Bioinformatics

CNR

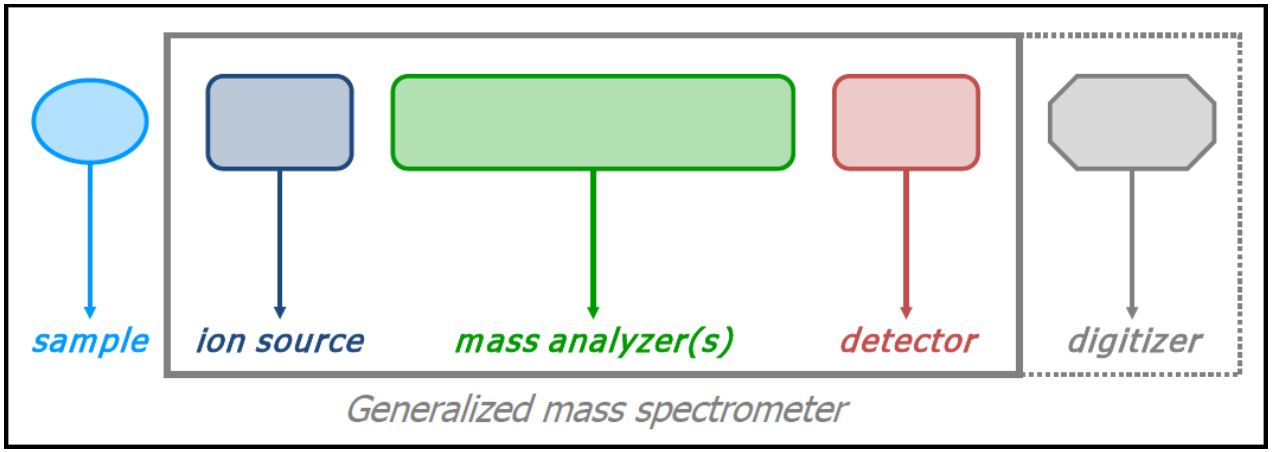
PeptidomicsDB: a new platform for sharing MS/MS data.

Federica Viti, Ivan Merelli, Dario Di Silvestre, Pietro Brunetti, Luciano Milanesi, Pierluigi Mauri

NETTAB2010

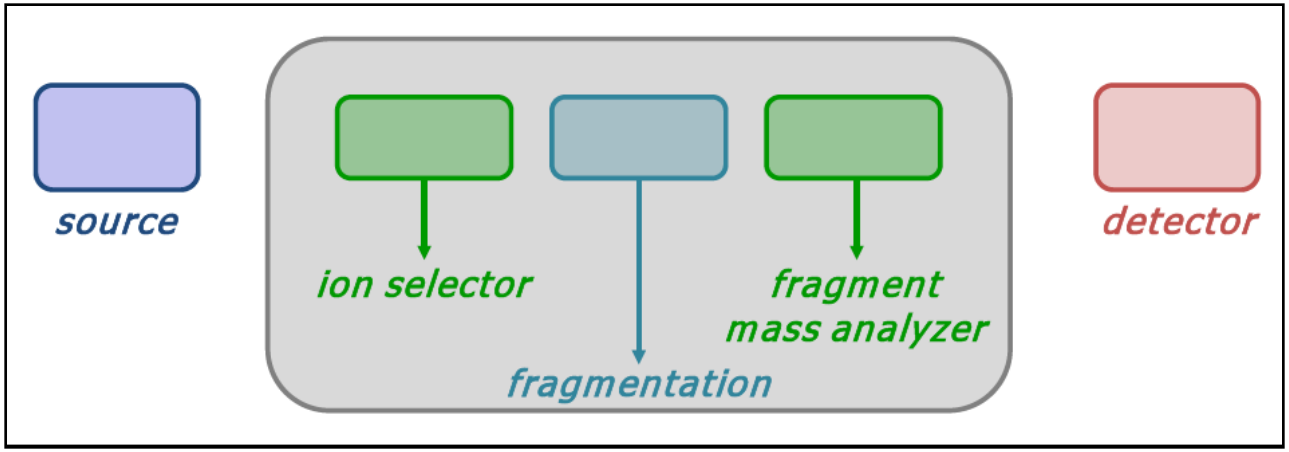
Napoli, 01/12/2010

Mass Spectrometry



The **ion source** ionizes molecules and brings them into the gas phase.
The **mass analyzer** operates on gas-phase ions using electromagnetic fields to detected mass-over-charge (m/z) ratio.
The **detector** is responsible for actually recording the presence of ions.

MS/MS data



Two MS in series

- the first MS performs the function of ion selector, by selectively allowing only ions of a given m/z to pass through;
- the second MS is situated after fragmentation and is used as a mass analyzer for the fragments.

This approach allows the sequencing of the peptide and consequently a more accurate protein recognition

Repositories for proteomics (1)

Peptide Atlas and **GPMdb**

-data reprocessing: uploaded raw data are not presented as they have been analysed by the owner but are processed again using pipelines developed expressly for the repository and based on PeptideProphet for PeptideAtlas and X!Tandem for GPMdb.

-both repositories provide protein annotations and proteotypic peptides prediction, identified as being highly related to the presence of the associated protein within the sample (unique requirement for GPMdb) and uniquely associated to a certain protein (additional requirement for PeptideAtlas).

Repositories for proteomics (2)

Proteomics Identifications Database (PRIDE) – EBI

- focused on the submission of proteins identification, while peptides spectra are optional.
- metadata are mandatory for the submission, in order to better understand experiments and data analysis and to perform queries on uploaded information (metadata schema has been developed according to the MIAPE standard).
- submitted data are maintained private until the submitter chooses to public them.
- it does not suggests how to enrich the protein list nor how to identify proteotypic peptides.

Repositories for proteomics (3)

Tranche

- organized as a filesystem
- accepts any proteomics-related files, regardless of their format
- simple repository design which do not allow advanced queries: after file uploading a unique hash key is retrieved, necessary to access the data.

Peptidome – NCBI

- organized into ‘Studies’ and ‘Samples’: the former are collections of related ‘samples’ and provide the description of the whole experiment; the second contain all data (lists of peptides and lists of proteins) related to the biological material processed through MS technology.

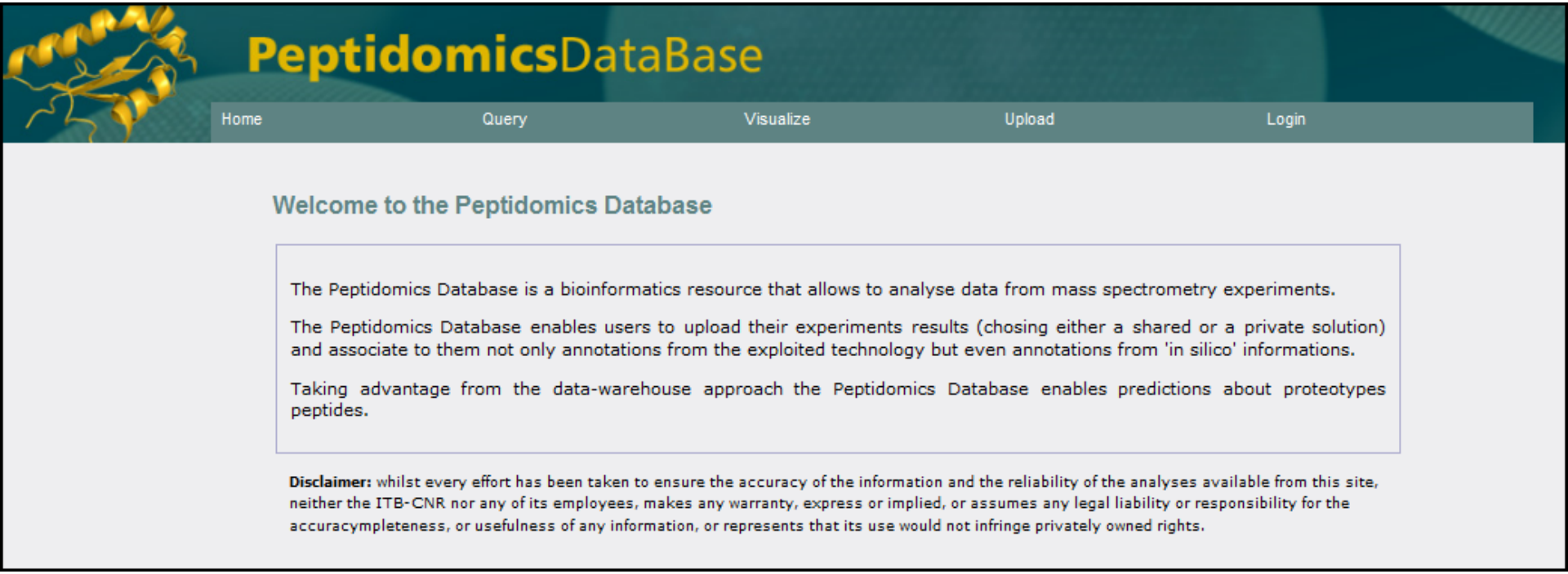
Aim of the work

Working in collaboration with the proteomics group of ITB-CNR we focused the need for a shared, analysis-oriented, MS/MS data repository.

The developed platform:

1. provides a storage solution for MS/MS data that can be used in its local version (MySQL can be customized to work in a federated mode) or in the web based one.
2. helps the identification of proteins present within a mixture, enriching the search engine output (that is often a single protein, as in Sequest).
3. supports the inference of proteotypic peptides.
4. enables collaboration and sharing within the proteomics community.

PeptidomicsDB

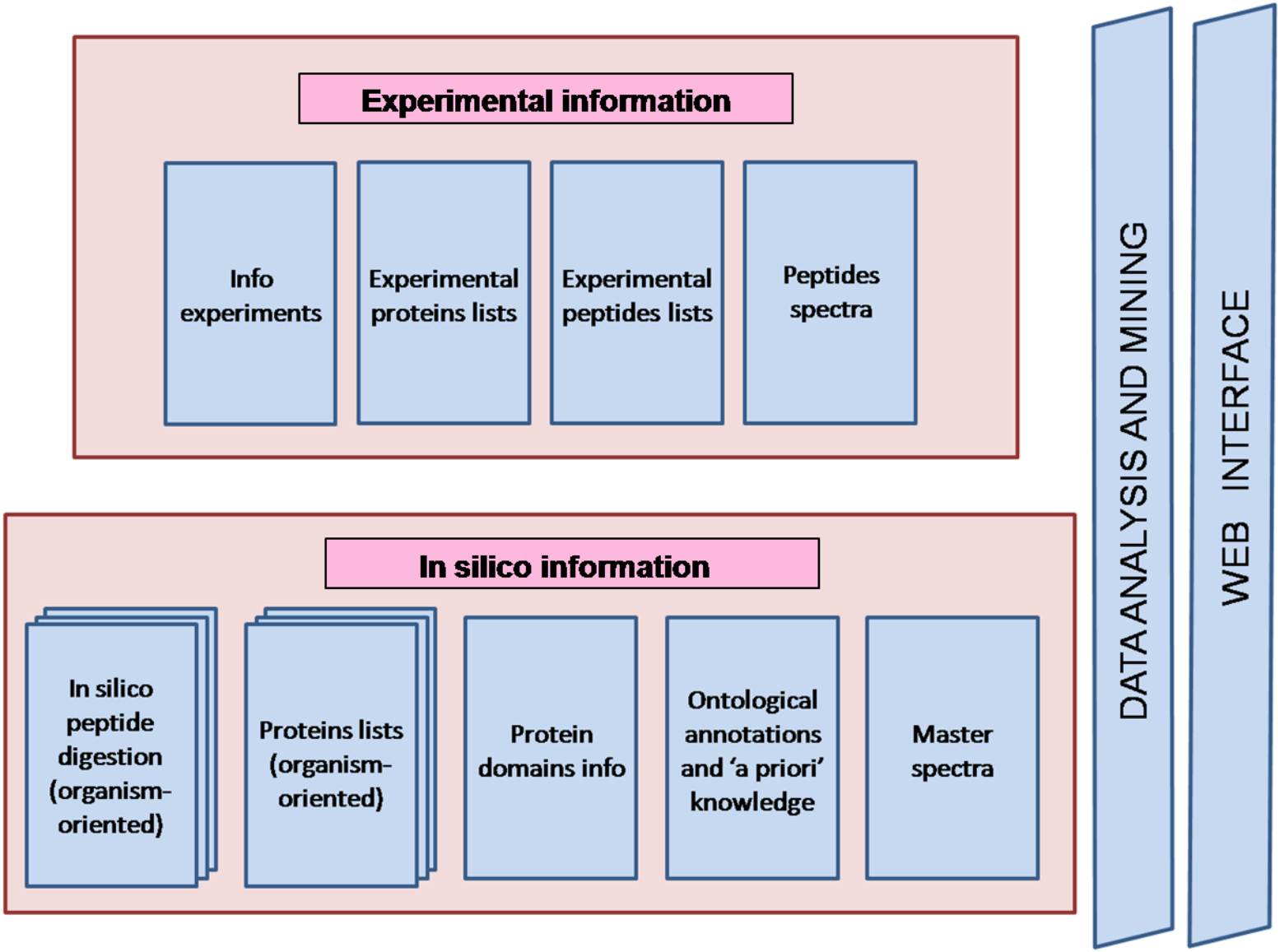


http://www.itb.cnr.it/peptidomics/

PeptidomicsDB features

- The database includes different proteomics data types, from experiments information to spectra, to peptides, to proteins.
- Spectra-peptides association is provided according to the currently available search engines (Sequest , Mascot, etc..).
- Information enrichment is performed about protein identification to overcome the one-peptide one-protein association.
- Both **in-silico** and **experimental** data are provided. In-silico data enable the re-annotation of the fragmented peptides, thus overcoming the limits of mass spectrometry software.
- In-silico information is available separately for each organism considered in the uploaded experiments.
- The database is accessible via web interface.

PeptidomicsDB design



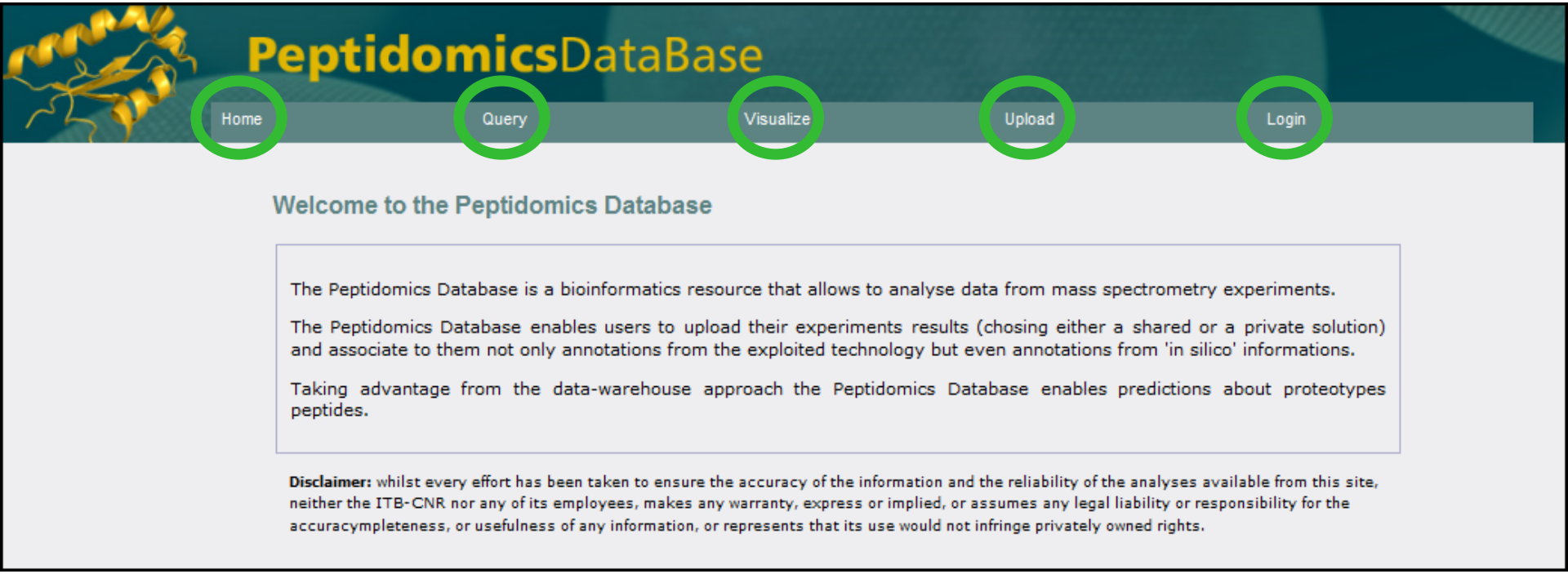
In-silico information

It enables the **re-annotation** of the fragmented peptides, thus overcoming the limits of mass spectrometry software that usually performs a ‘one peptide - one protein’ assignment.

In-silico data are collected into three kinds of tables, repeated **for each considered organism**.
Table are populated by following automated pipelines of scripts, which differ according to tables:

1. ‘In-silico protein’ table is a non-redundant list of proteins annotated with their sequence, Entrez gi identifier, reference name and description.
2. ‘Synonym’ table maintains a redundant list of the proteins that find a representative in the ‘In-silico protein’ table.
3. ‘In-silico peptide’ table is created from the ‘in-silico protein’ table, by digesting each reference protein sequence through a customized version of Proteogest perl script.

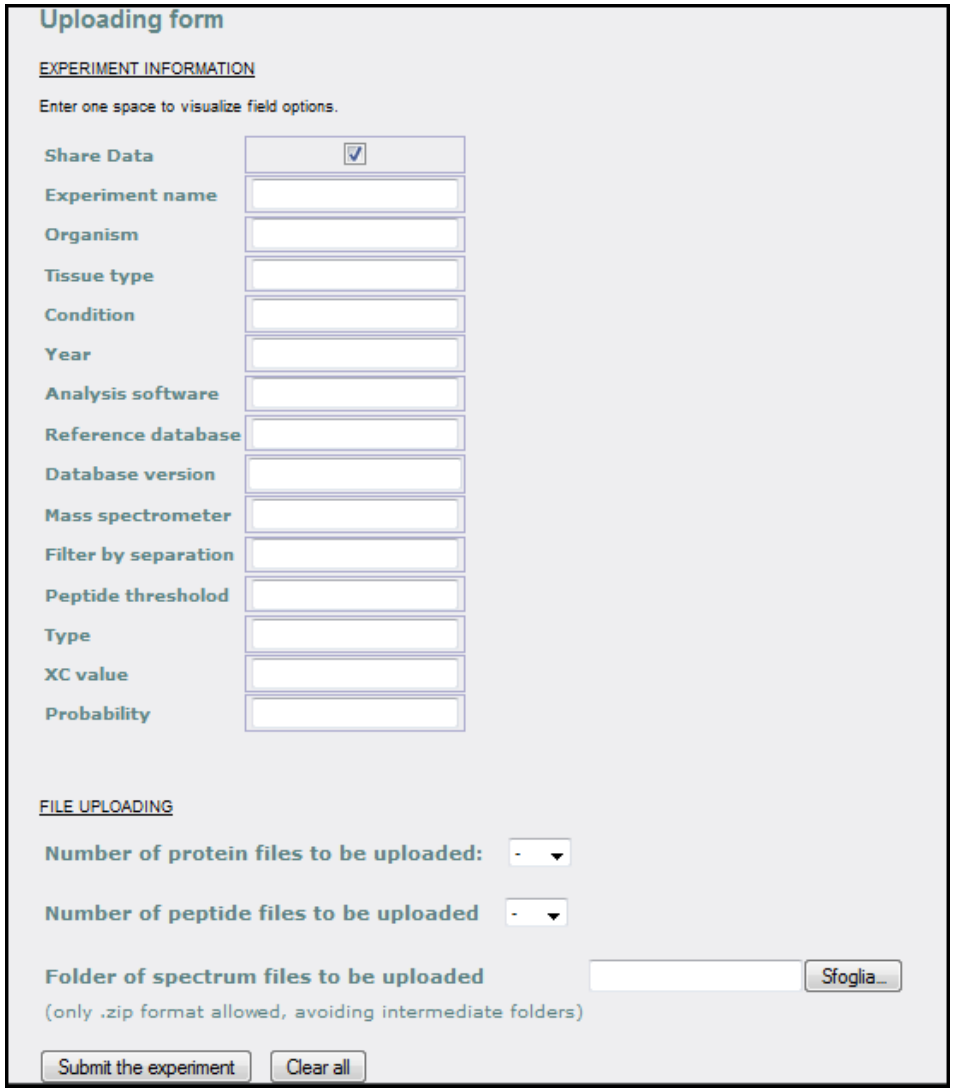
PeptidomicsDB



http://www.itb.cnr.it/peptidomics/

Upload

This section allows the submission of experiment characteristics and the upload of spectra, peptide list and protein list files. Data are recorded into database tables and associated to a-priori and in-silico knowledge, thus integrating the search engine results with other annotations and protein identification options.



Visualize

This tab allows to retrieve the list of uploaded experiments, ordered by organism, year of experiment performance or file owner.

For each experiment:

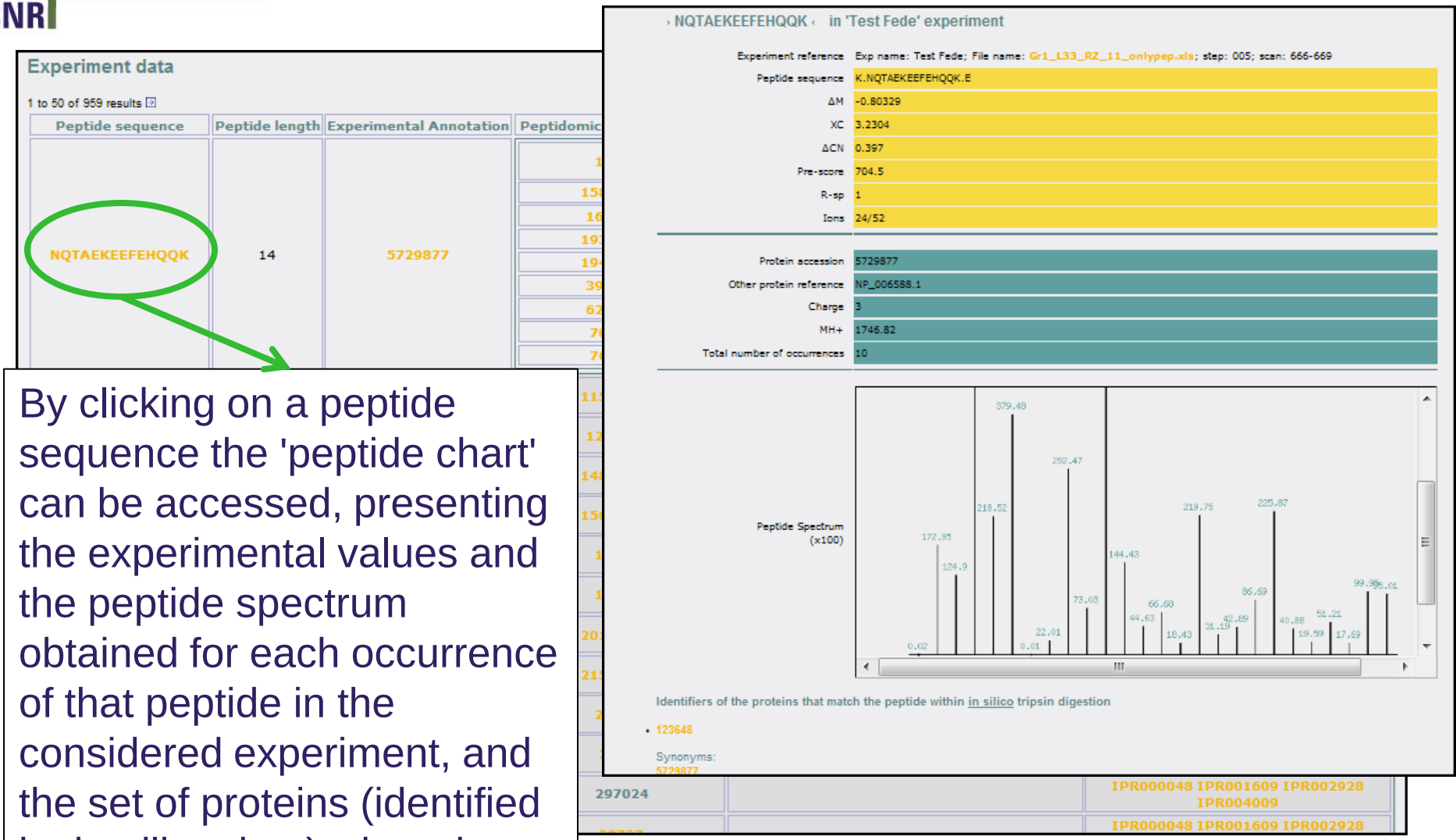
- peptides list
- identified proteins
- alternative proteins
- their synonyms
- associated protein domains.

Experiment data

1 to 50 of 959 results

Peptide sequence	Peptide length	Experimental Annotation	PeptidomicsDB annotation	Protein synonyms	Protein family / Protein domains
NQTAEKEEFEHQQK	14	5729877	123648	119587943 119587944 13273304 16740593 16741727 18043726 32467 5729877	IPR001023 IPR013126 IPR018181
			158257566		IPR001023 IPR013126 IPR018181
			16041670		IPR001023 IPR013126
			193788318		IPR001023 IPR013126 IPR018181
			194384180		IPR001023 IPR013126 IPR018181
			39645216		IPR001023 IPR013126
			62897129		IPR001023 IPR013126 IPR018181
			7020757		IPR001023 IPR013126
			7688965		IPR001023 IPR013126
KLAEKDEEMEQAK	13	297024	115496169	124302198 83304912	IPR000048 IPR001609 IPR002928 IPR004009 IPR015650
			12053672		IPR000048 IPR001609 IPR002928 IPR004009 IPR015650
			148342499		IPR000048 IPR001609 IPR002928 IPR004009 IPR015650
			156104908	124376530	IPR000048 IPR001609 IPR002928 IPR004009
			179510	179508	IPR000048 IPR001609 IPR002928 IPR004009 IPR015650
			188986	825694	IPR000048 IPR001609 IPR002928 IPR004009 IPR015650
			201067589		IPR000048 IPR001609 IPR002928 IPR004009 IPR015650
			215274256		IPR000048 IPR001609 IPR002928 IPR004009
			219524		IPR000048 IPR001609 IPR002928 IPR004009
			29468		IPR000048 IPR001609 IPR002928 IPR004009 IPR015650
			297024		IPR000048 IPR001609 IPR002928 IPR004009
			[illegible]		IPR000048 IPR001609 IPR002928

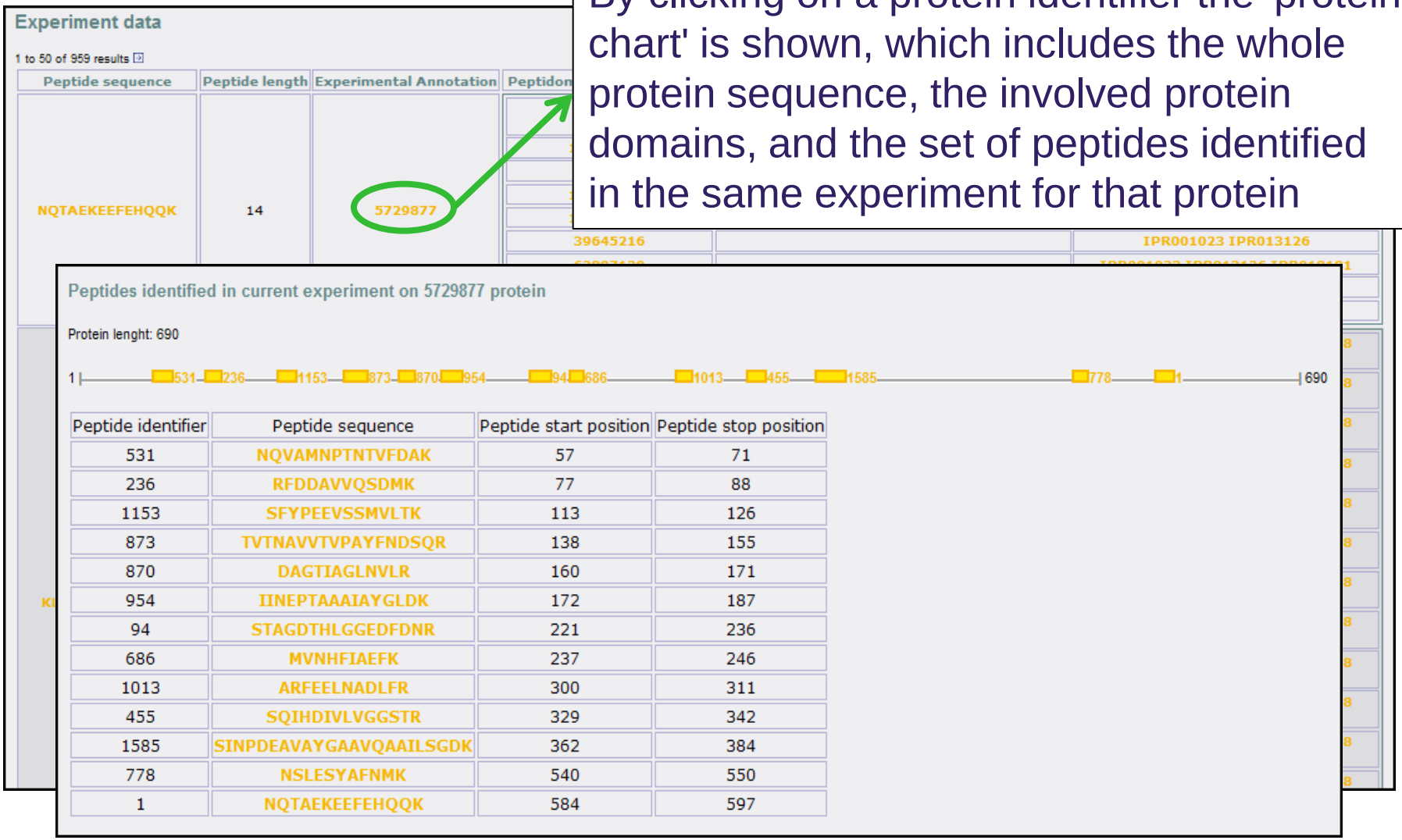
Peptide chart



By clicking on a peptide sequence the 'peptide chart' can be accessed, presenting the experimental values and the peptide spectrum obtained for each occurrence of that peptide in the considered experiment, and the set of proteins (identified by in-silico data) where it appears.

Protein chart

By clicking on a protein identifier the 'protein chart' is shown, which includes the whole protein sequence, the involved protein domains, and the set of peptides identified in the same experiment for that protein



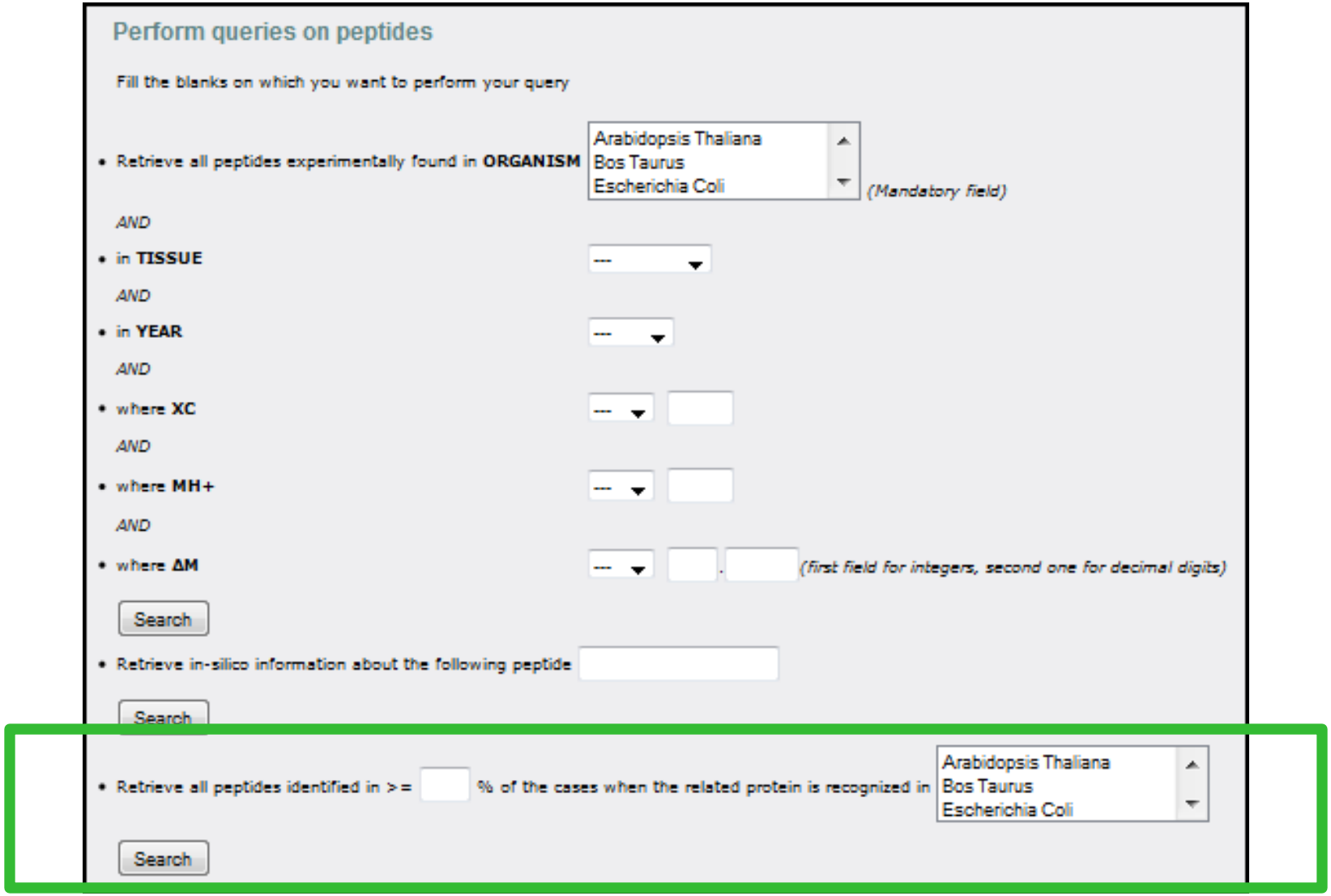
Peptide identifier	Peptide sequence	Peptide start position	Peptide stop position
531	NQVAMNPTNTVFDAK	57	71
236	RFDDAVVQSDMK	77	88
1153	SFYPEEVSSMVLTK	113	126
873	TVTNAVVTVPAYFNDSQR	138	155
870	DAGTIAGLNVLR	160	171
954	IINEPTAAAIAYGLDK	172	187
94	STAGDTHLGGEDFDNR	221	236
686	MVNHFIAEFK	237	246
1013	ARFEELNADLFR	300	311
455	SQIHDIVLVGGSTR	329	342
1585	SINPDEAVAYGAAVQAAILSGDK	362	384
778	NSLESYAFNMK	540	550
1	NQTAEKEEFEHQQK	584	597

Query on peptides

The 'Query' section provides the possibility to select a limited and focused number of experiments, proteins and peptides according to the specific interests of the user. Queries are available both on peptide and protein levels.

The **peptide** section allows to return
(i) peptides by parameters such as organism, tissue type, delta mass;
(ii) experimental features about a specific peptide;
(iii) peptides identified in a selected organism as associated to a defined protein in a certain percentage of cases.

Query on peptides



Proteotypic peptides

The definition of libraries of proteotypic peptide sequences is a crucial target, since they can be exploited to quickly scan through collections of tandem mass spectra for easily and unequivocally discovering the proteins present in the sample.

- Retrieve all peptides identified in >= 60 % of the cases when the related protein is recognized in

Escherichia Coli
Homo Sapiens
Mus Musculus

Search

List of peptides identified in >= 60 % of the cases when related protein appear in Homo Sapiens organism

Peptide sequence	Reference protein id	Percentage
AAAEVNQDYGLDPK	182794	60
AAALAHLDR	3510334	100
AAAVLPVLDLAQR	222080062	66.666666666667
AAFDDAIAELDTLSEESYK	5803225	60
AAFDDAIAELDTLSEESYKDSTLIMQLLR	5803225	60
AANDAGYFNDEMAPIEVK	167614485	100
AAQSQLSQGDLVVAIDGVNTDTMTHLEAQN	3327040	100
AASADSTTEGTPADGFTVLSTK	4758496	100
AAVEQLTEEQKNEFK	4507615	100
AAVPSGASTGIYEALELR	114665857	100
AAYFGVYDTAK	55749577	66.666666666667
ADHHATNGVVHLIDK	4507467	100
ADLINNLGTIAK	32486	100
ADVDAATLAR	55749932	60
AENKLHVSLMENYPGTLEALGEPIR	119590272	100
AENKLHVSLMENYPGTLQALGEPIR	71891703	100
AENNPWVTPIADQFQLGVSHVFEYIR	6912638	100
AEVQNLGGELVVSGVDSAMSLIQAAK	55770844	75
AFGPGLEGGLVNK	5419655	60
AFMTADLPNELIELLEK	4758012	100
AFQPWEDIQENFLYYEEK	5802978	60
AFYPEEISSMVLTK	4529894	75

Query on proteins

For what concerns **proteins**, selections can be performed
(i) by filtering collected proteins on experiment features such as organism, tissue type, probability, isoelectric point, molecular weight, even contemporary;
(ii) by obtaining peptides associated to a defined protein;
(iii) by listing all experiments where a target protein has been identified.

Query on proteins

Perform queries on proteins

Fill the blanks on which you want to perform your query.

- Retrieve all proteins experimentally recognised in **ORGANISM** — Escherichia Coli / Homo Sapiens / Mus Musculus *(Mandatory field)*

AND

- in **TISSUE** — ---

AND

- in **YEAR** — ---

AND

- presenting **probability** — --- --- --- *(Insert single value or range values)*

AND

- presenting **pI** — --- --- --- *(Insert single value or range values)*

AND

- presenting **MW** — --- --- --- *(Insert single value or range values)*

AND

- presenting **score** — ---

AND

- presenting **hits** — ---

Search

- Search for peptides associated to the following protein *(Search protein by its Entrez GI identifier)*

Search

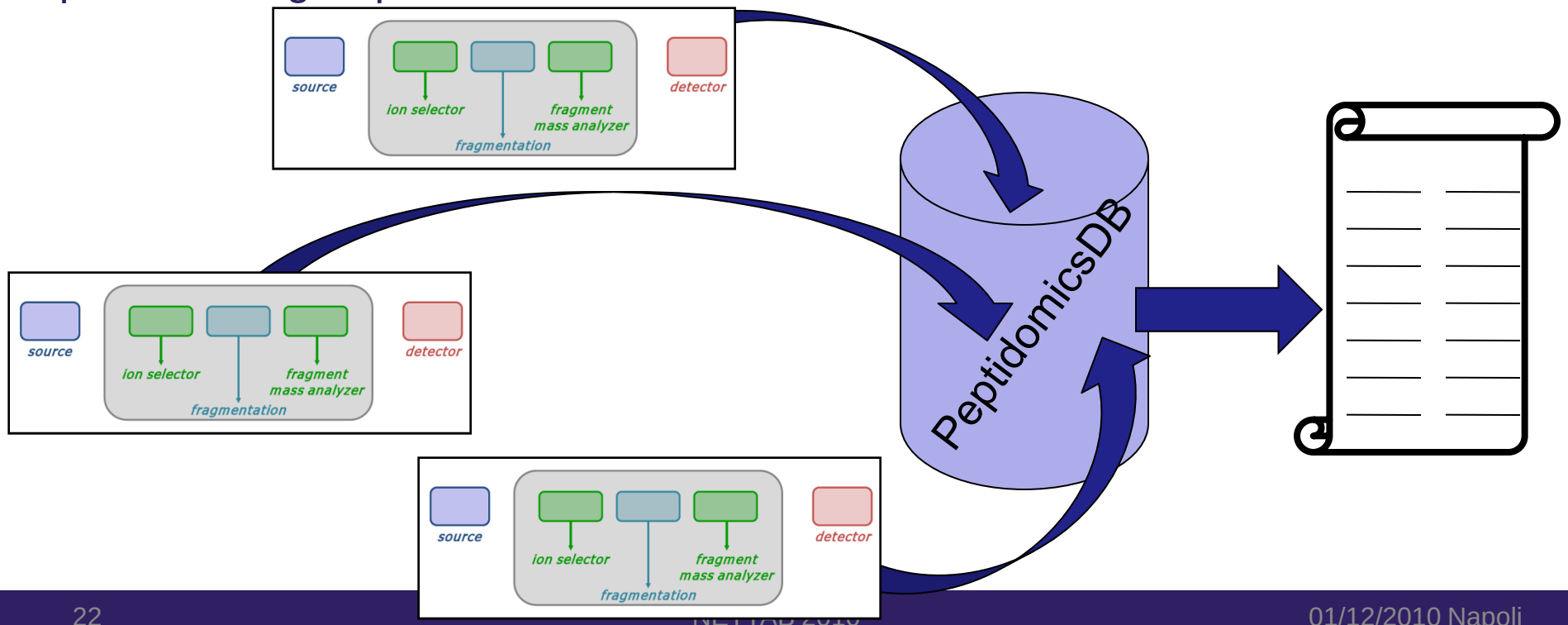
- Search for experiments where the following protein is detected *(Search protein by its Entrez GI identifier)*

Search

Work in progress

- We are paying particular attention to data enrichment through the integration of an ontological layer and a knowledge base about biomolecular processes in order to better qualify protein presence.

- We are available to collaborate with proteomics groups that would like to test our system and to share their experimental data with other proteomics groups.



Acknowledgements

Bioinformatics Division

Dr. Ivan Merelli

Dr. Luciano Milanesi

Proteomics Division

Dr. Dario Di Silvestre

Dr. Pietro Brunetti

Dr. Pierluigi Mauri

This work has been supported by the EGEE-III, BBMRI, EDGE European projects, by the MIUR FIRB ITALBIONET (RBPR05ZK2Z), BIOPOPGEN (RBIN064YAT), CNR-BIOINFORMATICS initiatives, and by the ACCORDO QUADRO TRA REGIONE LOMBARDIA - CNR.

THANKS FOR YOUR ATTENTION!

QUESTIONS?

federica.viti@itb.cnr.it