Resources and Tools for Protein-Protein Interaction

Master-Module Biological Networks
July 20, 2016

Emidio Capriotti
http://biofold.org/

Biomolecules
Folding and
Disease

Institute for Mathematical Modeling
of Biological Systems
Department of Biology

HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

Experimental Techniques

Wide variety of techniques and methods have been developed to generate PPI data and can be subdivided in:

- High throughput techniques
- Low throughput techniques

These techniques can be further divided in:

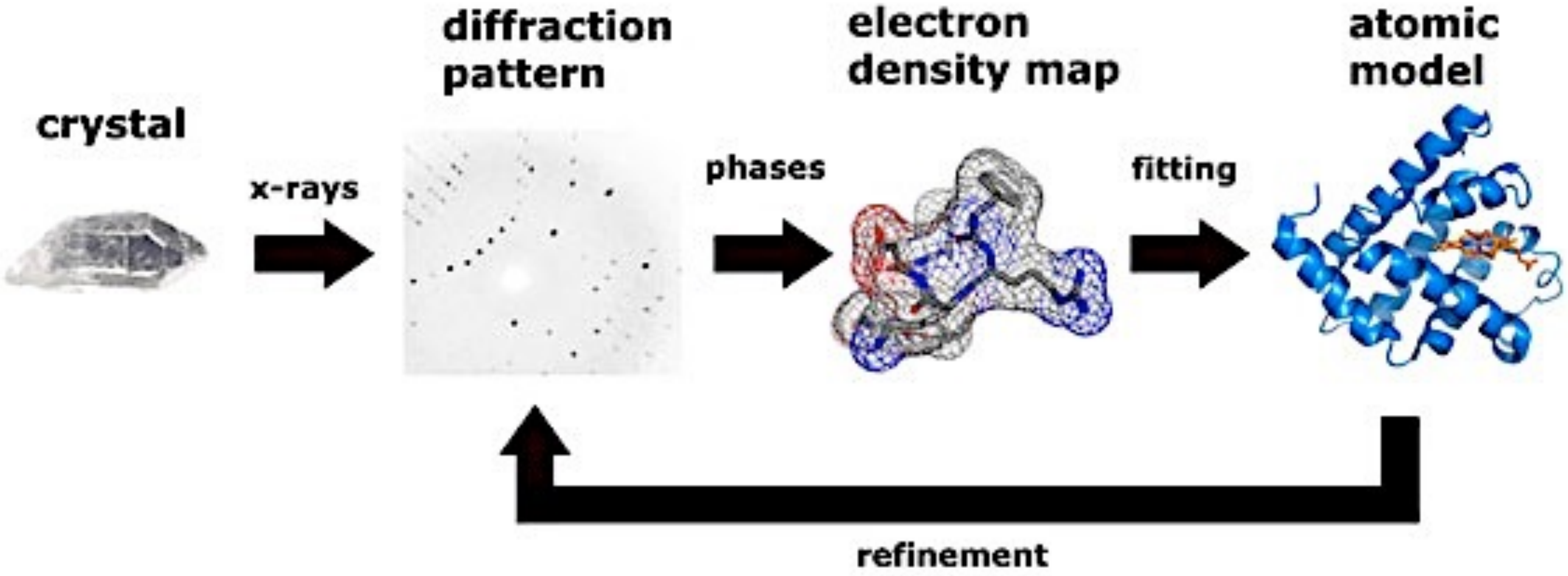
- techniques that detect direct physical interactions between two proteins, called binary methods
- techniques that detect interactions among groups of proteins that may not form physical contacts — co-complex methods.

Luana Licata, Systems Biology Course 2015

Low Throughput Techniques

Some low throughput techniques provide deeper insight certain characteristic of an interaction, such as FRET, NMR and X-ray crystallography.

X-ray crystallography is considered the gold standard for PPI, since provide high quality data of binding surfaces to the level of individual atoms and binding sites.

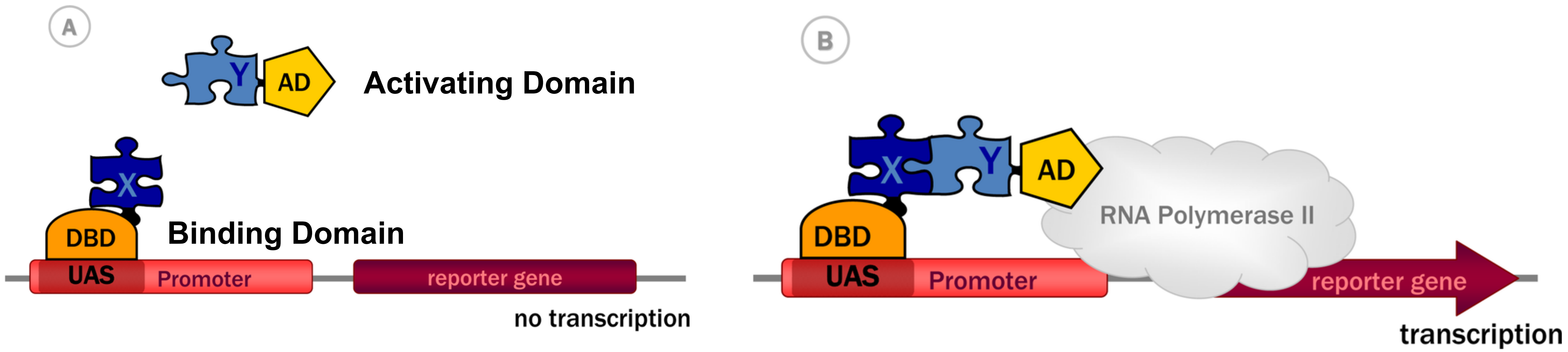


From Thomas Splettstoesser (www.scistyle.com)

High Throughput Techniques

The main binary methods for measuring of direct physical interactions between protein pairs is Yeast two-hybrid (Y2H).

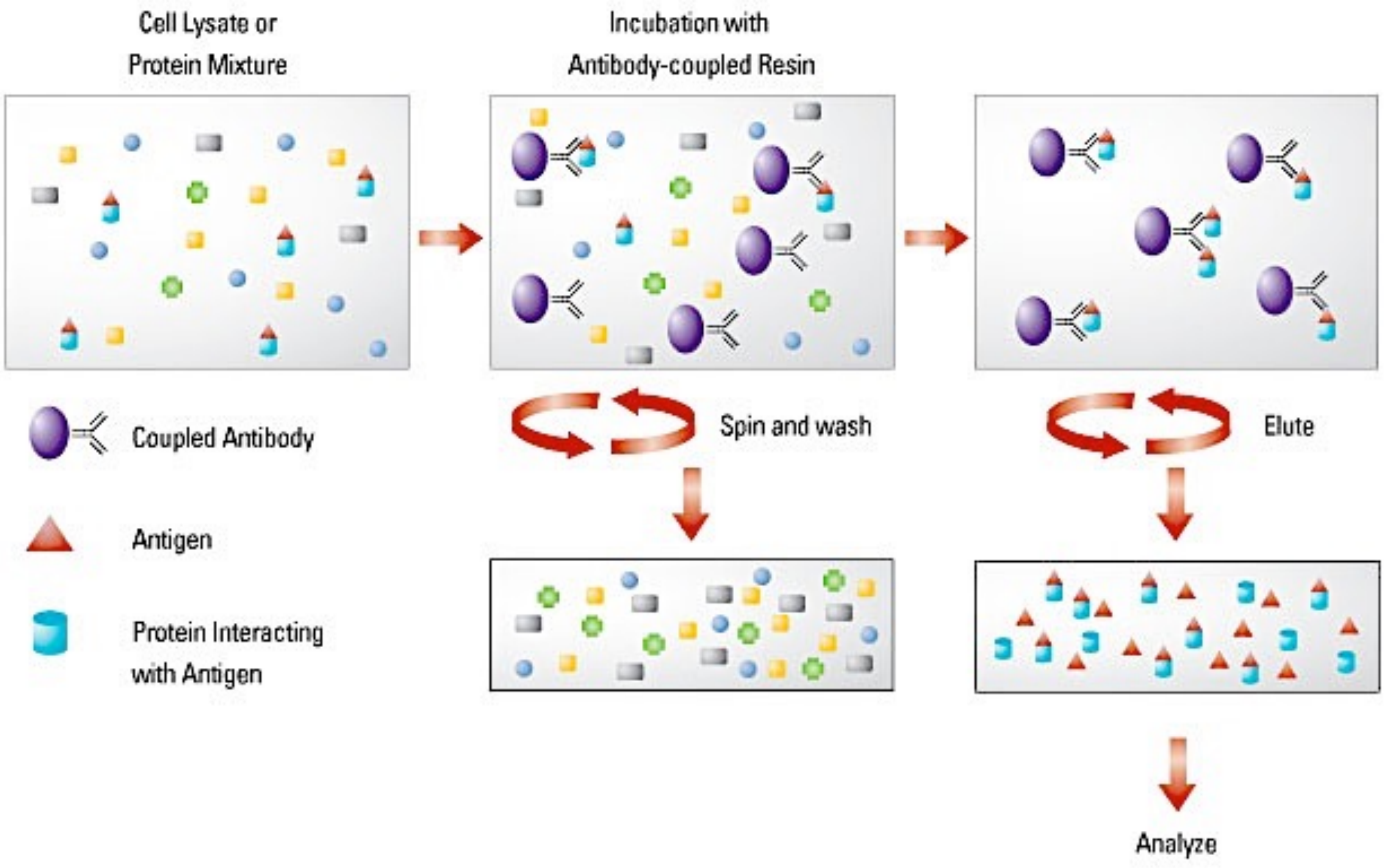
The strategy interrogates two proteins, called bait (X) and prey (Y), coupled to two halves of a transcription factor and expressed in yeast. If the proteins make contact, they reconstitute a transcription factor that activates a reporter gene.



Anna Brückner, et al., Int. J. Mol. Sci. 2009

Co-complex Method

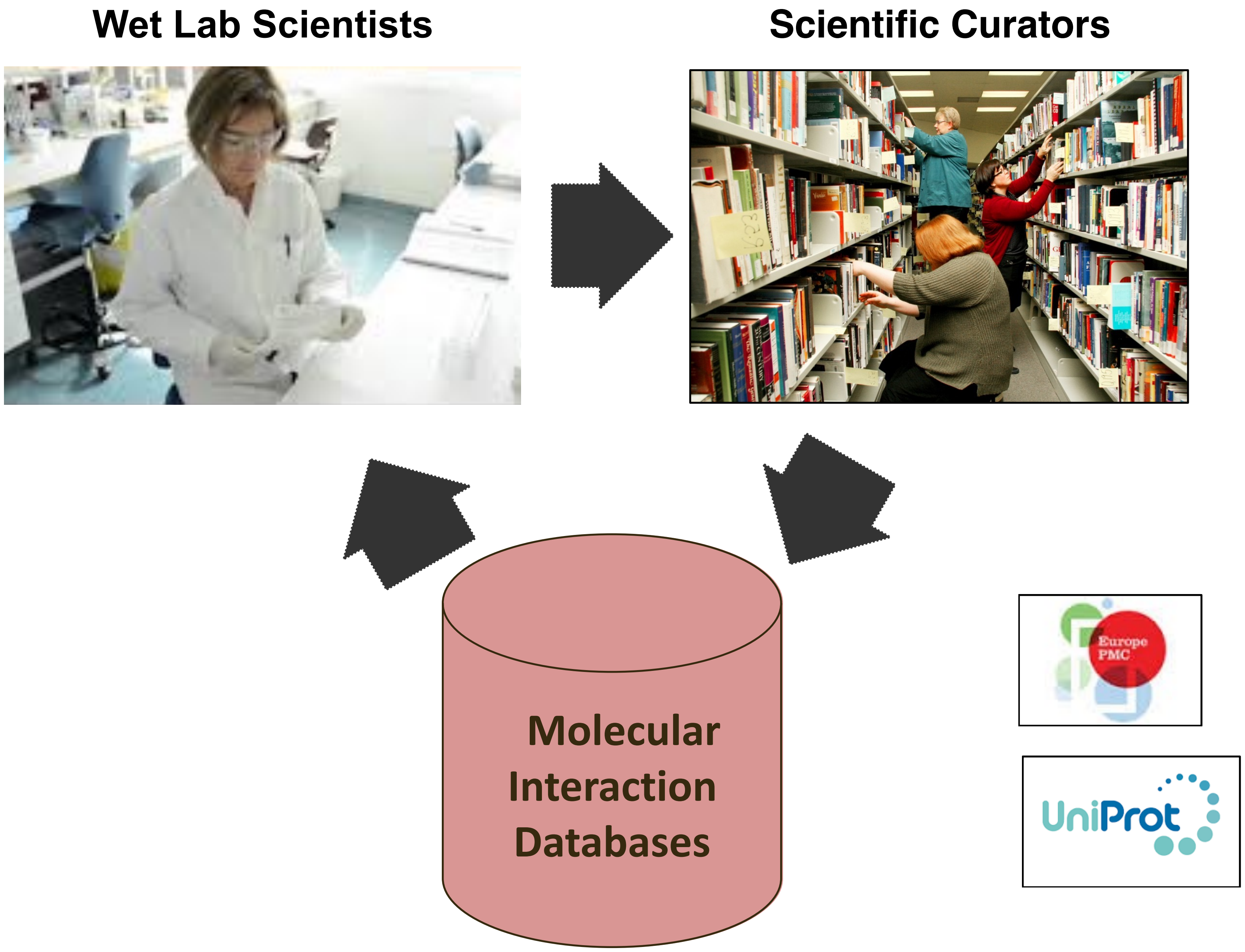
The most common co-complex method is co-immunoprecipitation (co-IP) coupled with mass spectrometry (MS). In this approach, the bait protein, usually expressed in the cell at *in vivo* conditions, is affinity purified and the interacting partners are detected by mass spectrometry.



Luana Licata, Systems Biology Course 2015

Curation and Databases

The results of experiments are published on scientific journal. The curators extract information from the literature and to develop curated databases.



Wet Lab Scientists

Scientific Curators

Molecular Interaction Databases

Europe PMC

PubMed

UniProt

National Center for Biotechnology Information NCBI

Interaction Databases

Molecular interaction databases have been established to archive and subsequently disseminate molecular interaction data in a structured format available to perform searches and bioinformatics analyses.

Molecular interaction databases can be divided in:

- Primary databases: experimentally proven protein interactions coming from either small-scale or large-scale published studies that have been manually curated

- Meta databases: experimentally proven PPIs obtained by consistent integration of several primary databases

- Prediction databases: mainly predicted PPIs derived using different approaches, combined with experimentally proven PPIs.

Luana Licata, Systems Biology Course 2015

Database Classification

Type of data captured:

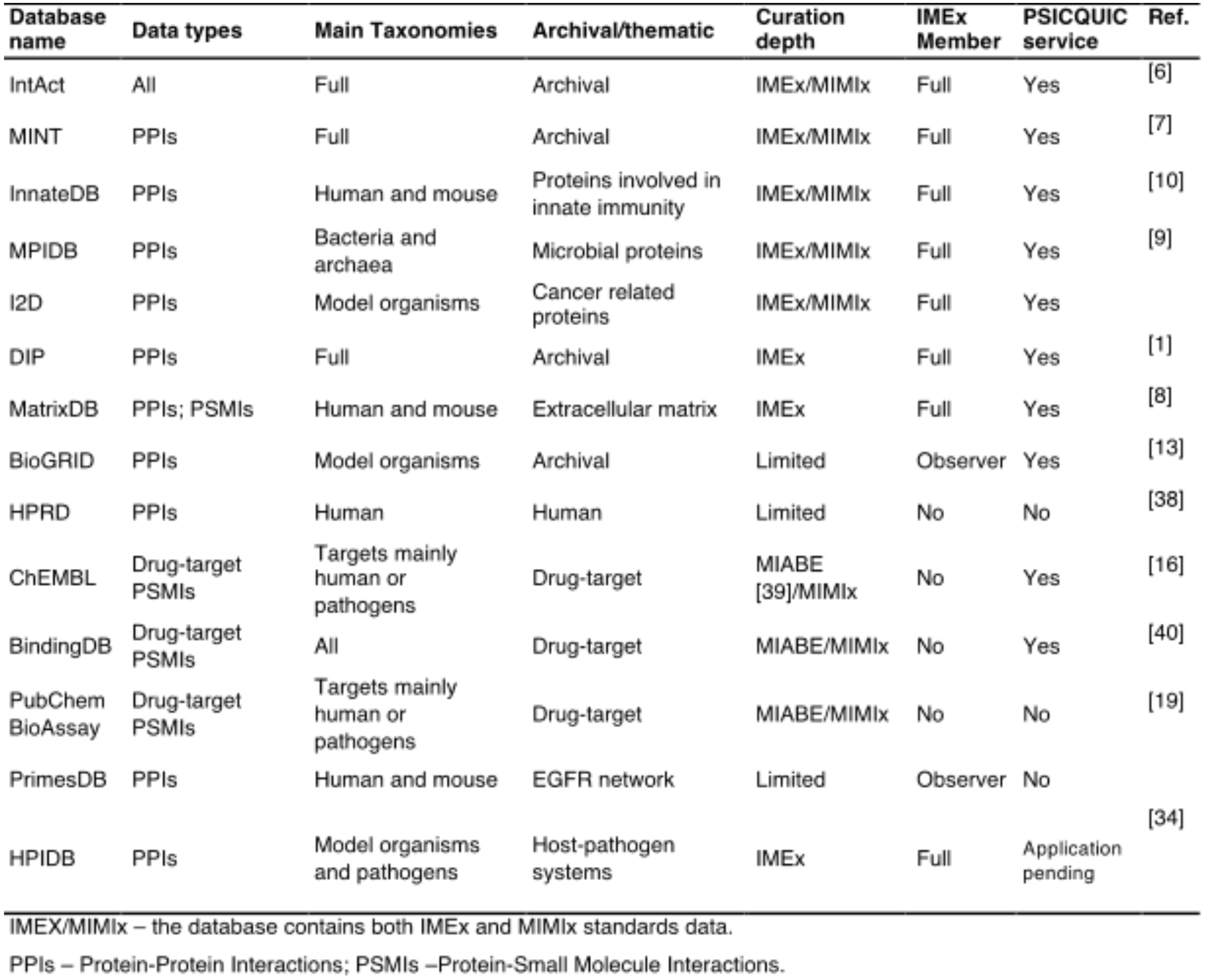
- Only PPIs information as MINT and DIP.
- Interactions between proteins and other molecular types (DNA, RNA, small molecules) as IntAct and MatrixDB.
- PPIs and genetic interactions as BIOGRID.
- Only PPIs related to a specific scientific topic such as : InnateDB (PPIs in the immune system), MPIDB (PPIs in microbes) and MatrixDB (extracellular PPIs).

Type of curation Policy:

- Databases describing PPIs with low level of curation details and quality control procedures
- Databases describing PPIs with high level of curation details and high accuracy of quality control procedures such as IMEx databases.

Important Databases

A complete list of molecular interaction databases is available at: http://www.pathguide.org.



Database name	Data types	Main Taxonomies	Archival/thematic	Curation depth	IMEx Member	PSICQUIC service	Ref.
IntAct	All	Full	Archival	IMEx/MIMIx	Full	Yes	[6]
MINT	PPIs	Full	Archival	IMEx/MIMIx	Full	Yes	[7]
InnateDB	PPIs	Human and mouse	Proteins involved in innate immunity	IMEx/MIMIx	Full	Yes	[10]
MPIDB	PPIs	Bacteria and archaea	Microbial proteins	IMEx/MIMIx	Full	Yes	[9]
I2D	PPIs	Model organisms	Cancer related proteins	IMEx/MIMIx	Full	Yes	
DIP	PPIs	Full	Archival	IMEx	Full	Yes	[1]
MatrixDB	PPIs; PSMIs	Human and mouse	Extracellular matrix	IMEx	Full	Yes	[8]
BioGRID	PPIs	Model organisms	Archival	Limited	Observer	Yes	[13]
HPRD	PPIs	Human	Human	Limited	No	No	[38]
ChEMBL	Drug-target PSMIs	Targets mainly human or pathogens	Drug-target	MIABE [39]/MIMIx	No	Yes	[16]
BindingDB	Drug-target PSMIs	All	Drug-target	MIABE/MIMIx	No	Yes	[40]
PubChem BioAssay	Drug-target PSMIs	Targets mainly human or pathogens	Drug-target	MIABE/MIMIx	No	No	[19]
PrimesDB	PPIs	Human and mouse	EGFR network	Limited	Observer	No	
HPIDB	PPIs	Model organisms and pathogens	Host-pathogen systems	IMEx	Full	Application pending	[34]

IMEX/MIMIx – the database contains both IMEx and MIMIx standards data.

PPIs – Protein-Protein Interactions; PSMIs –Protein-Small Molecule Interactions.

Luana Licata, Systems Biology Course 2015

IMEx Consortium

- An international collaboration between a group of major public interaction data providers who have agreed to share curation effort (www.imexconsortium.org)

- 12 active molecular interaction databases dedicated to producing high quality, annotated data, curated to the same standards and following the same curation rules

- Data is curated once at a single centre then exchanged between partners

- Users can query a single website to obtain all data

Imex Central

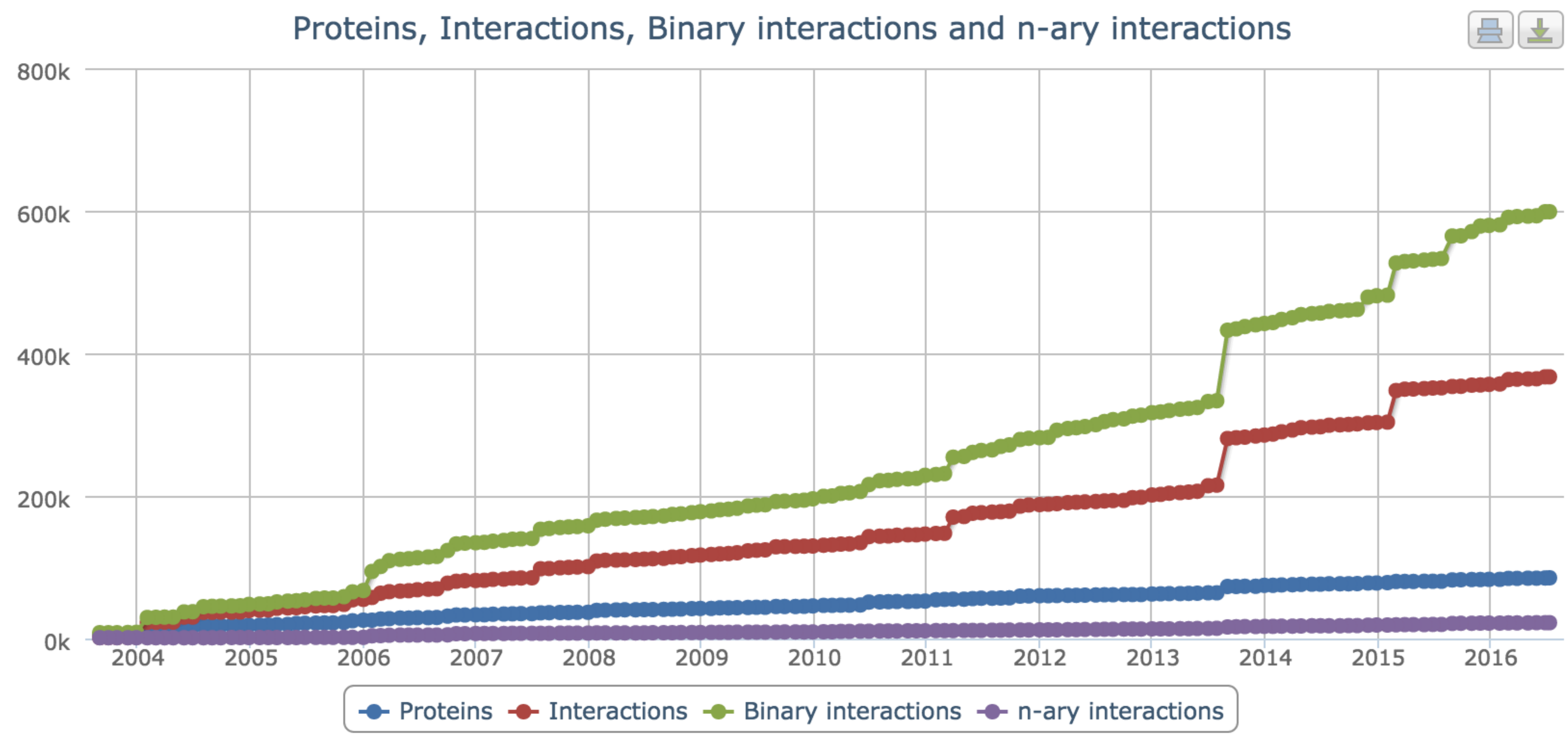
The web service IMEx Central (https://imexcentral.org/icentralbeta/) is a central resource to assign IMEx IDs to the publications curated by IMEx members (version BETA-0.93 has been recently released).

Curators can check by using the NCBI PubMed identifier (PMID) if other IMEx members have curated an already published paper and therefore it allows avoiding work duplication.

MIntAct Project

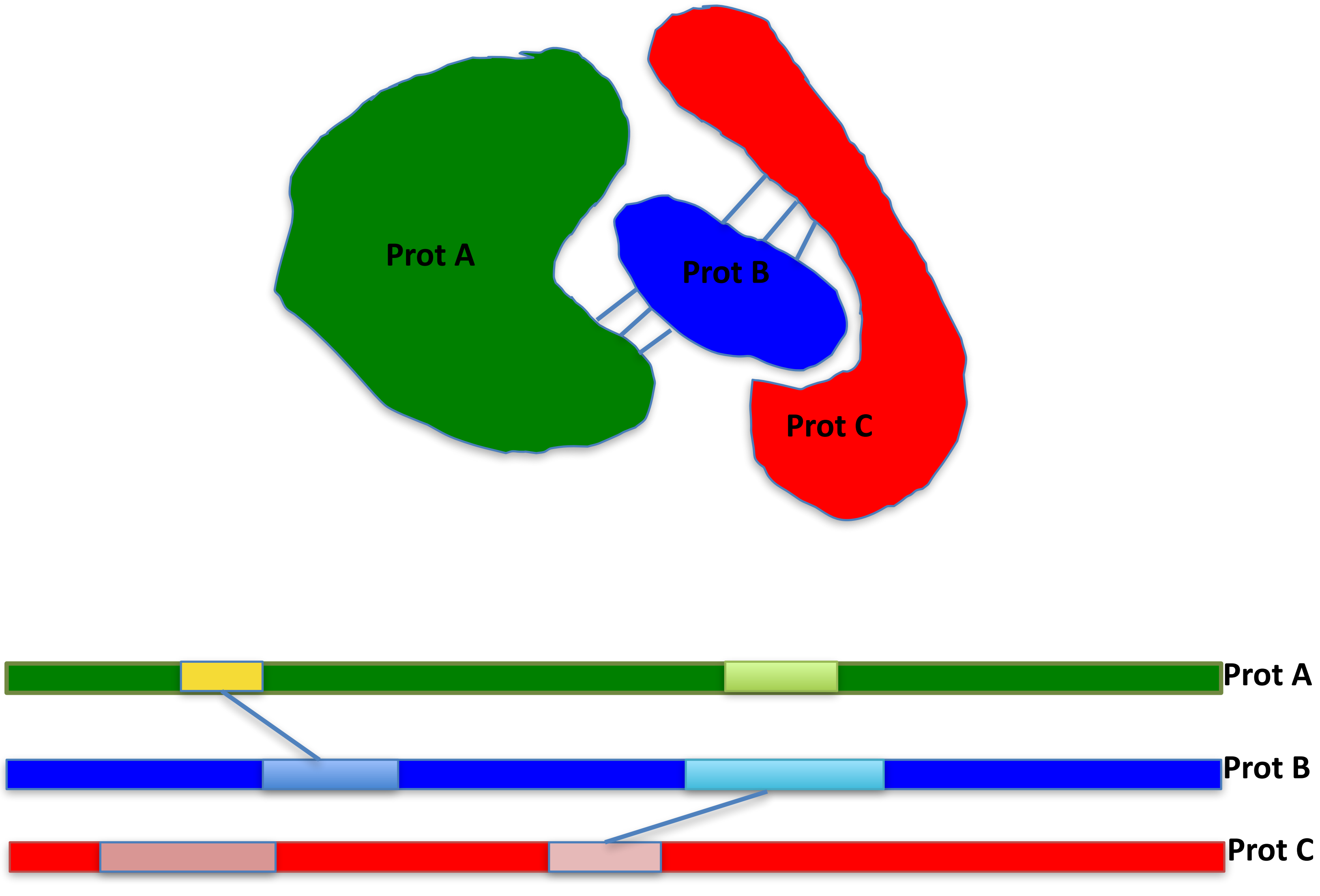
- MINT and IntAct databases were two of the largest databases (number of manuscripts curated and the number of non-redundant interactions).
- Both adopted the highest possible data quality standards.
- Both were founder members of the IMEx Consortium.

IntAct and MINT joined forces to create a single resource to improve curation and software development efforts.



PPI Representation

Representation of binding domain of interacting proteins in IMEx databases

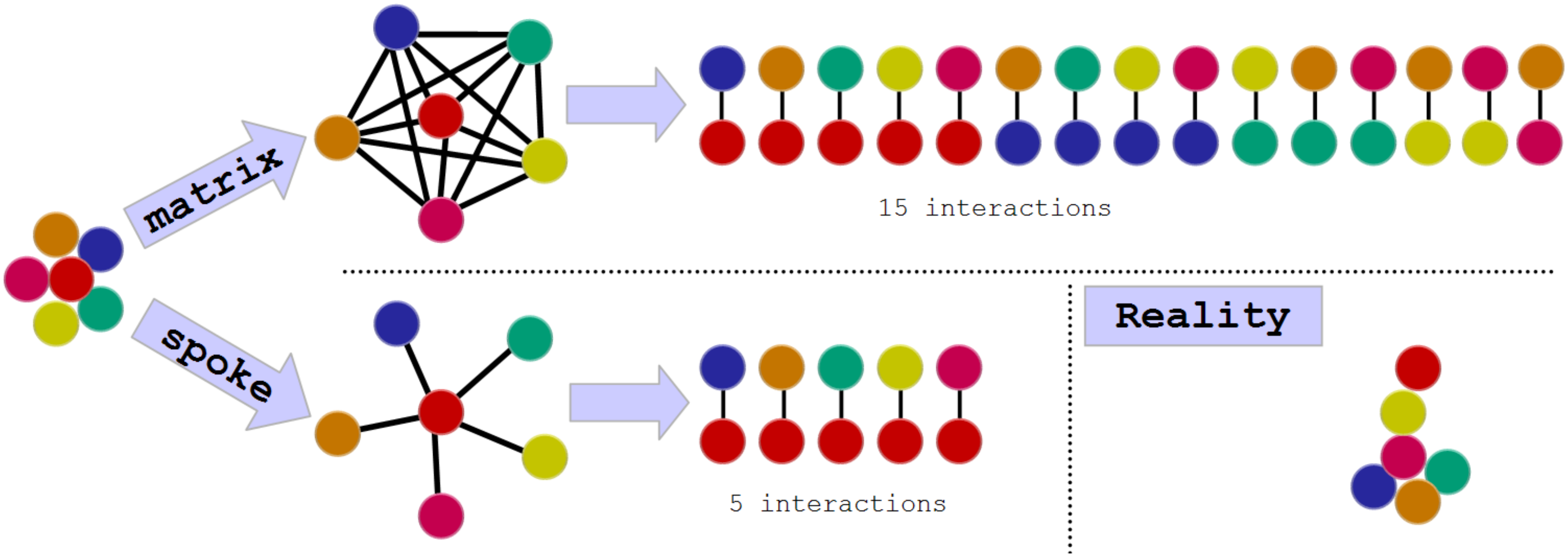


Luana Licata, Systems Biology Course 2015

Complex Representation

- Several experimental techniques produce complex data: Eg. co-IP coupled with MS

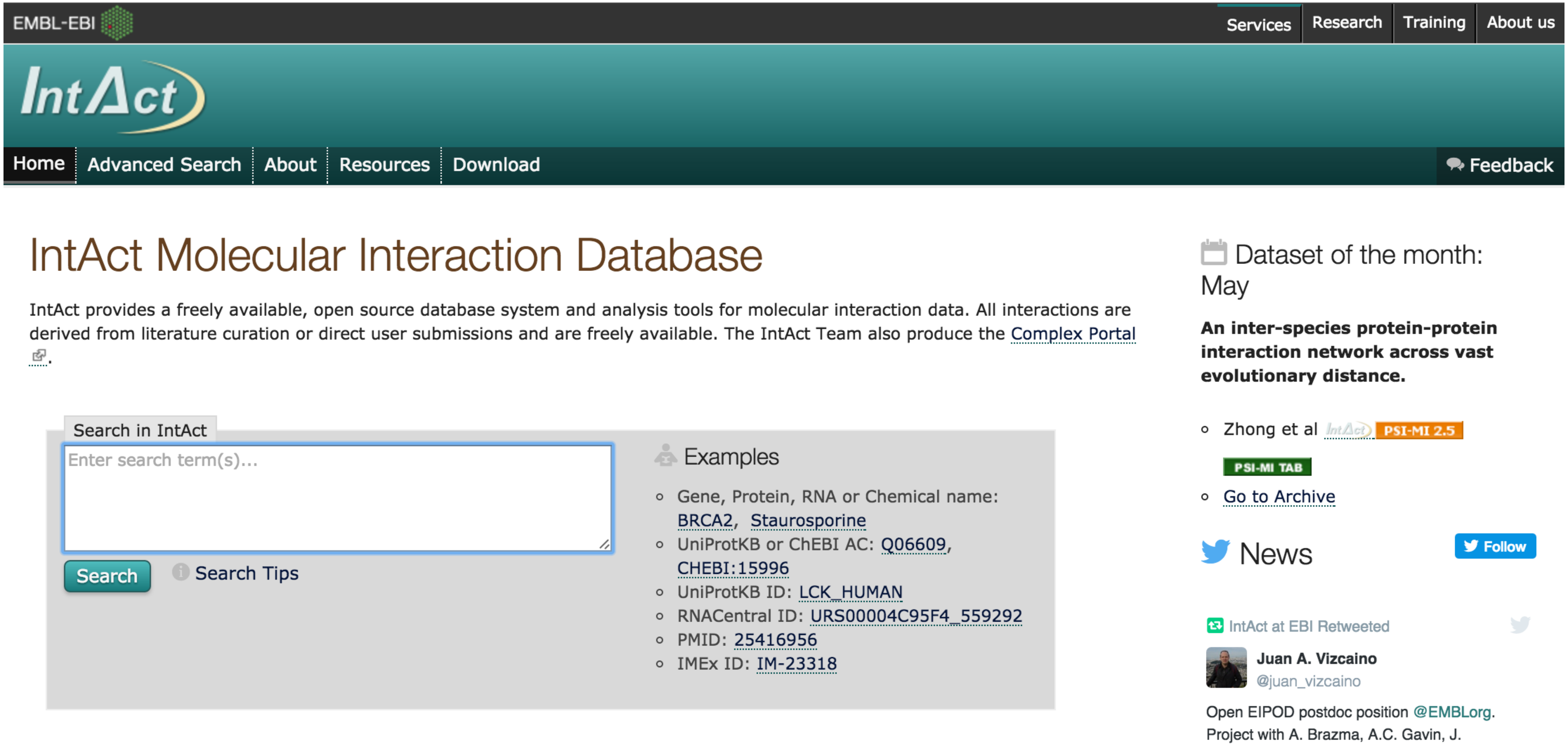
- There are two algorithms available to convert complexes into binary interactions



http://www.ebi.ac.uk

IntAct Interface

Use the input window to search for the interactions of the CREB1 protein



IntAct Molecular Interaction Database

IntAct provides a freely available, open source database system and analysis tools for molecular interaction data. All interactions are derived from literature curation or direct user submissions and are freely available. The IntAct Team also produce the Complex Portal.

Search in IntAct

Enter search term(s)...

Examples

- Gene, Protein, RNA or Chemical name: BRCA2, Staurosporine
- UniProtKB or ChEBI AC: Q06609, CHEBI:15996
- UniProtKB ID: LCK_HUMAN
- RNACentral ID: URS00004C95F4_559292
- PMID: 25416956
- IMEx ID: IM-23318

Dataset of the month: May

An inter-species protein-protein interaction network across vast evolutionary distance.

- Zhong et al IntAct PSI-MI 2.5 PSI-MI TAB
- Go to Archive

News

IntAct at EBI Retweeted

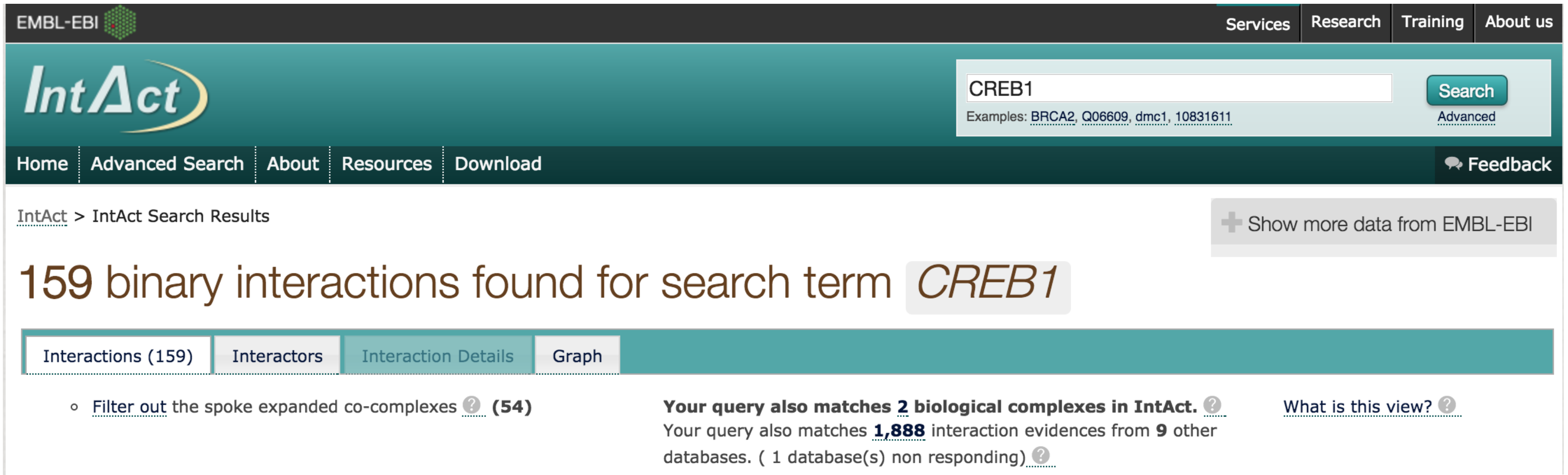
Juan A. Vizcaino
@juan_vizcaino

Open EIPOD postdoc position @EMBLorg. Project with A. Brazma, A.C. Gavin, J.

http://www.ebi.ac.uk

IntAct Output

CREB1 has 159 possible interaction, 91 of which are with proteins



Examples: BRCA2, Q06609, dmc1, 10831611

IntAct > IntAct Search Results

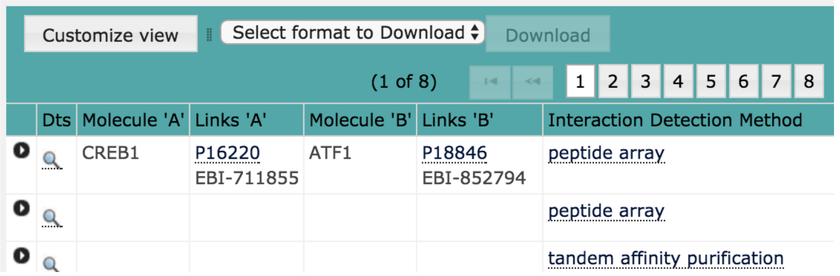
159 binary interactions found for search term *CREB1*

- Filter out the spoke expanded co-complexes (54)

Your query also matches 2 biological complexes in IntAct.

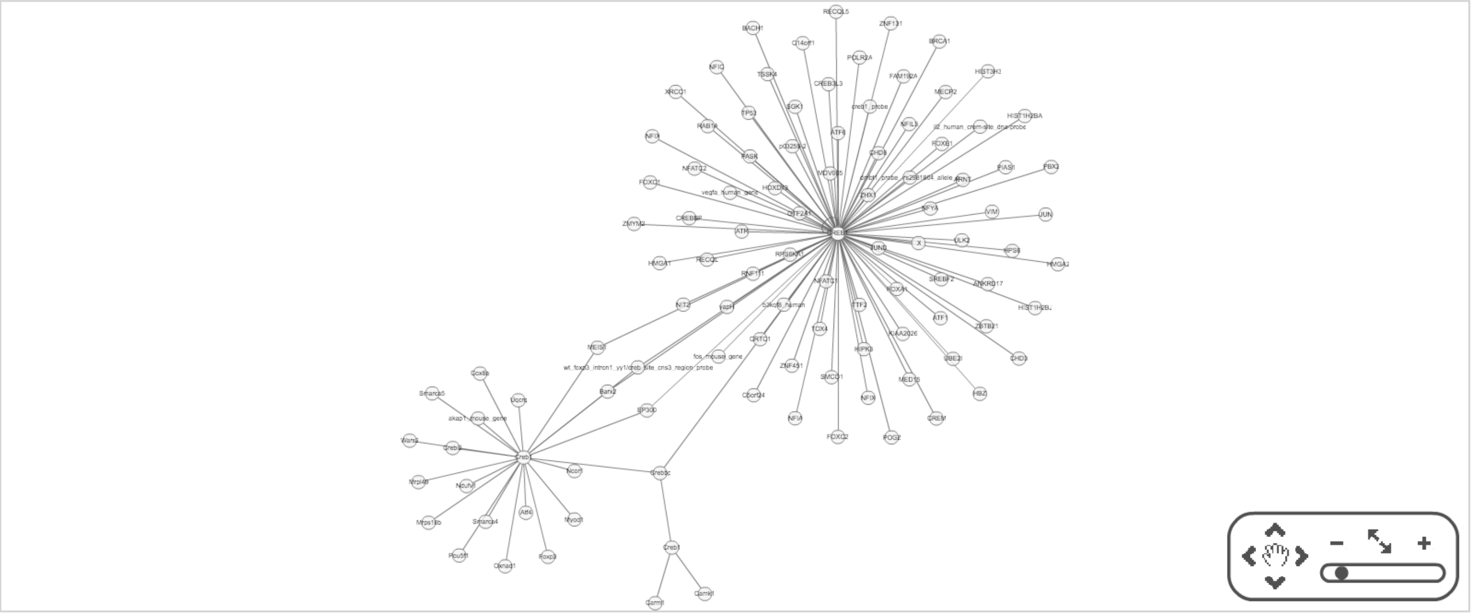
Your query also matches **1,888** interaction evidences from **9** other databases. (1 database(s) non responding)

What is this view?



	Dts	Molecule 'A'	Links 'A'	Molecule 'B'	Links 'B'	Interaction Detection Method
●	🔍	CREB1	P16220 EBI-711855	ATF1	P18846 EBI-852794	peptide array
●	🔍					peptide array
●	🔍					tandem affinity purification

Network visualisation



http://www.ebi.ac.uk

PPI Data Format

The first molecular interaction databases independently established their own dataset formats and curation strategies:

In 2002, The HUPO-Proteomics Standards Initiative (HUPO-PSI) defined community standards for data representation of proteomics data to facilitate data comparison, exchange and verification.

The development of PSI-MI XML schema has facilitated the description of protein-protein interactions.

An Excel-compatible, tab-delimited format, MITAB, has been developed for users who require only minimal information but in a more accessible configuration.

MITAB File

MI-TAB 2.7 Standard Culomns (+27)

- ID(S) INTERACTORS
- ALT. ID(S) INTERACTORS
- ALIAS(ES) INTERACTORS
- INTERACTION DETECTION METHOD(S)
- PUBLICATION FIRST AUTHOR(S)
- PUBLICATION IDENTIFIER(S)
- TAXID INTERACTORS
- INTERACTION TYPE
- SOURCE DATABASE(S)
- INTERACTION IDENTIFIER(S)
- CONFIDENCE VALUE(S) EXPANSION METHOD(S)
- BIOLOGICAL ROLE(S)
- EXPERIMENTAL ROLE(S)
- TYPE OF INTERACTORS
- PROPERTIES (CROSS REFERENCES) OF INTERACTORS/INTERACTION
- ANNOTATION OF INTERACTORS/INTERACTION
- HOST ORGANISM(S)
- PARAMETER OF INTERACTION
- FEATURE(S) INTERACTORS
- STOICHIOMETRY(S) INTERACTORS
- PARTECIPANT IDENTIFICATION METHODS

Ontology

- In computer science, ontology is a way to capture knowledge in a written and computable form.

- It is a formal naming and definition of the types, properties, and interrelationships of the entities.

- A common ontology defines the vocabulary with which queries and assertions are exchanged.

- In the PPI field, a common controlled vocabulary (CV) has been developed to standardize interaction data and to allow the systematic capture of the majority of experimental detail.

- Controlled vocabularies have a hierarchical structure and each object can be mapped to both parent and child terms.

Controlled Vocabulary

- The adoption of the CV enables users to search the data without having to select the correct synonym for a term (two hybrid or 2-hybrid or Y2H) or worry about alternative spelling, and allows the curators to uniformly annotate each experimental detail.

- Using the Interaction Type CV, it is possible to specify whether the experimental evidences have shown if the interaction between two molecules is direct (direct interaction, MI:0407) or only that the molecules are part of a large affinity complex (association, MI:0914).

- New experimental methodologies can be captured by the simple addition of an appropriate CV term, without a change to the data interchange format.

The Gene Ontology

- The Gene Ontology (GO) project is a major bioinformatics initiative to develop a computational representation of how genes encode biological functions at the molecular, cellular and tissue levels.

- The project has developed formal ontologies that represent over 40,000 biological concepts, and are constantly being revised to reflect new discoveries.

- The Gene Ontology project provides:

- Creation of a structured controlled vocabularies (ontologies) that describes gene products in terms of their associated biological processes, cellular components and molecular functions

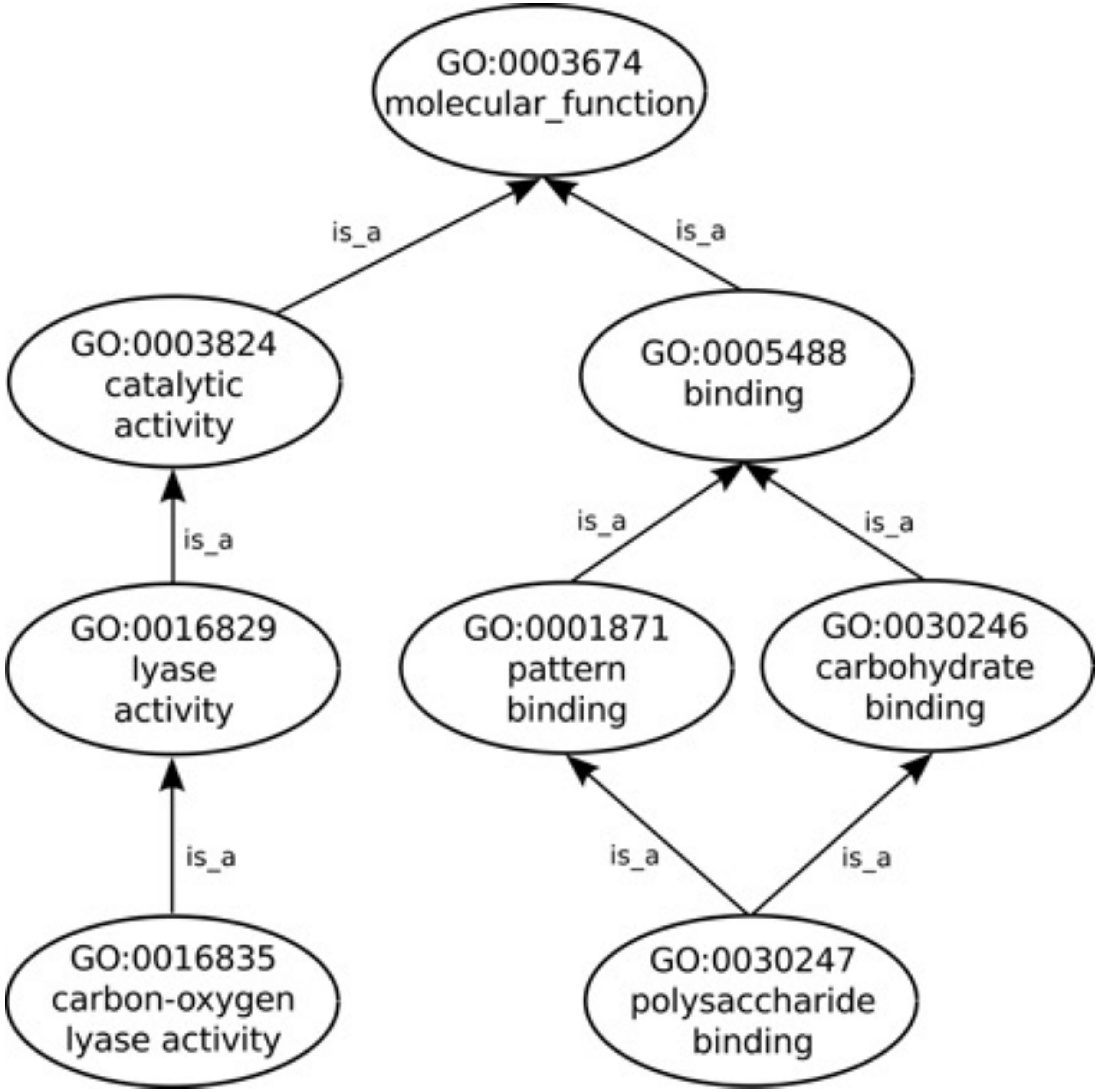
- The annotation of gene products using this ontology.

Gene Ontology Structure

GO terms are related within a hierarchy

GO terms are divided into three parts:

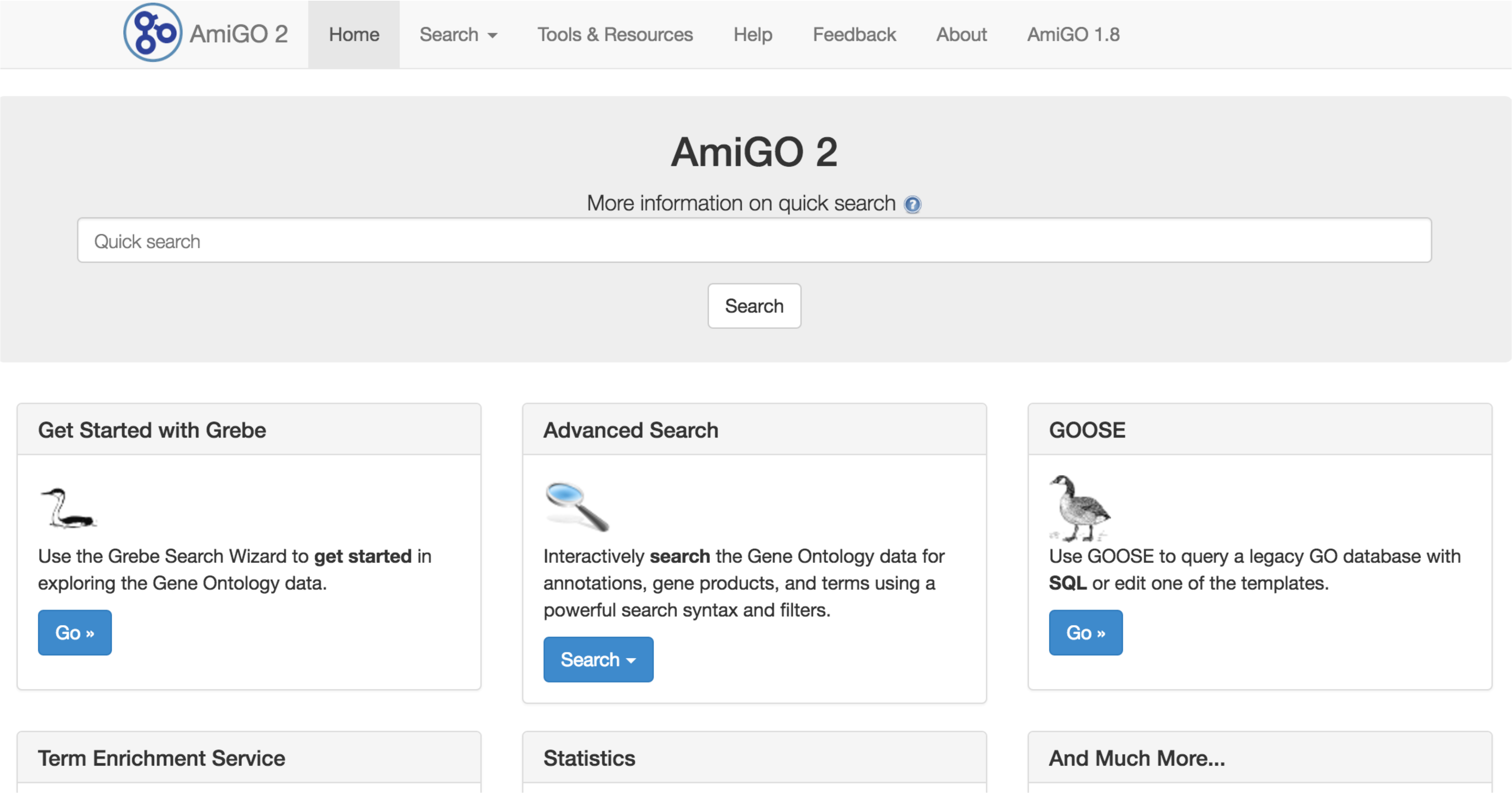
- Biological Process
  Pathways and larger processes made up of the activities of multiple genes
- Cellular Component
  Where does it act?
- Molecular Function
  Molecular activities of gene products



Luana Licata, Systems Biology Course 2015

AmiGO

AmiGo is a tool for browsing the Gene Ontology. Each function is represented by a code such as GO:0043403



AmiGO 2

More information on quick search

Get Started with Grebe

Use the Grebe Search Wizard to **get started** in exploring the Gene Ontology data.

Advanced Search

Interactively **search** the Gene Ontology data for annotations, gene products, and terms using a powerful search syntax and filters.

GOOSE

Use GOOSE to query a legacy GO database with **SQL** or edit one of the templates.

Term Enrichment Service

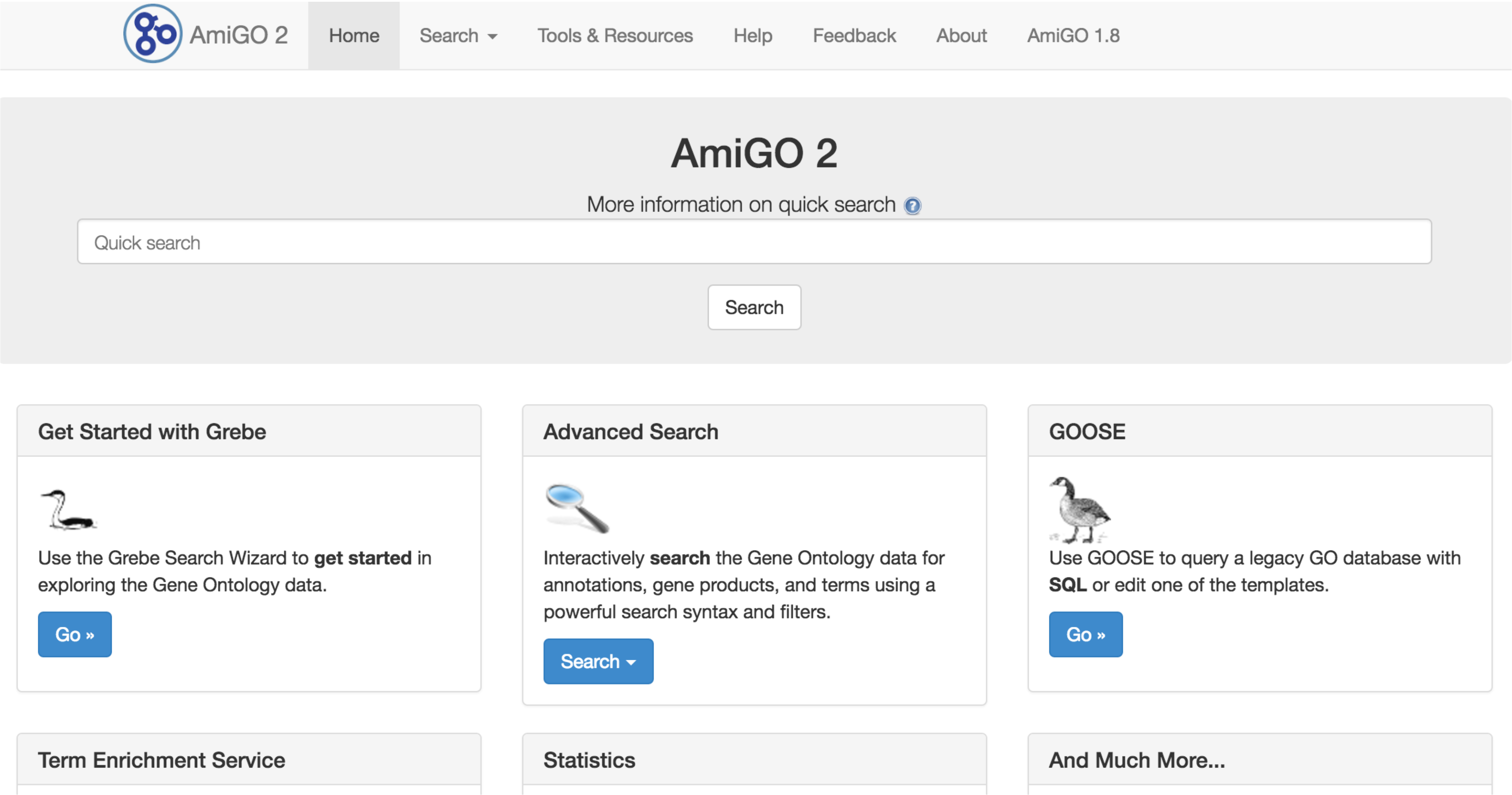
Statistics

And Much More...

http://amigo.geneontology.org/amigo/

AmiGO

AmiGo is a tool for browsing the Gene Ontology. Each function is represented by a code such as GO:0043403



AmiGO 2

More information on quick search

Get Started with Grebe

Use the Grebe Search Wizard to **get started** in exploring the Gene Ontology data.

Advanced Search

Interactively **search** the Gene Ontology data for annotations, gene products, and terms using a powerful search syntax and filters.

GOOSE

Use GOOSE to query a legacy GO database with **SQL** or edit one of the templates.

Term Enrichment Service

Statistics

And Much More...

http://amigo.geneontology.org/amigo/

Why the Gene Ontology

GO is used for several purposes:

- finding functional similarities in genes that are over-expressed or under-expressed in a specific condition (Enrichment analysis)
- integrating proteomic information from different organisms
- assigning functions to protein domains
- analyzing groups of genes that are co-expressed during development

Enrichment Analysis

Used to detect enrichment for a particular type of function in a set of genes.

The method uses statistical approaches to identify significantly enriched or depleted groups of genes.

Omics experiments always results often identify thousands of genes which are used for the analysis.

This analysis results in the calculation of an expected p-value, which indicates the over/underrepresentation for each term.

Exercise

- Search for the interactions of the MEKK1 protein.

How many interaction you can find? Are all this referring to the same protein?

- Refine your search using the UniProtID Q13233.

How many interaction you have now?

- Download the MI-TAB 2.7 file and search for the interaction with BRAF

How many experiments are supporting the existence of this interaction?

- Generate a list of the interacting genes.

which “biological process” is strongly enriched among those genes?