Babelomics

Protein-protein interactions and functional enrichment analysis
SNOW: Studying Networks in the Omic World

Buenos Aires, October 2011

Javier Santoyo-Lopez
jsantoyo@cipf.es
http://bioinfo.cipf.es
Genomics Department
Centro de Investigacion Principe Felipe (CIPF)
(Valencia, Spain)
Scenario

- Gene selection: clustering, differential expression
- Gene sorting: differential expression

- GO terms
- KEGG pathways
- Interpro motifs
- miRNAs
- TFBS
- Text-mining
- ...

Functional profiling

Protein-protein interactions

- Preprocessing
- Normalization
Protein-Protein Interactions (ppi)

Ppis are a central point at almost every level of cell function:

- **Structure** of subcellular organelles (structural proteins)
- **Transport** machinery across the various biological membranes (nuclear pore importins)
- **Packing** the chromatine (histones)
- **Signal transduction** (important in many diseases, eg. cancer)
- **Regulation of gene expression** (transcription and translation factors)
- **Protein modifications** (kinases)
Protein-Protein Interactions (ppi)

What we are going to explore in this session?

• How to extract information about *sets* of genes

• How to perform **functional enrichment analyses** using *protein-protein interactions* as annotation source:
  
  • Looking for *modules* of proteins
    
    1. With a cooperative behaviour
    2. With a statistical evaluation
What's a ppi?

A ppi is a physical interactions between two proteins.

Classification:
- Transient (weak) – eg. signalling cascades
- Stable (strong) – eg. protein complexes

Ppis are defined by pairwise interactions.

The Interactome is the complete net of interactions in a cellular system.
The interactome

The interactions are the real source of complexity of the cell.

So, ppi information can be the clue to explain your phenotype!

- 20,935 protein coding genes
- 650,000 predicted interactions
- 20,389 protein coding genes
- 240,000 predicted interactions
There is **not a common** repository for ppis. There are many different databases. Basically, **you have to create your own curated annotation**.

**Standard format to submit ppis**


Defines community standards for data representation in proteomics to facilitate data comparison, exchange and verification.

**Controlled Vocabulary: Molecular Interactions Ontology**

```
[Term]
id: MI:0018
name: two hybrid
def: "The classical two-hybrid system is a method that uses transcriptional activity as a measure of protein-protein interaction. It relies on the modular nature of many site-specific transcriptional activators (GAL 4), which consist of a DNA-binding domain and a transcriptional activation domain. The DNA-binding domain serves to target the activator to the specific genes that will be expressed, and the activation domain contacts other proteins of the transcriptional machinery to enable transcription to occur. The two-hybrid system is based on the observation that the two domains of the activator need to be non-covalently brought together by the interaction of any two proteins. The application of this system requires the expression of two hybrid. Generally this assay is performed in yeast cell, but it can also be carried out in other organism."
[PMID:10967325, PMID:12634794, PMID:1946372]
related_synonym: "2-hybrid" []
related_synonym: "2H" []
related_synonym: "2h" []
related_synonym: "classical two hybrid" []
related_synonym: "Gal4 transcription regeneration" []
related_synonym: "two-hybrid" []
related_synonym: "yeast two hybrid" []
extact_synonym: "2 hybrid" []
is_a: MI:0232 ! transcriptional complementation assay
```
Interactome Generation: DBs

Many databases, one controlled vocabulary, a standard format (PSI-MI) but ... still under development, still a challenging task

Problems with databases:

- No standardization of protein/gene ids.
- No applications of format standards.
- No good experiment annotation.
- No overlapping among databases.

ppis

ppis with experimental support, no predictions.

By now, only data from human.

Main 5 public DBs: HPRD, BIND, DIP, MINT, IntAct

HPRD: http://www.hprd.org
BIND: http://bond.unleashedinformatics.com/
DIP: http://dip.doe-mbi.ucla.edu/
MINT: http://mint.bio.uniroma2.it/mint/Welcome.do
IntAct: http://www.ebi.ac.uk/intact/
BRCA2 is your favourite protein, where do you find information about its interactions?

From:
- IntAct - http://www.ebi.ac.uk/intact
- HPRD - http://www.hprd.org
- MINT - http://mint.bio.unroma2.it
- BIND - http://bond.unleashedinformatics.com
- DIP - http://dip.doe-mbi.ucla.edu
- BioGrid - http://thebiogrid.org
• Information about a single protein

BRCA2 is your favourite protein, where do you find information about its interactions?

From:
- **IntAct** - [http://www.ebi.ac.uk/intact](http://www.ebi.ac.uk/intact)
- **HPRD** - [http://www.hprd.org](http://www.hprd.org)
- **MINT** - [http://mint.bio.uniroma2.it](http://mint.bio.uniroma2.it)
- **BIND** - [http://bond.unleashedinformatics.com](http://bond.unleashedinformatics.com)
- **DIP** - [http://dip.doe-mbi.ucla.edu](http://dip.doe-mbi.ucla.edu)
- **BioGrid** - [http://thebiogrid.org](http://thebiogrid.org)

93 interactions
Information about a single protein

BRCA2 is your favourite protein, where do you find information about its interactions?

From:
- IntAct - http://www.ebi.ac.uk/intact
- HPRD - http://www.hprd.org
- MINT - http://mint.bio.uniroma2.it
- BIND - http://bond.unleashedinformatics.com
- DIP - http://dip.doe-mbi.ucla.edu
- BioGrid - http://thebiogrid.org

24 interactions
Information about a single protein

BRCA2 is your favourite protein, where do you find information about its interactions?

From:
- IntAct - http://www.ebi.ac.uk/intact
- HPRD - http://www.hprd.org
- MINT - http://mint.bio.uniroma2.it
- BIND - http://bond.unleashedinformatics.com
- DIP - http://dip.doe-mbi.ucla.edu
- BioGrid - http://thebiogrid.org

6 interactions
Information about a single protein

BRCA2 is your favourite protein, where do you find information about its interactions?

From:
- IntAct - http://www.ebi.ac.uk/intact
- HPRD - http://www.hprd.org
- MINT - http://mint.bio.uniroma2.it
- BIND - http://bond.unleashedinformatics.com
- DIP - http://dip.doe-mbi.ucla.edu
- BioGrid - http://thebiogrid.org

<table>
<thead>
<tr>
<th>Mol A Short Label</th>
<th>Mol B Short Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>FANCG</td>
<td>BRCA2</td>
</tr>
<tr>
<td>BRCA2</td>
<td>hsFLNa</td>
</tr>
<tr>
<td>FANCD2</td>
<td>BRCA2</td>
</tr>
<tr>
<td>BRCA2</td>
<td>BRCA2</td>
</tr>
<tr>
<td>BRCA2</td>
<td>RAD51</td>
</tr>
<tr>
<td>SMAD3</td>
<td>BRCA2</td>
</tr>
<tr>
<td>BRCA2</td>
<td>DSS1</td>
</tr>
<tr>
<td>CDK2/CCNA2</td>
<td>BRCA2</td>
</tr>
<tr>
<td>BRCA2</td>
<td>RAD51</td>
</tr>
<tr>
<td>BRCA2</td>
<td>RAD51</td>
</tr>
<tr>
<td>BRCA2</td>
<td>SHFDG1</td>
</tr>
<tr>
<td>SHFDG1</td>
<td>BRCA2</td>
</tr>
<tr>
<td>Brh2</td>
<td>Rad51</td>
</tr>
<tr>
<td>RAD51</td>
<td>RAD51</td>
</tr>
<tr>
<td>BRCA2</td>
<td>RAD51</td>
</tr>
<tr>
<td>RAB163</td>
<td>RAD51</td>
</tr>
<tr>
<td>Pol II</td>
<td>BRCA2 promoter</td>
</tr>
</tbody>
</table>

59 interactions
• Information about a single protein

BRCA2 is your favourite protein, where do you find information about its interactions?

From:
- IntAct - http://www.ebi.ac.uk/intact
- HPRD - http://www.hprd.org
- MINT - http://mint.bio.uniroma2.it
- BIND - http://bond.unleashedinformatics.com
- DIP - http://dip.doe-mbi.ucla.edu
- BioGrid - http://thebiogrid.org

<table>
<thead>
<tr>
<th>Protein: BRCA2 protein</th>
<th>Cross Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Binary</td>
<td>Complex</td>
</tr>
<tr>
<td>DIP:57452E</td>
<td>DIP:29383N</td>
</tr>
<tr>
<td>DIP:40108E</td>
<td>DIP:462N</td>
</tr>
<tr>
<td>DIP:40109E</td>
<td>DIP:368N</td>
</tr>
<tr>
<td>DIP:76301E</td>
<td>DIP:5971N</td>
</tr>
<tr>
<td>DIP:103802E</td>
<td>DIP:38427N</td>
</tr>
</tbody>
</table>
Information about a single protein

BRCA2 is your favourite protein, where do you find information about its interactions?

From:
- IntAct - http://www.ebi.ac.uk/intact
- HPRD - http://www.hprd.org
- MINT - http://mint.bio.uniroma2.it
- BIND - http://bond.unleashedinformatics.com
- DIP - http://dip.doe-mbi.ucla.edu
- BioGrid - http://thebiogrid.org

Displaying 37 total unique interactors

**RAD51** | HRAD51, RECA, BRCC5, HsT16930, HsRad51, RAD51A
RAD51 homolog (RecA homolog, E. coli) (S. cerevisiae)

**HMG20B** | SMARCE1r, HMGXB2, FLJ26127, SOXL, PP7706, HMGX2, BRAF25, BRAF35, pp8857
SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily E, member 1-related

**BRCA1** | RNF53, IRIS, BRCC1, PSCP, PNCA4, BRCAI, BROVCA1
breast and ovarian cancer susceptibility protein 1
Information about a single protein

BRCA2 is your favourite protein, where do you find information about its interactions? Choose your favourite!!
• Information about a single protein

Some resources that collect (and predict) interactions (not only physical)

✔ STRING - http://string-db.org/
✔ GeneMania - http://genemania.org
✔ APID - http://bioinfow.dep.usal.es/apid
✔ GeneCards - http://www.genecards.org
No one has a complete coverage of the known ppis.

There is not a complete interactome in a single DB.
Interactome generation: Type of interactome

We generated two types of interactome

- **Non-curated interactome**: All interactions (no filters, but no redundancies).
- **Curated interactome**: Filter on the methodologies used for detecting the interactions.
We use Molecular Interactions (MI) Ontology to identify the type of experiment that detect the ppi. MI has a tree structure.

Some experiments can't be considered totally independent.

Different databases - different criteria and different depth of annotation.

- **HPRD** (the largest DB) uses MI in a very restricted way, its annotations lay just into three categories (*in vivo, in vitro and y2h*).
Interactome generation: Methods

Yeast two-hybrid assay

A. Regular transcription of the reporter gene

B. One fusion protein only (Gal4-BD + Bait) - no transcription

C. One fusion protein only (Gal4-AD + Prey) - no transcription

D. Two fusion proteins with interacting Bait and Prey

If bait catch the prey (interaction) a reporter gene is expressed.
Techniques to explore protein-protein interactions

**Non-screening techniques:** Co-immunoprecipitation, fluorescence resonance energy transfer, Dual polarization interferometry.

But most of the data we find in the databases come from **high-throughput techniques:**

- Bio-molecular fluorescence complementation.
- Yeast two-hybrid (Y2H).
- Tandem affinity purification (TAP).
- High-throughput mass spectrometry.
- Cross-linking.

from the Molecular Interactions Ontology
Interactome generation: Methods

Molecular Interactions (MI) Ontology Browser

- **Associated information**
  - **definition**: Methods based on laboratory experiments to determine an interaction
  - **preferred name**: experimental interaction detection
  - **Unique short label curated by PSI-MI synonym**: experimental interaction detection
  - **subset**: Subset of PSI-MI
  - **xref_definition**: PMID:14755292

- **Term Hierarchy**
  - **Paths to Root**: molecular interaction
  - **Child relationships**: part_of
    - interaction detection method
      - experimental interaction detection
        - biophysical
        - genetic interference
        - post transcriptional interference
        - imaging technique
        - interaction prediction
        - inference
        - unspecified method
        - participant identification method
        - feature detection method
        - feature type
        - interaction type
        - alias type
        - interactor type
        - feature range status
        - experimental preparation
        - cross-reference type
        - database citation
        - experimental role
        - biological role
        - attribute name
        - parameter type
        - parameter unit
        - curation quality

Legend:
- **is a**
- **develops from**
- **part of**
- **other**
We selected the 6 main parent types of interaction experiments plus in vivo and in vitro (from MI ontology).

To each interaction in a database we assign one of these categories ascending in the tree.

Then just get the interactions that are annotated with two or more of these top categories.
Interactome generation: Methods

Molecular Interactions (MI) Ontology:

**biophysical**: The application of *physical principles* and methods to biological experiments.

**protein complementary assay**: The function of numerous proteins or ribonucleic particles (enzymes, transcription factors, and others) can be rationally dissected into two fragments that fold autonomously but cannot complement to reconstitute the complex function, unless they are located in close proximity. In a two hybrid experiment, *restoration of the activity by complementation of the two fragments* when expressed as fusion with two polypeptides is taken as an evidence that the two polypeptides interact together.

**genetic interference**: This term refers to methods that aim at *interfering with the activity of a specific gene by altering the gene regulatory or coding sequences*. This goal can be achieved either by a classical genetic approach (random mutagenesis followed by phenotype characterization and genetic mapping) or by a reverse genetics approach where a gene of interest is modified by directed mutagenesis.

**post transcriptional interference**: This term refers to methods designed to *interfere with gene expression at post-transcriptional level rather than with the gene itself*.

**biochemical**: The application of *chemical principles* and methods to biological experiments.

**imaging techniques**: Methods that provide images of molecules at various resolution depending on the technology used.
Transcripts and “Genes” Interactomes

**Transcripts** have a **one to one** relationship with proteins.

We chose transcripts as our core ID because their possibilities to convert to other IDs and to genes using Ensembl.

**Genes** have a **one to many** relationship with proteins. There is not a real interactome for genes but users sometimes have gene lists.

Mapping genes (many potential proteins) into a proteins interactome may give a fake highly connected network.

To avoid this situation we generated a proteins and a “genes” interactome.
# Interactome generation: Interactomes

<table>
<thead>
<tr>
<th></th>
<th>Non-curated interactome</th>
<th>Curated interactome</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Transcripts</td>
<td>Genes</td>
</tr>
<tr>
<td>Nodes</td>
<td>16.799</td>
<td>10.027</td>
</tr>
<tr>
<td>Edges</td>
<td>109.709</td>
<td>46.799</td>
</tr>
</tbody>
</table>
Graph Theory

Nodes = proteins
Edges = interaction events

Set of proteins interacting

Undirected graph

structured data
Graph theory may help us to study protein networks

Some interesting parameters

- **Degree (connectivity or connections)**: number of edges connected to a node. Nodes with high degree are called **hubs**.

- **Betweenness**: A measure of centrality of a node, it is defined by:

  \[ C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}} \]

  \( \sigma_{st} \) is total number of shortest paths in the graph.

  \( \sigma_{st}(V) \) is the number of shortest paths that pass through node V
- **Clustering coefficient** (of a node): A measure of how interconnected the neighbours of that node are. Proportion of links between the nodes within its neighbourhood divided by the number of links that could possibly exist between them.

\[
C_i = \frac{2e_i}{n_i(n_i - 1)}
\]

- $e_i$ is the number of edges among the nodes connected to node $i$
- $n_i$ is the number of neighbours of node $i$

To differentiate between star-shaped nets and more interconnected nets.
Graph Theory

Some Graph Theory Concepts

**Shortest path:** The path with less edges that connects two nodes.

**Component:** A group of nodes connected among them.

**Bicomponent:** A group of nodes connected to other group of nodes by only an edge. The edge that joins two bicomponents is called *articulation point.*
Graph Theory

Components, bicomponents & articulation points
- **Interactome.** Complete collection of protein-protein interactions in the cell.

- **Transcriptome** determines the real interactome.
How we evaluate the cooperative behaviour of a list of proteins/genes in terms of its ppi network parameters?

Two different approximations

- Importance in complete interactome
- Cooperative behaviour - Minimal Connected Network
The list of proteins mapped into the complete interactome may provide clues about the importance of certain proteins.

Importance in complete interactome

Hubs: nodes with high degree of connections

Nodes very central (high betweenness)

External proteins
Comparison of parameters distributions of the lists versus complete interactome distributions applying a Kolmogorov-Smirnov test.

- **betweenness**: pval = 0.0058
- **connections degree**: pval = 0.0013
- **clustering coefficient**: pval = 0.0055

Results indicate whether the set of prots/genes are collectively ... 

- In a central position in the interactome (betweenness).
- If they are highly connected (degree of connection).
- If they are in a very connected area (cluster coefficient).
Moreover, nets may reveal a **cooperative behaviour**.

Proteins external in the complete interactome **but** they reveal an interesting subnetwork.

It seems we have have found something about the cooperative behaviour of our list!!!
Interactome & transcriptome: MCN

**Minimal connected network (MCN)**

Minimal network that connects a set of proteins.

**Minimal Connected Network generation**

- Find all shortest paths for all the pairs of nodes.
- Accept paths that connect two proteins in the list either directly or through a predetermined number of not-in-list proteins (0-1).
MCN evaluation

Parameters to evaluate: degree, betweenness, clustering coefficient & number of components.

Comparison of parameters distributions of the network versus set of same size random distributions applying a Kolmogorov-Smirnov test.

You can use lists up to 200 nodes (prots/genes) with interactomic data (the list can be bigger).
Interactome & transcriptome: MCN

betweenness

connections degree

clustering coefficient

pval < 0.001

pval < 0.001

pval < 0.001

Minimal Connected Network topological evaluation

- Number of components with more than 1 node: 1
- Number of components [95% confidence interval]: 12 [19, 34]
Interactome & transcriptome: MCN

This tells us whether the complete list of proteins is ...

- more/less connected than a random network (degree, connections)
- have more/less nodes with high centrality than a random network (betweenness)
- well interconnected or there are just a few nodes that contribute to the connections (cluster coefficient)
- a compact network or a set of small networks (number of components)

Moreover...
Finding statistically significance in the different parameters points to different possible topologies of the network.

High Connections degree
Low Clustering coefficient

High Betweenness
Low Connections

hub

signaling cascade
SNOW

Originally implemented as an stand-alone application

http://snow.bioinfo.cipf.es/cgi-bin/snow.cgi

Currently fully integrated in Babelomics v4.

http://snow.babelomics.org

Functional analysis > protein-protein interaction > SNOW

Recommended parameters to run Snow:
- Interactome: Interactions detected by at least two methodologies
- Maximum number of external proteins introduced: 1
SNOW:
Web form
Snow - Results

Role in the interactome

betweenness  |  connections degree  |  clustering coefficient

betweenness  |  connections degree  |  clustering coefficient

pval = 0.0058  |  pval = 0.0013  |  pval = 0.0055

Evaluation of the MCN

betweenness  |  connections degree  |  clustering coefficient

betweenness  |  connections degree  |  clustering coefficient

pval < 0.001  |  pval < 0.001  |  pval < 0.001

Information about components

- **Minimal Connected Network topological evaluation**
  - Number of components with more than 1 node: 1
  - Number of components (95% confidence interval): 12 [19, 34]
Snow - Network visualization

Better with Google Chrome!!!

Gene Ontology (GO)

**Go biological process**

- **GO:0006511**: The chemical reactions and pathways resulting in the breakdown of a protein or peptide by hydrolysis of its peptide bonds, initiated by the covalent attachment of a ubiquitin moiety, or multiple ubiquitin moieties, to the protein.
- **GO:0031145**: The chemical reactions and pathways resulting in the breakdown of a protein or peptide by hydrolysis of its peptide bonds, initiated by the covalent attachment of ubiquitin, with ubiquitin–protein ligation catalyzed by the anaphase-promoting complex, and mediated by the proteasome.
- **GO:0051436**: Any process that stops, prevents or reduces the frequency, rate or extent of ubiquitin ligase activity during the mitotic cell cycle.
- **GO:0051437**: Any process that activates, maintains or increases the rate of ubiquitin ligase activity during the mitotic cell cycle.
- **GO:0070536**: A protein deubiquitination process in which a K63-linked ubiquitin chain, i.e. a polymer of ubiquitin formed by linkages between lysine residues at position 63 of the ubiquitin monomers, is removed from a protein.

**Go cellular component**

- **GO:0005002**: A large multisubunit complex which catalyzes protein degradation. This complex consists of the barrel-shaped proteasome core complex and one or two
Snow - Editing your network

Search for any field in the nodes

Apoptosis

- GO:0006916: A process which directly inhibits any of the steps required for cell death by apoptosis.
- GO:0042981: Any process that modulates the occurrence or rate of cell death by apoptosis.
- GO:0006917: A process that directly activates any of the steps required for cell death by apoptosis.
- GO:0008630: A cascade of processes initiated by the detection of DNA damage and resulting in the induction of apoptosis (programmed cell death).
- GO:0043065: Any process that activates or increases the frequency, rate or extent of cell death by apoptosis.

Edit

- Show or hide nodes:
  - Show
  - Hide
- Change node shape:
  - Square
  - Circle
- Change color nodes:
  - #123456
Snow - Editing your network
Compare GO enrichment analysis versus ppi enrichment analysis

We run FatiGO to a set of lists from microarray experiments and to co-expression modules in cancer.

Compare “positive” results given by both analysis ... but

• A positive result in FatiGO is when we find a GO over-represented.
• A positive result in Snow is ... ¿?

What is a positive result?

We decided to stablish a criteria to decide when a Snow analysis of a list of proteins give a positive result.

Our criteria:
MCN has more degree (connections) and less components than the set of random networks.

Betweenness & clustering coefficient are not included because the are more related to the topology of the network.
Snow Exercises

http://babelomics.bioinfo.cipf.es/functional.html

Functional analysis > protein-protein interaction > SNOW

On-line examples

- Example 1: Downregulated in fibroblasts from old individuals, compared to young
- Example 2: Upregulated by induction of exogenous BRCA1 in EcR-293 cells
2. Some datasets to run SNOW

Here are several examples of lists of genes selected to differentiate two samples in microarray experiments. The description of the experiment is given.

The SNOW parameters used to perform the analyses were:

- Interactome of reference: ppis detected by two methods.
- Maximum number of external proteins: 1
- Nature of the lists: Genes

Download the lists and perform your own SNOW analyses choosing same or different parameters. For a reference we give the results pages as you will obtain them, have a look at them and compare them with SNOW results using different parameters taking into account that results shown here may have been run with different version of ppi data.

<table>
<thead>
<tr>
<th>Example number</th>
<th>Dataset</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>brca1_overexp_up</td>
<td>Upregulated by induction of exogenous BRCA1 in EcR-293 cells</td>
</tr>
<tr>
<td>2.2</td>
<td>brca1_overexp_dn</td>
<td>Downregulated by induction of exogenous BRCA1 in EcR-293 cells</td>
</tr>
<tr>
<td>2.3</td>
<td>serum_fibroblast_cellcycle</td>
<td>Cell-cycle dependent genes regulated following exposure to serum in a variety of human fibroblast cell lines</td>
</tr>
<tr>
<td>2.4</td>
<td>ageing_brain_dn</td>
<td>Age-downregulated in the human frontal cortex</td>
</tr>
<tr>
<td>2.5</td>
<td>brca1_sw480_up</td>
<td>Up-regulated by infection of human colon adenocarcinoma cells (SW480) with Ad-BRCA1 , versus Ad-LacZ control</td>
</tr>
</tbody>
</table>