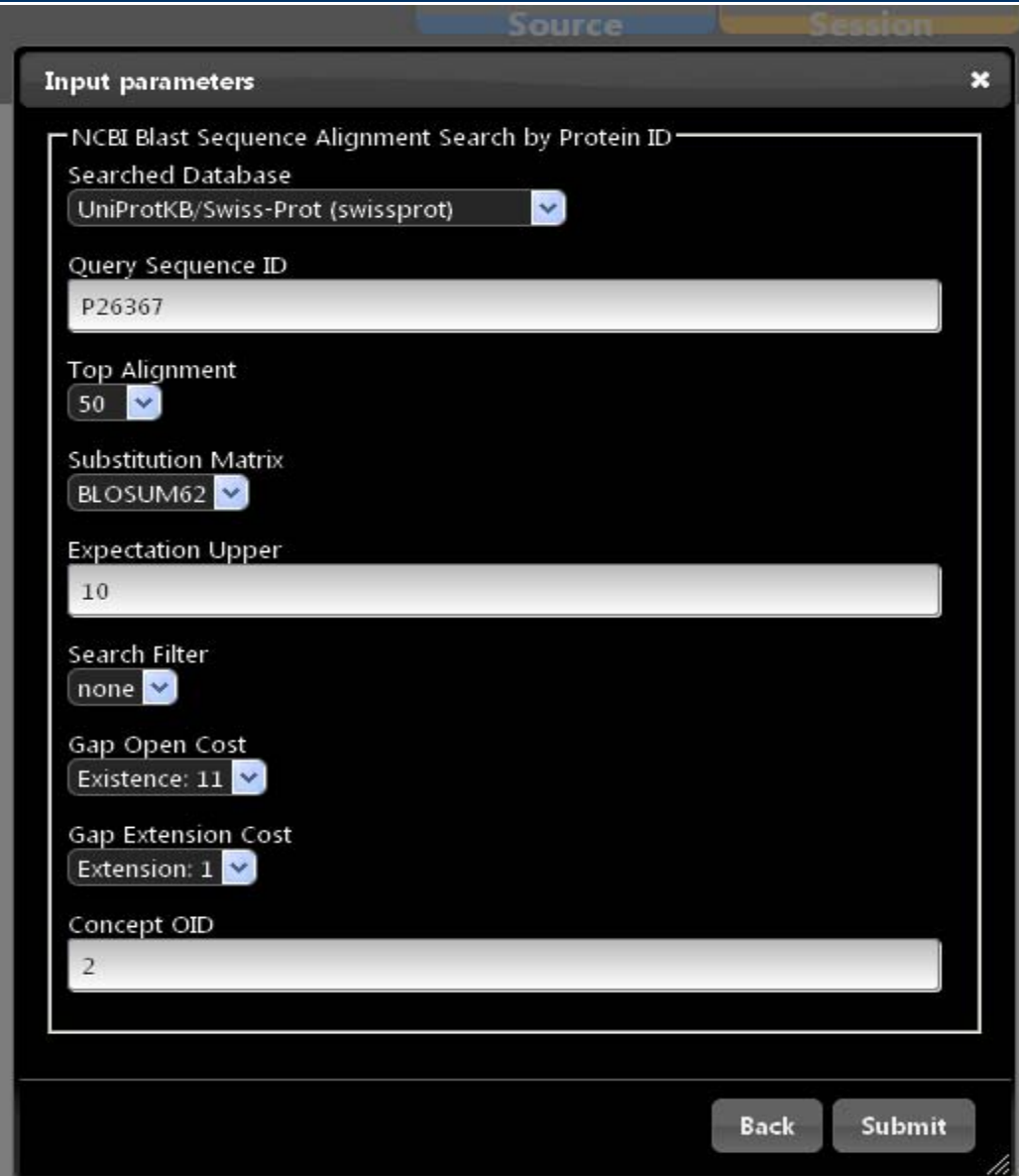




Results of sequence alignment search on NCBI-BLAST

“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**



The screenshot shows the NCBI-BLAST search interface. At the top, there are tabs for 'Source' and 'Session'. Below this is a window titled 'Input parameters' with a close button (X). The window contains the following fields and options:

- NCBI Blast Sequence Alignment Search by Protein ID**
- Searched Database:** UniProtKB/Swiss-Prot (swissprot) (dropdown menu)
- Query Sequence ID:** P26367 (text input field)
- Top Alignment:** 50 (dropdown menu)
- Substitution Matrix:** BLOSUM62 (dropdown menu)
- Expectation Upper:** 10 (text input field)
- Search Filter:** none (dropdown menu)
- Gap Open Cost:** Existence: 11 (dropdown menu)
- Gap Extension Cost:** Extension: 1 (dropdown menu)
- Concept OID:** 2 (text input field)

At the bottom right of the window are two buttons: 'Back' and 'Submit'.



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**

NCBI Blast Sequence Alignment Search by Protein ID									
Global Tuple Data									
	Global Score	Tuple Score	Found Sequence Symbol	Found Sequence ID	Found Sequence ID Name	Found Sequence Length	Found Sequence Description	Best Alignment Expectation	Concept OID
☐	1.00000	1.00000	PAX6_HUMAN	P26367	uniprot	422	paired box protein Pax-6 isoform a [Homo sapiens]	1.0E-300	2
☐	1.00000	1.00000	PAX6_BOVIN	Q1LZF1	uniprot	422	RecName: Full=Paired box protein Pax-6; AltName: Full=Oculorhombin	1.0E-300	2
☐	1.00000	1.00000	PAX6_RAT	P63016	uniprot	422	RecName: Full=Paired box protein Pax-6; AltName: Full=Oculorhombin	1.0E-300	2
☐	1.00000	1.00000	PAX6_COTJA	P47238	uniprot	416	RecName: Full=Paired box protein Pax-6; AltName: Full=Pax-QNR	1.0E-300	2
☐	1.00000	1.00000	PAX6_XENLA	P55864	uniprot	422	RecName: Full=Paired box protein Pax-6	1.0E-300	2
☐	1.00000	1.00000	PAX6_DANRE	P26630	uniprot	437	RecName: Full=Paired box protein Pax-6; AltName: Full=Pax[Zf-a]	1.0E-300	2
☐	1.00000	1.00000	PAX6_ORYLA	O73917	uniprot	437	RecName: Full=Paired box protein Pax-6	1.0E-300	2
☐	0.50638	0.50638	PAX6_CHICK	P47237	uniprot	216	RecName: Full=Paired box protein Pax-6	1.22218E-152	2



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*

Bio Search Computing

Visualization Source Session

Session 0 Session 1

Menu

Global Tuple Data			GPDW Gene by Protein ID			
	Global Score	Tuple Score	Gene ID	Gene ID Name	Gene Concept OID	Concept OID
☐	1.00000	1.00000	8620	hgnc	769478066	1
☐	1.00000	1.00000	5080	entrez_gene	760897928	1
☐	1.00000	1.00000	ENSG000000007372	ensembl	760897928	1

Actions

More All More One

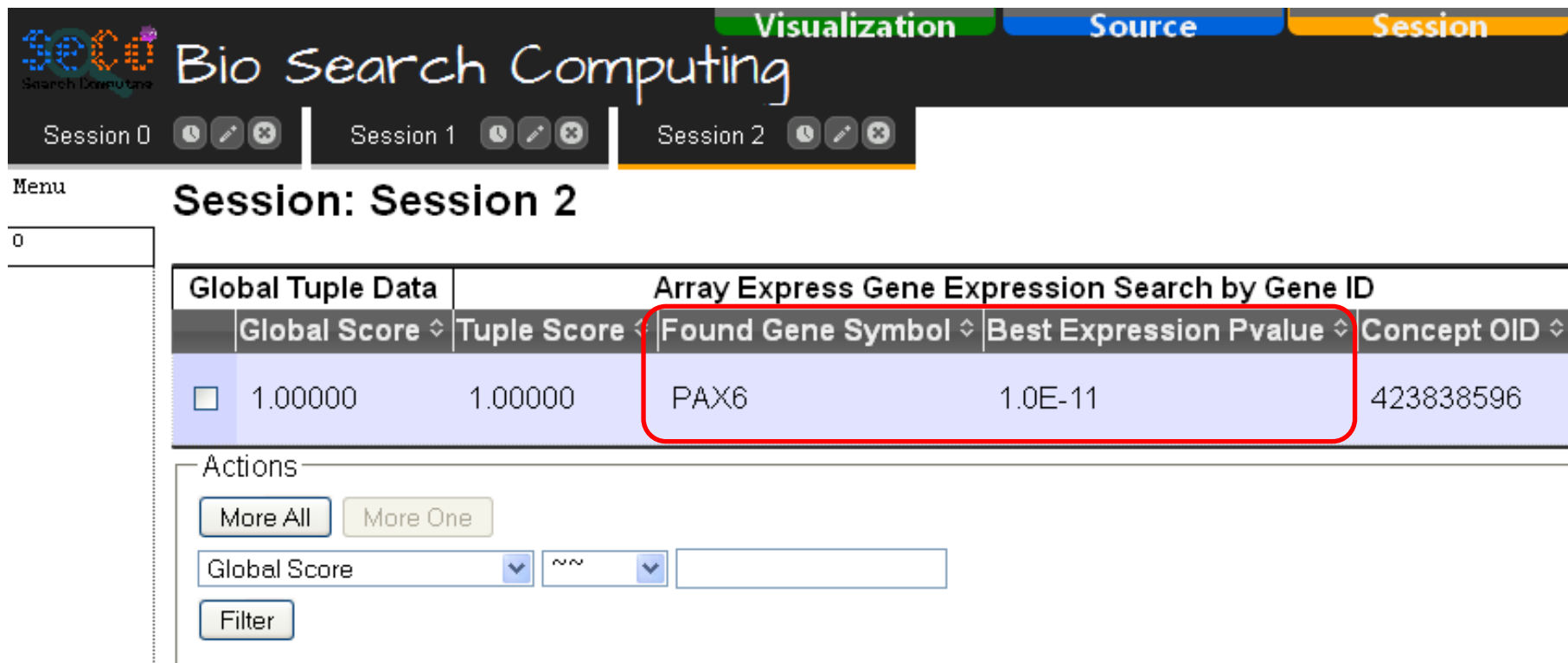
Global Score ~ ~

Filter

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*



The screenshot shows the Bio Search Computing interface. At the top, there are tabs for Visualization, Source, and Session. Below these, there are session controls for Session 0, Session 1, and Session 2. The current session is Session 2. On the left, there is a menu with a search bar containing the number 0. The main content area displays the results of a search for PAX6. The results are shown in a table with columns: Global Tuple Data, Global Score, Tuple Score, Found Gene Symbol, Best Expression Pvalue, and Concept OID. The first row shows a Global Score of 1.00000, a Tuple Score of 1.00000, the gene symbol PAX6, a Best Expression Pvalue of 1.0E-11, and a Concept OID of 423838596. Below the table, there are action buttons: More All, More One, and a Filter button. There is also a dropdown menu for Global Score and a search bar.

Global Tuple Data		Array Express Gene Expression Search by Gene ID			
	Global Score	Tuple Score	Found Gene Symbol	Best Expression Pvalue	Concept OID
☐	1.00000	1.00000	PAX6	1.0E-11	423838596

Actions:

Global Score



Results of gene2biologicalFunctionFeature search on GPDW

“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*

Global Tuple Data		GPDW Biological Function Feature by Gene ID and Feature Name				
	Global Score	Tuple Score	Biological Function Feature ID	Biological Function Feature ID Name	Biological Function Feature Name	Biological Function Feature Definition
☐	1.00000	1.00000	GO:0019220	go	regulation of phosphate metabolic process	Any process that modulates the frequency, rate or extent of the chemical reactions and pathways involving phosphates.
☐	1.00000	1.00000	GO:0031323	go	regulation of cellular metabolic process	Any process that modulates the frequency, rate or extent of the chemical reactions and pathways by which individual cells transform chemical substances.
☐	1.00000	1.00000	GO:0032268	go	regulation of cellular protein metabolic process	Any process that modulates the frequency, rate or extent of the chemical reactions and pathways involving a protein, occurring at the level of an individual cell.
☐	1.00000	1.00000	GO:0051174	go	regulation of phosphorus metabolic process	Any process that modulates the frequency, rate or extent of the chemical reactions and pathways involving phosphorus or compounds containing phosphorus.
☐	1.00000	1.00000	GO:0051246	go	regulation of protein metabolic process	Any process that modulates the frequency, rate or extent of the chemical reactions and pathways involving a protein.



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*



Combined search results



POLITECNICO
DI MILANO



Bio Search Computing

Visualization

Source

Session

Session 0

Menu

- 0
- 1
- 2
- 3
- 4
- 5
- 6

NCBI Blast Sequence Alignment Search by Protein ID

	Tuple Score	Found Sequence Symbol	Found Sequence ID	Best Alignment Expectation
☐	0.25289	PAX7_HUMAN	P23759	1.35413E-76
☐	0.23255	PAX2_HUMAN	Q02962	1.72295E-70
☐	0.22831	PAX8_MOUSE	Q00288	3.22281E-69
☐	0.22311	PAX8_HUMAN	Q06710	1.16475E-67

GPDW Gene by Protein ID

	Gene ID	Gene ID Name
☐	ENSG00000009709	ensembl
☐	ENSG00000075891	ensembl
☐	ENSMUSG00000026976	ensembl
☐	ENSG00000125618	ensembl
☐	18510	entrez_gene

Menu

Array Express Gene Expression Search by Gene ID

	Tuple Score	Found Gene Symbol	Best Expression Pvalue
☐	1.00000	PAX7	1.0E-11
☐	1.00000	PAX2	1.0E-11
☐	1.00000	Pax8	1.0E-11
☐	1.00000	PAX8	1.0E-11
☐	0.19063	emb2444	0.0080
☐	0.17146	AT5G18510	0.013
☐	0.13016	AT3G18510	0.037

Menu

GPDW Biological Function Feature by Gene Concept

	Biological Function Feature ID	Biological Function Feature Name
☐	GO:0043067	regulation of programmed cell death
☐	GO:0043067	regulation of programmed cell death
☐	GO:0043067	regulation of programmed cell death
☐	GO:0043067	regulation of programmed cell death



See **Bio-SeCo** online at

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

Tomorrow **DEMO**

Thank you for your attention!

Any question?