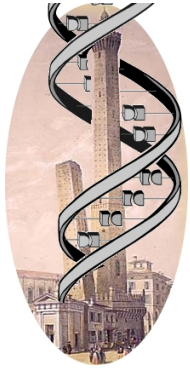


ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Integration of PPI sources and prediction methods

NET-GE: a novel NETwork-based Gene Enrichment for detecting biological processes associated to Mendelian diseases

Pietro Di Lena

pietro.dilena@unibo.it

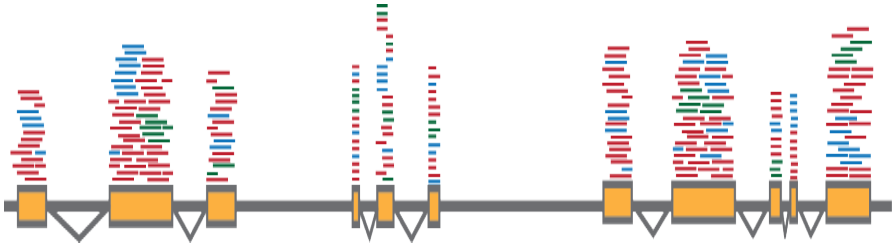
Department of Computer Science and Engineering
University of Bologna

December 18, 2015

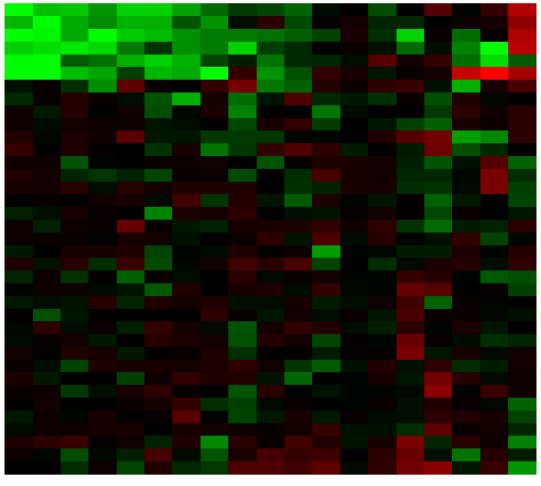
Outline

1. Background on enrichment analysis.
   - What is gene set enrichment?
   - How to perform gene set enrichment analysis?
2. Background on network-based enrichment analysis.
   - How can we exploit network information for gene set enrichment?
   - How to perform network-based gene set enrichment analysis?
   - NET-GE: NETwork-based Gene Enrichment
3. Tutorial on NET-GE (presented by Samuele Bovo).

Introduction



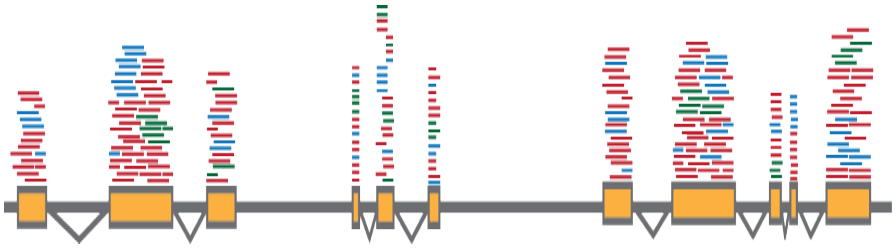
RNA-Seq/ChIP-Seq



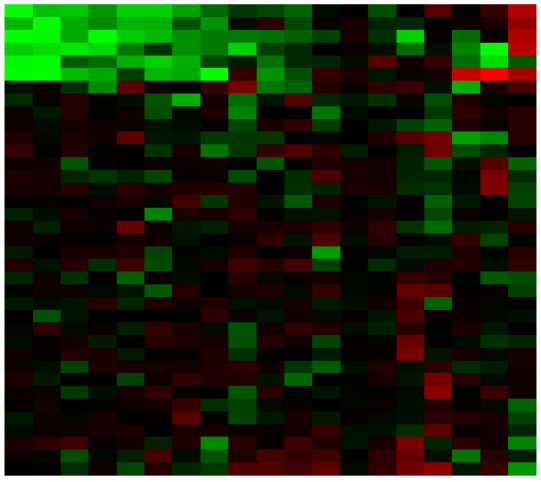
Micro-arrays

Long list of ranked genes

Introduction



RNA-Seq/ChIP-Seq

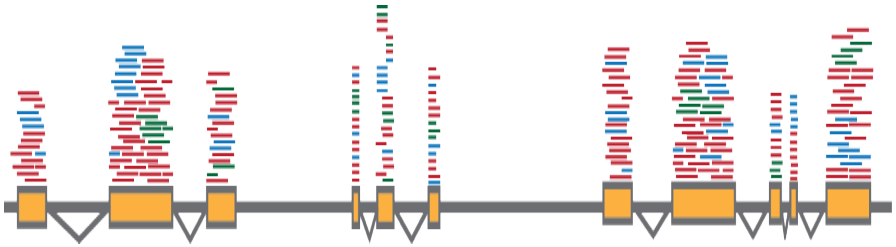


Micro-arrays

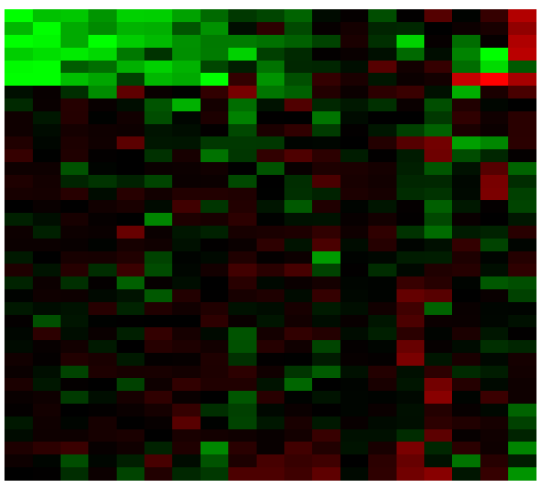
Gene	Log Ratio	p-value
Abcg1	-2.09614	4.72E-07
Adamts5	2.483321	1.33E-07
Alox12b	-2.41347	3.59E-07
Arg1	-2.27214	3.06E-07
AU018091	2.048711	4.62E-07
Bex1	2.591349	4.08E-07
Degs2	-2.46253	1.54E-07
Klk7	-2.18902	3.77E-07
Krt78	-2.89916	2.18E-07
Ly6c1	3.085592	9.41E-08
Ly6g6c	-2.55108	3.62E-07
Sdr16c6	-2.16277	4.05E-07
Sdr9c7	-2.25984	2.63E-07
Sept5	-2.08797	6.31E-07
Kprp	-2.34542	6.77E-07
Ly6a	2.839925	6.04E-07
Slc2a3	2.199118	6.52E-07
Sprr2i	-2.22872	5.67E-07
Mxd1	-1.77522	9.66E-07
Cidea	-1.93749	1.20E-06
Krt16	-1.91642	1.24E-06
Krt8	2.057569	1.22E-06
Trex2	-1.71243	1.29E-06
Aldh3b2	-1.7556	2.63E-06
Asprv1	-1.56796	2.35E-06

Long list of ranked genes

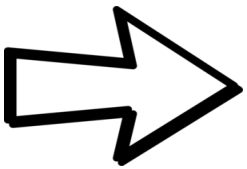
Introduction



RNA-Seq/ChIP-Seq

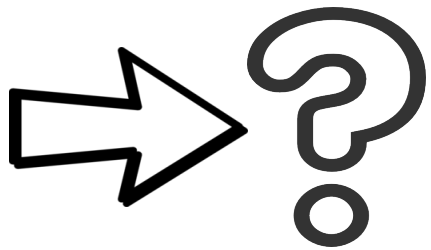


Micro-arrays



Gene	Log Ratio	p-value
Abcg1	-2.09614	4.72E-07
Adamts5	2.483321	1.33E-07
Alox12b	-2.41347	3.59E-07
Arg1	-2.27214	3.06E-07
AU018091	2.048711	4.62E-07
Bex1	2.591349	4.08E-07
Degs2	-2.46253	1.54E-07
Klk7	-2.18902	3.77E-07
Krt78	-2.89916	2.18E-07
Ly6c1	3.085592	9.41E-08
Ly6g6c	-2.55108	3.62E-07
Sdr16c6	-2.16277	4.05E-07
Sdr9c7	-2.25984	2.63E-07
Sept5	-2.08797	6.31E-07
Kprp	-2.34542	6.77E-07
Ly6a	2.839925	6.04E-07
Slc2a3	2.199118	6.52E-07
Sprr2i	-2.22872	5.67E-07
Mxd1	-1.77522	9.66E-07
Cidea	-1.93749	1.20E-06
Krt16	-1.91642	1.24E-06
Krt8	2.057569	1.22E-06
Trex2	-1.71243	1.29E-06
Aldh3b2	-1.7556	2.63E-06
Asprv1	-1.56796	2.35E-06

Long list of ranked genes



How can we extract knowledge from a *noisy* list of genes?

Enrichment analysis

- **Enrichment analysis** is a technique to identify whether a set of genes or proteins *share some common property*.

Enrichment analysis

- **Enrichment analysis** is a technique to identify whether a set of genes or proteins *share some common property*.
- Enrichment analysis tools make use of **statistical approaches** to identify **functional annotations** that are *overrepresented* in a set of genes or proteins.

Enrichment analysis

- **Enrichment analysis** is a technique to identify whether a set of genes or proteins *share some common property*.
- Enrichment analysis tools make use of **statistical approaches** to identify **functional annotations** that are *overrepresented* in a set of genes or proteins.
- **Functional annotations**:
  - Gene Ontology
  - Biological pathways (KEGG, REACTOME, ..)
  - In general, any gene or protein classification scheme.

Enrichment analysis

- **Enrichment analysis** is a technique to identify whether a set of genes or proteins *share some common property*.
- Enrichment analysis tools make use of **statistical approaches** to identify **functional annotations** that are *overrepresented* in a set of genes or proteins.
- **Functional annotations**:
  - Gene Ontology
  - Biological pathways (KEGG, REACTOME, ..)
  - In general, any gene or protein classification scheme.
- **Statistical approaches**:
  - Fisher's Exact Test (Hypergeometric Test)
  - Kolmogorov-Smirnov (KS) Test
  - Z-test
  - Chi-square Test
  - All statistical approaches require *multiple testing correction*.

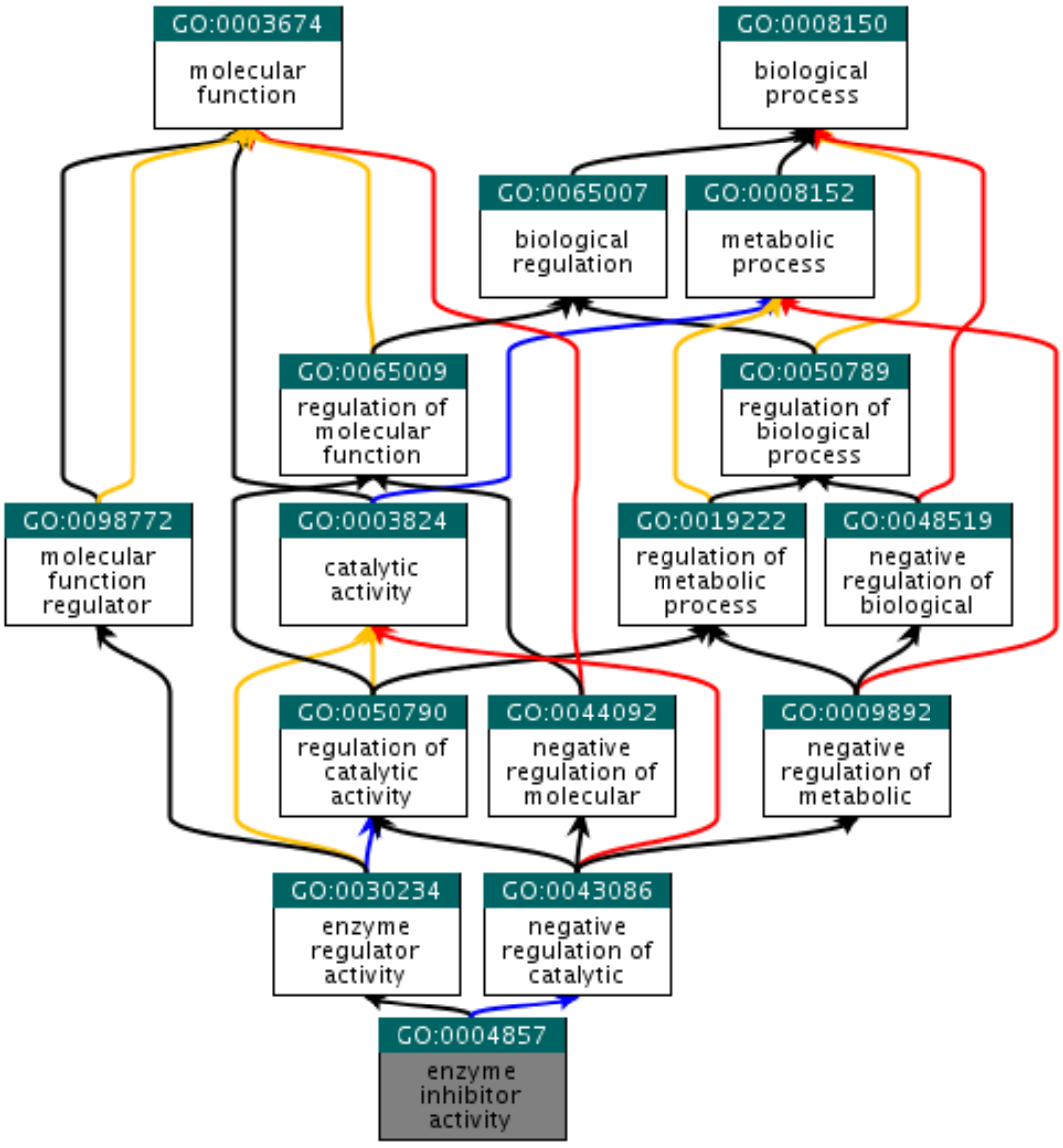
Functional annotations: Gene Ontology

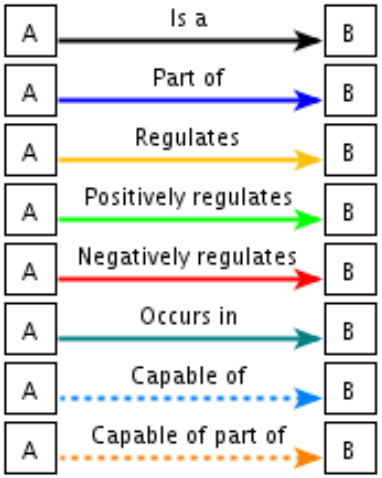
- **Gene Ontology (GO) project**
  - Major collaborative initiative, started in 1998, with the aim to unify the representation of gene and gene product attributes across all species.
- **Goals of the GO project**
  1. development and maintenance of a controlled vocabulary (ontologies) of functional attributes,
  2. annotation of genes in terms of the their associated attributes.

Functional annotations: Gene Ontology

- **Gene Ontology (GO) project**
  - Major collaborative initiative, started in 1998, with the aim to unify the representation of gene and gene product attributes across all species.
- **Goals of the GO project**
  1. development and maintenance of a controlled vocabulary (ontologies) of functional attributes,
  2. annotation of genes in terms of the their associated attributes.
- **Gene ontology directed acyclic graph (DAG)**
  - **Biological Process** (BP). A biological process term describes a series of events accomplished by one or more organized assemblies of molecular functions. A biological process is not equivalent to a pathway.
  - **Molecular Function** (MF). Molecular Function terms describes activities that occur at the molecular level.
  - **Celular Component** (CC). Cellular Component terms describe a component of a cell that is part of a larger object, such as an anatomical structure or a gene product group.

Functional annotations: Gene Ontology

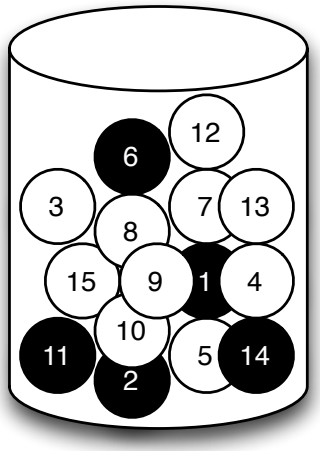


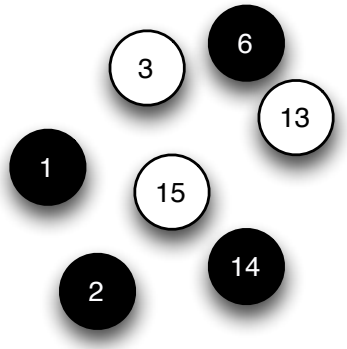


Statistical approaches: Fisher's Exact Test

- **Fisher's exact test** (1922) is a statistical significance test.
- Exact test: the significance of the deviation from the **null hypothesis** can be computed exactly.
- Very useful to assess the significance of the association between two kinds of classifications.
- Very useful when sample sized are small.
- When the sample sizes are large a **chi-squared test** provides a good approximation.

Statistical approaches: Fisher's Exact Test





- *What is the probability of getting exactly 4 black balls if we randomly select 7 balls from the bowl on the left?*

	Black	White
Selected	4	3
Non Selected	1	7

$$prob_4 = \frac{\binom{4+3}{4}\binom{1+7}{1}}{\binom{4+3+1+7}{4+1}} = 0.093$$

- *What is the probability of getting exactly 5 black balls if we randomly select 7 balls from the bowl on the left?*

	Black	White
Selected	5	2
Non Selected	0	8

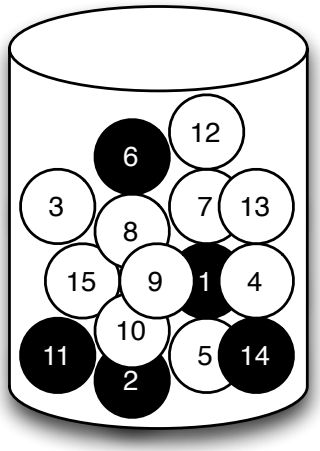
$$prob_5 = \frac{\binom{5+2}{5}\binom{0+8}{8}}{\binom{5+3+0+8}{5+0}} = 0.007$$

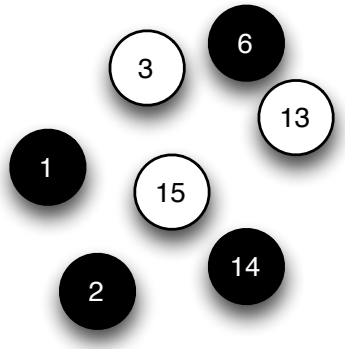
- **Are black balls over-represented in the set of 7 balls on the right?**

$$prob_4 + prob_5 = 0.093 + 0.007 = 0.1$$

$\Longrightarrow$ *We expect to to see this configuration one out of ten tries.*

Statistical approaches: Fisher's Exact Test





- *What is the probability of getting exactly 4 black balls if we randomly select 7 balls from the bowl on the left?*

	Black	**White**
Selected	4	3
Non Selected	1	7

$$prob_4 = \frac{\binom{4+3}{4}\binom{1+7}{1}}{\binom{4+3+1+7}{4+1}} = 0.093$$

- *What is the probability of getting exactly 5 black balls if we randomly select 7 balls from the bowl on the left?*

	Black	**White**
Selected	5	2
Non Selected	0	8

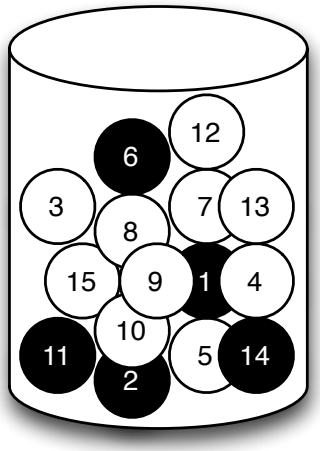
$$prob_5 = \frac{\binom{5+2}{5}\binom{0+8}{8}}{\binom{5+3+0+8}{5+0}} = 0.007$$

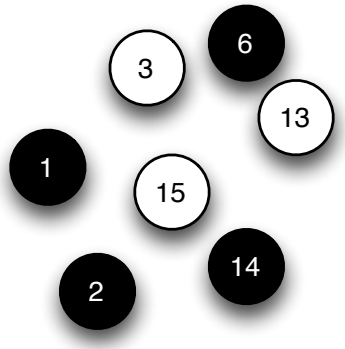
- **Are black balls over-represented in the set of 7 balls on the right?**

$$prob_4 + prob_5 = 0.093 + 0.007 = 0.1$$

$\Longrightarrow$ *We expect to to see this configuration one out of ten tries.*

Statistical approaches: Fisher's Exact Test





- *What is the probability of getting exactly 4 black balls if we randomly select 7 balls from the bowl on the left?*

	Black	White
Selected	4	3
Non Selected	1	7

$$prob_4 = \frac{\binom{4+3}{4}\binom{1+7}{1}}{\binom{4+3+1+7}{4+1}} = 0.093$$

- *What is the probability of getting exactly 5 black balls if we randomly select 7 balls from the bowl on the left?*

	Black	White
Selected	5	2
Non Selected	0	8

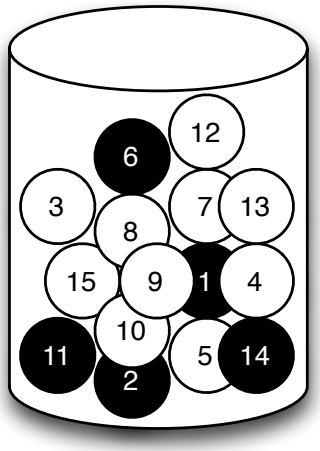
$$prob_5 = \frac{\binom{5+2}{5}\binom{0+8}{8}}{\binom{5+3+0+8}{5+0}} = 0.007$$

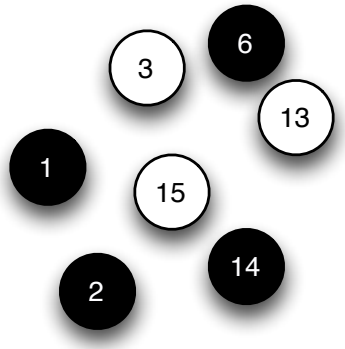
- **Are black balls over-represented in the set of 7 balls on the right?**

$$prob_4 + prob_5 = 0.093 + 0.007 = 0.1$$

$\Longrightarrow$ *We expect to to see this configuration one out of ten tries.*

Statistical approaches: Fisher's Exact Test





- *What is the probability of getting exactly 4 black balls if we randomly select 7 balls from the bowl on the left?*

	Black	White
Selected	4	3
Non Selected	1	7

$$prob_4 = \frac{\binom{4+3}{4}\binom{1+7}{1}}{\binom{4+3+1+7}{4+1}} = 0.093$$

- *What is the probability of getting exactly 5 black balls if we randomly select 7 balls from the bowl on the left?*

	Black	White
Selected	5	2
Non Selected	0	8

$$prob_5 = \frac{\binom{5+2}{5}\binom{0+8}{8}}{\binom{5+3+0+8}{5+0}} = 0.007$$

- **Are black balls over-represented in the set of 7 balls on the right?**

$$prob_4 + prob_5 = 0.093 + 0.007 = 0.1$$

$\Longrightarrow$ *We expect to to see this configuration one out of ten tries.*

Statistical approaches: enrichment analysis with Fisher's Exact Test

- **Problem**.
  - We want to asses whether there is some **statistically overrepresented** GO term in a list of Human proteins.
- In order to perform a Fisher test, we need the following information:
  1. **Classes**.
     - List of GO terms we want to consider for the enrichment analysis.
  2. **Background**.
     - List of Human proteins and their GO annotations.
  3. **Test set**.
     - Set of Human proteins on which to perform enrichment analysis.

Example of enrichment analysis with Fisher's Test 1/2

- **Class**. The *cellular response to erythropoietin* (GO:0036018) GO term.
- **Background**. List of of 20,814 human proteins in the UniProtKB.
- **Test set**. We consider the following list of 10 proteins (UniProt AC).

  P80297 P02795 P02810 P04921 P05060
  P05162 P05204 P06028 P0C2L3 P0C2S0

- Is the GO:0036018 term overrepresented in the test set?
  - Two test proteins (P80297, P02795) associated with GO:0036018.
  - Four proteins in UniprotKB associated with the GO:0036018 term.

	GO:0036018	**not GO:0036018**
Selected	2	8
not Selected	2	20802

$$\text{Fisher's test p-value} = 1.25 \times 10^{-6}$$

  - Conclusion: the probability of randomly selecting from the UniProtKB a set of 10 human proteins, containing two or more proteins associated with the GO:0036018 term is lower than 1.25×10^{-6} (**very unlikely**).

Example of enrichment analysis with Fisher's Test 1/2

- **Class**. The *cellular response to erythropoietin* (`GO:0036018`) GO term.
- **Background**. List of of 20,814 human proteins in the UniProtKB.
- **Test set**. We consider the following list of 10 proteins (UniProt AC).

`P80297 P02795 P02810 P04921 P05060`
`P05162 P05204 P06028 P0C2L3 P0C2S0`

- Is the `GO:0036018` term overrepresented in the test set?
  - Two test proteins (`P80297`, `P02795`) associated with `GO:0036018`.
  - Four proteins in UniprotKB associated with the `GO:0036018` term.

	GO:0036018	**not GO:0036018**
Selected	2	8
not Selected	2	20802

$$\text{Fisher's test p-value} = 1.25 \times 10^{-6}$$

  - Conclusion: the probability of randomly selecting from the UniProtKB a set of 10 human proteins, containing two or more proteins associated with the `GO:0036018` term is lower than 1.25×10^{-6} (**very unlikely**).

Example of enrichment analysis with Fisher's Test 1/2

- **Class**. The *cellular response to erythropoietin* (GO:0036018) GO term.
- **Background**. List of of 20,814 human proteins in the UniProtKB.
- **Test set**. We consider the following list of 10 proteins (UniProt AC).

P80297 P02795 P02810 P04921 P05060
P05162 P05204 P06028 P0C2L3 P0C2S0

- Is the GO:0036018 term overrepresented in the test set?
  - Two test proteins (P80297, P02795) associated with GO:0036018.
  - Four proteins in UniprotKB associated with the GO:0036018 term.

	GO:0036018	**not GO:0036018**
Selected	2	8
not Selected	2	20802

$$\text{Fisher's test p-value} = 1.25 \times 10^{-6}$$

- Conclusion: the probability of randomly selecting from the UniProtKB a set of 10 human proteins, containing two or more proteins associated with the GO:0036018 term is lower than 1.25×10^{-6} (**very unlikely**).

Example of enrichment analysis with Fisher's Test 1/2

- **Class**. The *cellular response to erythropoietin* (GO:0036018) GO term.
- **Background**. List of of 20,814 human proteins in the UniProtKB.
- **Test set**. We consider the following list of 10 proteins (UniProt AC).

P80297 P02795 P02810 P04921 P05060
P05162 P05204 P06028 P0C2L3 P0C2S0

- Is the GO:0036018 term overrepresented in the test set?
  - Two test proteins (P80297, P02795) associated with GO:0036018.
  - Four proteins in UniprotKB associated with the GO:0036018 term.

	GO:0036018	**not GO:0036018**
Selected	2	8
not Selected	2	20802

$$\text{Fisher's test p-value} = 1.25 \times 10^{-6}$$

  - Conclusion: the probability of randomly selecting from the UniProtKB a set of 10 human proteins, containing two or more proteins associated with the GO:0036018 term is lower than 1.25×10^{-6} (**very unlikely**).

Example of enrichment analysis with Fisher's Test 2/2

- **Class**. The *cellular response to interleukin-3* (GO:0036016) GO term.
- **Background**. List of of 20,814 human proteins in the UniProtKB.
- **Test set**. We consider the following list of 10 proteins (UniProt AC).

P80297 P02795 P02810 P04921 P05060
P05162 P05204 P06028 P0C2L3 P0C2S0

- *Is the GO:0036016 term overrepresented in the test set?*
  - One test protein (P02795) associated with the GO:0036016 term.
  - Five proteins in UniprotKB associated with the GO:0036016 term.

	GO:0036016	not GO:0036016
Selected	1	9
not Selected	4	20800

Fisher's test p-value $= 0.0024$

  - Conclusion: the probability of randomly selecting from the UniProtKB a set of 10 human proteins, containing one or more proteins associated with the GO:0036016 term is lower than 0.0024 (**unlikely**).

Example of enrichment analysis with Fisher's Test 2/2

- **Class**. The *cellular response to interleukin-3* (`GO:0036016`) GO term.
- **Background**. List of of 20,814 human proteins in the UniProtKB.
- **Test set**. We consider the following list of 10 proteins (UniProt AC).

`P80297 P02795 P02810 P04921 P05060`
`P05162 P05204 P06028 P0C2L3 P0C2S0`

- *Is the `GO:0036016` term overrepresented in the test set?*
  - One test protein (`P02795`) associated with the `GO:0036016` term.
  - Five proteins in UniprotKB associated with the `GO:0036016` term.

	GO:0036016	**not GO:0036016**
Selected	1	9
not Selected	4	20800

Fisher's test p-value $= 0.0024$

  - Conclusion: the probability of randomly selecting from the UniProtKB a set of 10 human proteins, containing one or more proteins associated with the `GO:0036016` term is lower than 0.0024 (**unlikely**).

Example of enrichment analysis with Fisher's Test 2/2

- **Class**. The *cellular response to interleukin-3* (`GO:0036016`) GO term.
- **Background**. List of of 20,814 human proteins in the UniProtKB.
- **Test set**. We consider the following list of 10 proteins (UniProt AC).

`P80297 P02795 P02810 P04921 P05060`
`P05162 P05204 P06028 P0C2L3 P0C2S0`

- *Is the `GO:0036016` term overrepresented in the test set?*
  - One test protein (`P02795`) associated with the `GO:0036016` term.
  - Five proteins in UniprotKB associated with the `GO:0036016` term.

	GO:0036016	**not GO:0036016**
Selected	1	9
not Selected	4	20800

Fisher's test p-value $= 0.0024$

  - Conclusion: the probability of randomly selecting from the UniProtKB a set of 10 human proteins, containing one or more proteins associated with the `GO:0036016` term is lower than 0.0024 (**unlikely**).

Example of enrichment analysis with Fisher's Test 2/2

- **Class**. The *cellular response to interleukin-3* (`GO:0036016`) GO term.
- **Background**. List of of 20,814 human proteins in the UniProtKB.
- **Test set**. We consider the following list of 10 proteins (UniProt AC).

`P80297 P02795 P02810 P04921 P05060`
`P05162 P05204 P06028 P0C2L3 P0C2S0`

- *Is the `GO:0036016` term overrepresented in the test set?*
  - One test protein (`P02795`) associated with the `GO:0036016` term.
  - Five proteins in UniprotKB associated with the `GO:0036016` term.

	GO:0036016	**not GO:0036016**
Selected	1	9
not Selected	4	20800

Fisher's test p-value $= 0.0024$

  - Conclusion: the probability of randomly selecting from the UniProtKB a set of 10 human proteins, containing one or more proteins associated with the `GO:0036016` term is lower than 0.0024 (**unlikely**).

Statistical analysis: Multiple Testing Correction

- Enrichment analysis involves multiple testings at the same time.
  - Given a list of candidate genes, we typically perform statistical tests against a large number of GO terms or pathways, not a single one.

Statistical analysis: Multiple Testing Correction

- Enrichment analysis involves multiple testings at the same time.
  - Given a list of candidate genes, we typically perform statistical tests against a large number of GO terms or pathways, not a single one.
- Multiple statistical comparisons need p-value adjustments.
  - Suppose we want to perform enrichment analysis against a huge number of GO terms.
  - As more GO terms are considered, it becomes more likely that a term results enriched in the test set.

Statistical analysis: Multiple Testing Correction

- Enrichment analysis involves multiple testings at the same time.
  - Given a list of candidate genes, we typically perform statistical tests against a large number of GO terms or pathways, not a single one.
- Multiple statistical comparisons need p-value adjustments.
  - Suppose we want to perform enrichment analysis against a huge number of GO terms.
  - As more GO terms are considered, it becomes more likely that a term results enriched in the test set.
- There are several statistical techniques to account for the multiple comparison problem.
  - Bonferroni correction (1961)
  - Šidárk correction (1967)
  - Benjamini-Hockberg correction (1995)
  - Storey correction (2003)
  - ... many more

Example of multiple testing correction in enrichment analysis

- **Bonferroni correction**.
  - Procedure: multiply the p-value by the number of tested hypothesis.
  - Problems: very conservative, high risk of introducing *false negatives*.

Example of multiple testing correction in enrichment analysis

- **Bonferroni correction**.
  - Procedure: multiply the p-value by the number of tested hypothesis.
  - Problems: very conservative, high risk of introducing *false negatives*.
- Recall the two previous examples.
- Assume we are testing 12,785 GO BP terms (as of September 2014).
  1. Statistical significance of the `GO:0036018` term.

$$\begin{aligned} \textbf{Fisher p-value} &= 1.25 \times 10^{-6} \\ \textbf{Corrected p-value} &= 1.25 \times 10^{-6} \times 12785 = 0.016 \end{aligned}$$

Example of multiple testing correction in enrichment analysis

- **Bonferroni correction**.
  - Procedure: multiply the p-value by the number of tested hypothesis.
  - Problems: very conservative, high risk of introducing *false negatives*.
- Recall the two previous examples.
- Assume we are testing 12,785 GO BP terms (as of September 2014).
  1. Statistical significance of the `GO:0036018` term.

$$\begin{aligned} \textbf{Fisher p-value} &= 1.25 \times 10^{-6} \\ \textbf{Corrected p-value} &= 1.25 \times 10^{-6} \times 12785 = 0.016 \end{aligned}$$

  2. Statistical significance of the `GO:0036016` term.

$$\begin{aligned} \textbf{Fisher p-value} &= 0.0024 \\ \textbf{Corrected p-value} &= 0.0024 \times 12785 = 30.68 > 1 \end{aligned}$$

Example of multiple testing correction in enrichment analysis

- **Bonferroni correction**.
  - Procedure: multiply the p-value by the number of tested hypothesis.
  - Problems: very conservative, high risk of introducing *false negatives*.
- Recall the two previous examples.
- Assume we are testing 12,785 GO BP terms (as of September 2014).
  1. Statistical significance of the `GO:0036018` term.

$$\begin{aligned} \textbf{Fisher p-value} &= 1.25 \times 10^{-6} \\ \textbf{Corrected p-value} &= 1.25 \times 10^{-6} \times 12785 = 0.016 \end{aligned}$$

  2. Statistical significance of the `GO:0036016` term.

$$\begin{aligned} \textbf{Fisher p-value} &= 0.0024 \\ \textbf{Corrected p-value} &= 0.0024 \times 12785 = 30.68 > 1 \end{aligned}$$

- Conclusions. In our test set of 10 Human proteins:
  1. it is unlikely to observe the `GO:0036018` term by chance;
  2. it is not surprising at all to observe the `GO:0036016` term.

Enrichment analysis tools

Table 1. List of 68 enrichment tools

Enrichment tool name	Year of release	Key statistical method	Category
FunSpec	2002	Hypergeometric	Class I
Onto-express	2002	Fisher's exact; hypergeometic; binomial; chi-square	Class I
EASE	2003	Fisher's exact (modified as EASE score)	Class I
FatiGO/FatiWise/FatiGO+	2003	Fisher's exact	Class I
FuncAssociate	2003	Fisher's exact	Class I
GARBAN	2003	Hypergeometric	Class I
GeneMerge	2003	Hypergeometric	Class I
GoMiner	2003	Fisher's exact	Class I
MAPPFinder	2003	*Z*-score; hypergeometric	Class I
CLENCH	2004	Hypergeometric; chi-square; binomial	Class I
GO::TermFinder	2004	hypergeometric	Class I
GOAL	2004	Permutation	Class I
GOArray	2004	Hypergeometric; *Z*-score; permutation	Class I
GOStat	2004	Fisher's exact; chi-squre	Class I
GoSurfer	2004	Chi-square	Class I
OntologyTraverser	2004	Hypergeometric; Fisher's exact	Class I
THEA	2004	Hypergeometric	Class I
BiNGO	2005	Hypergeometric; binomial	Class I
FACT	2005	Adopt GeneMerge and GO::TermFinder statistical modules	Class I
gfinder	2005	Fisher's exact	Class I
Gobar	2005	Hypergeometric	Class I
GOCluster	2005	Hypergeometric	Class I
GOSSIP	2005	Fisher's exact	Class I
L2L	2005	Binomial; hypergeometric	Class I
WebGestalt	2005	Hypergeometric	Class I
BayGO	2006	Bayesian; Goodman and Kruskal's gamma factor	Class I
eGOn/GeneTools	2006	Fisher's exact	Class I
Gene Class Expression	2006	*Z*-statistics	Class I
GOALIE	2006	Hidden Kripke model	Class I
GOFFA	2006	Fisher's inverse chi-square	Class I
GOLEM	2006	Hyerpgeometric	Class I
IPreGO	2006	Fisher's exact; Kolmogorov–Smirnov test	Class I

Huang, Sherman, Lempicki. *Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists*. **Nucleic Acids Research** 2009, 37, 1-17.

Network-based enrichment

- **Standard enrichment**.
  - Genes/proteins are treated as isolated objects.
  - Completely neglects the different types of relations among molecules.

Network-based enrichment

- **Standard enrichment**.
  - Genes/proteins are treated as isolated objects.
  - Completely neglects the different types of relations among molecules.
- **Network-based enrichment**
  - Gene/proteins are treated in the context of their interaction networks.
  - Physical interaction networks, gene regulatory networks, metabolic and signaling pathways can help in extracting new biological information.

Network-based enrichment

- **Standard enrichment**.
  - Genes/proteins are treated as isolated objects.
  - Completely neglects the different types of relations among molecules.
- **Network-based enrichment**
  - Gene/proteins are treated in the context of their interaction networks.
  - Physical interaction networks, gene regulatory networks, metabolic and signaling pathways can help in extracting new biological information.
  - *What kind of interaction networks?*
  - *How do we extract information from interaction networks?*

Interaction/pathway/ontology databases

Table 1. Overview of common data resources for the functional interpretation of protein lists

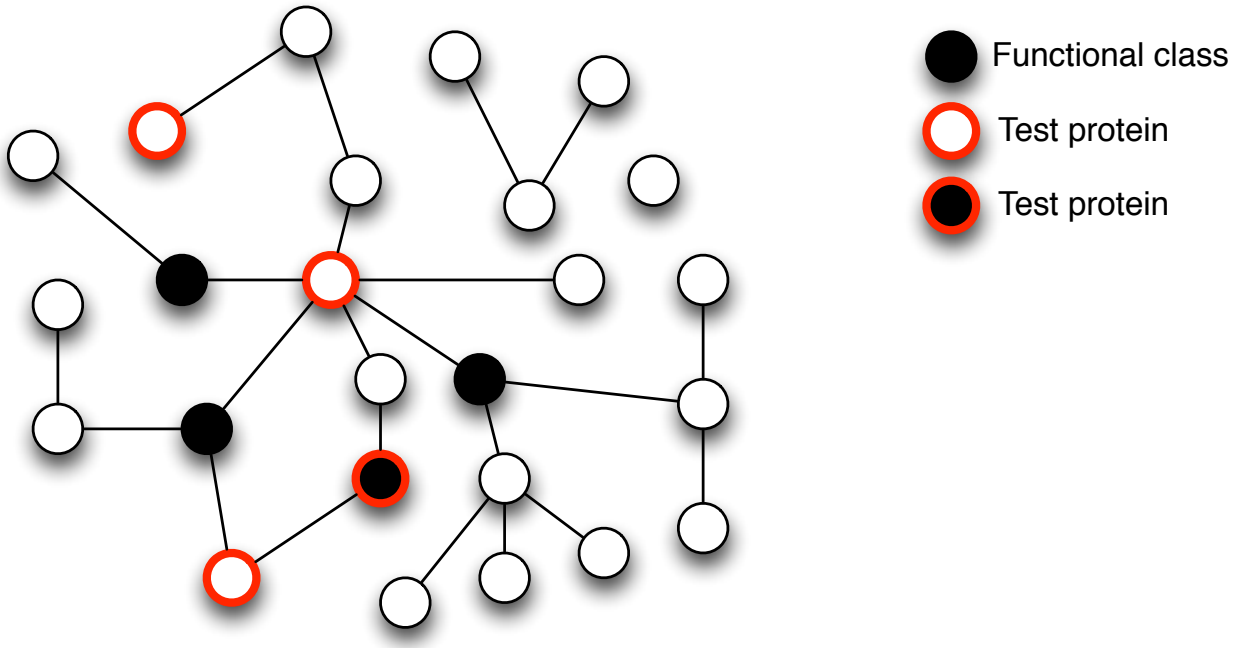
Content	Name	Reference	Availability[a)]	Link
Interaction databases				
Protein interactions	BioCarta	[173]	Mixed	http://www.biocarta.com/
	BioGrid	[134]	Public	http://thebiogrid.org
	DIP	[132]	Public	http://bond.unleashedinformatics.com/Action?
	HPRD	[133]	Public	http://www.hprd.org
	IntAct	[135]	Public	http://www.ebi.ac.uk/intact/
	MINT	[136]	Public	http://mint.bio.uniroma2.it/mint/Welcome.do
	String	[142]	Public	http://string-db.org
Pathways	BIND Translation	[174]	Public	http://baderlab.org/BINDTranslation
	BioCyc	[120]	Public	http://biocyc.org
	Ingenuity IPA		Commercial	http://www.ingenuity.com/products/ipa
	INOH	[175]	Public	http://inoh.hgc.jp/download.html
	KEGG	[72]	Mixed	http://www.genome.jp/kegg/kegg2.html
	NetPro		Commercial	http://www.molecularconnections.com/home/en/home/products/netPro
	Pathway Commons	[121]	Public	http://www.pathwaycommons.org/about/
	Panther Pathways	[122]	Public	http://www.pantherdb.org/pathway/
	Reactome	[118]	Public	http://www.reactome.org
	WikiPathways	[119]	Public	http://wikipathways.org/index.php/WikiPathways
Ontology databases				
Functions	Gene Ontology	[42]	Public	http://www.geneontology.org
Protein evolution	Protein Ontology	[143]	Public	http://pir.georgetown.edu/pro/pro.shtml
Aggregated databases				
	UniProt	[41]	Public	http://www.uniprot.org
	NCBI Protein	[176]	Public	http://www.ncbi.nlm.nih.gov/protein

a) Availability under the assumption of academic, noncommercial use.

Laukens et al. *Bioinformatics approaches for the functional interpretation of protein lists: From ontology term enrichment to network analysis*. **Proteomics**, 15, 981-996.

Network-based enrichment approaches 1/2

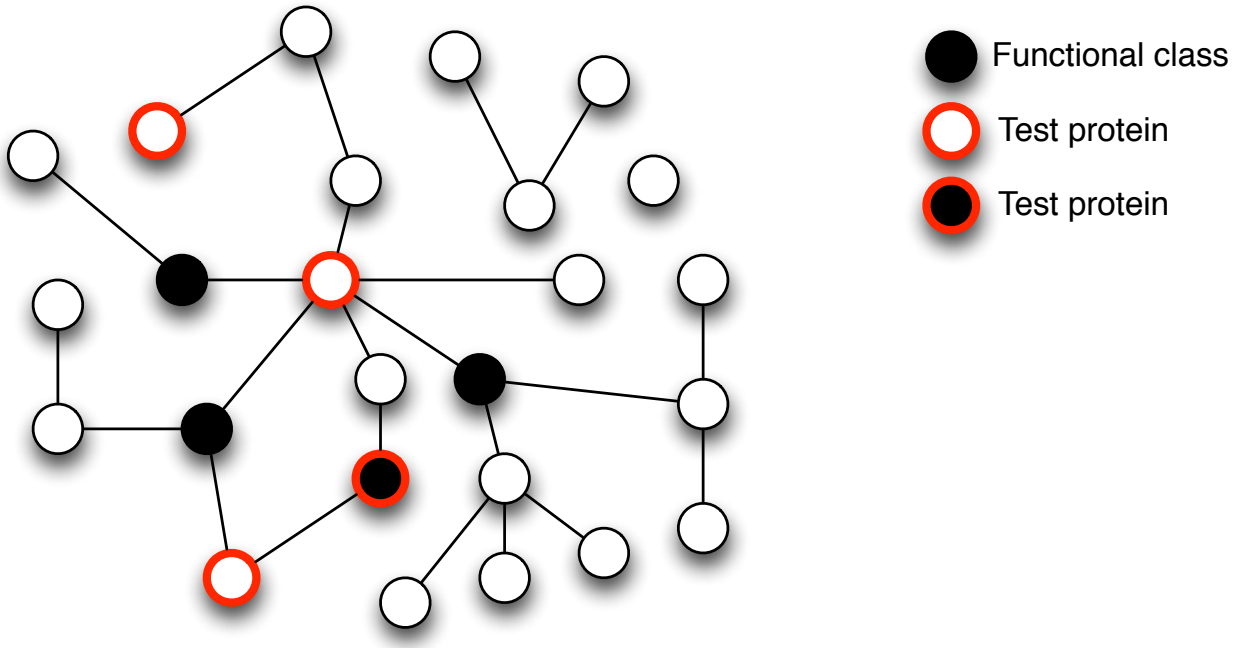
- **Class A**
  - Use the topology of the interaction network to infer how much similar distinct sets of gene/proteins are.



How **distant** is the set of test proteins (red-bordered circles) from a functional class of proteins (black circles)?

Network-based enrichment approaches 1/2

- **Class A**
  - Use the topology of the interaction network to infer how much similar distinct sets of gene/proteins are.

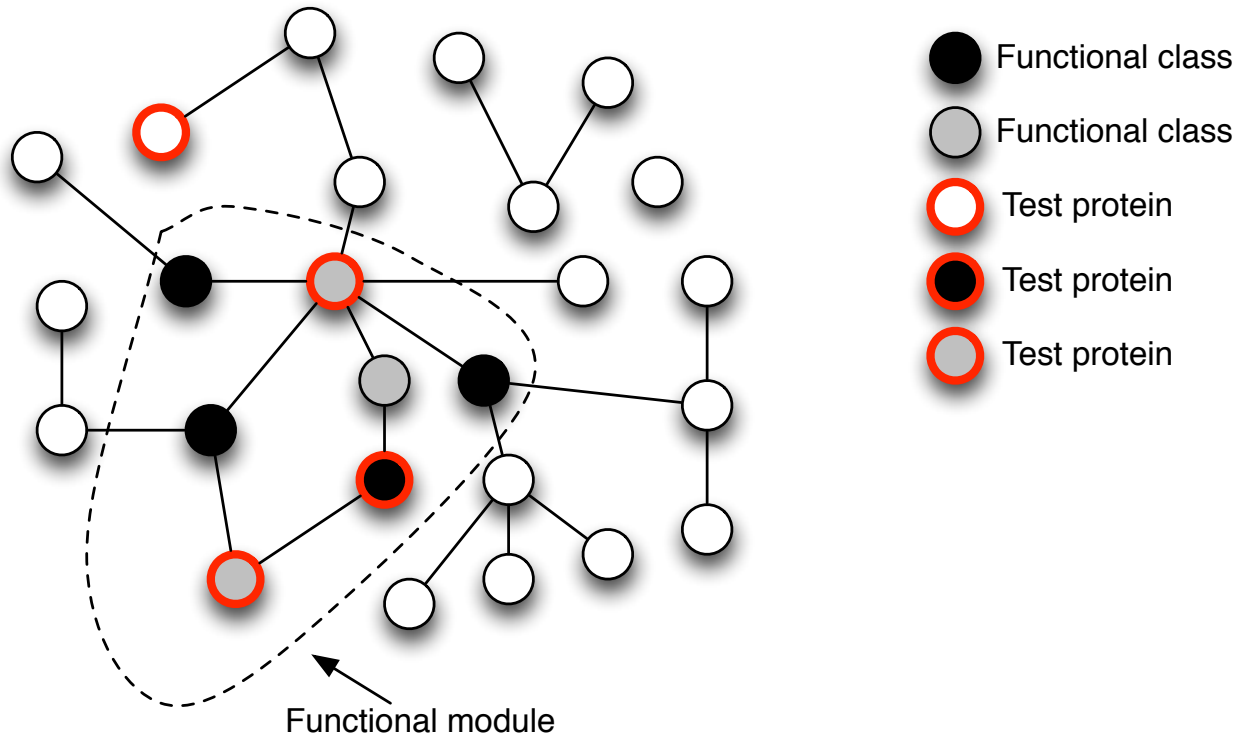


How **distant** is the set of test proteins (red-bordered circles) from a functional class of proteins (black circles)?

- Problem: *how to define a graph-based distance?*

Network-based enrichment approaches 2/2

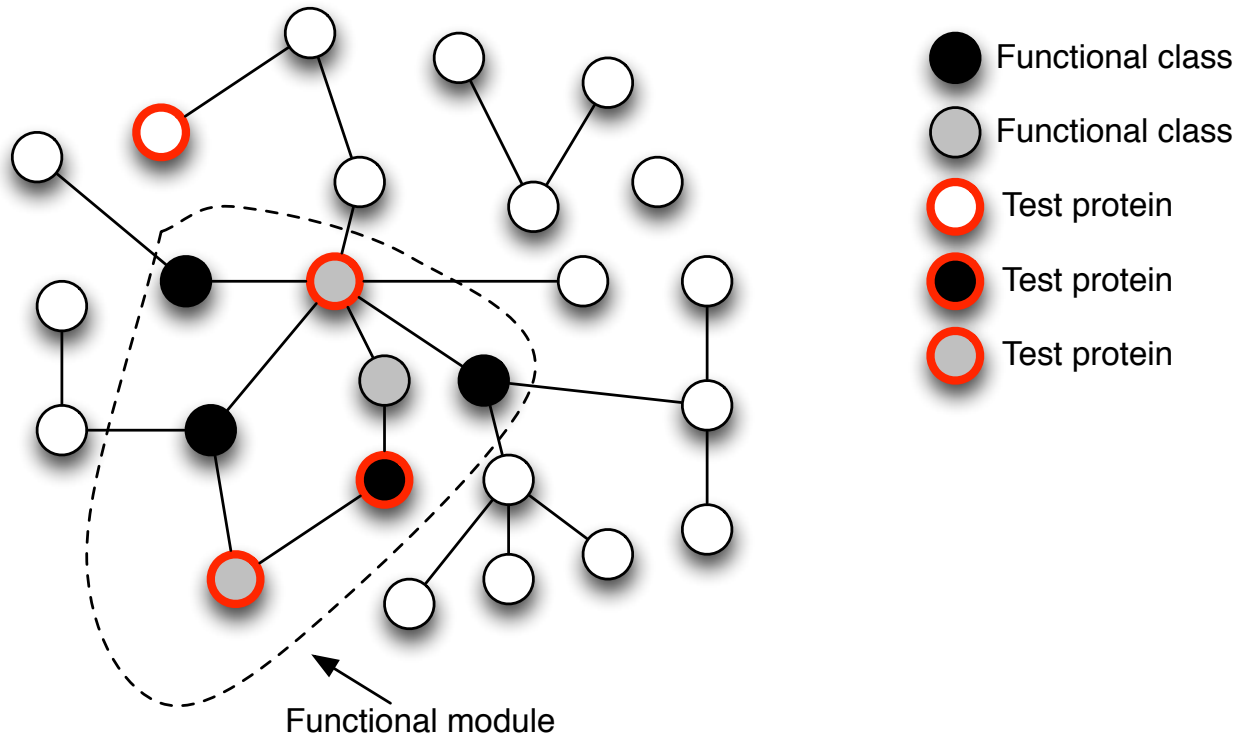
- **Class B**
  - Identify functionally-related modules in interaction net- works and then infer protein/gene biological roles from such modules.



Is the set of test proteins (red-bordered circles) overrepresented in a **functional module** (black and gray circles)?

Network-based enrichment approaches 2/2

- **Class B**
  - Identify functionally-related modules in interaction net- works and then infer protein/gene biological roles from such modules.



Is the set of test proteins (red-bordered circles) overrepresented in a **functional module** (black and gray circles)?

- Problem: *how to extract a functional module in an interaction graph?*

Network-based enrichment analysis tools

Table 2. Overview of commonly used computational tools for the functional interpretation of protein lists

Section	Name	Reference	Availability[a)]	Link
Keyword enrichment	agriGO	[55]	Public	http://bioinfo.cau.edu.cn/agriGO/
	BinGO	[49]	Public	http://apps.cytoscape.org/apps/bingo
	ClueGO	[54]	Public	http://apps.cytoscape.org/apps/cluego
	DAVID	[52]	Public	http://david.abcc.ncifcrf.gov
	FuncAssociate	[46]	Public	http://llama.mshri.on.ca/cgi/func/funcassociate
	GenMAPP	[51]	Public	http://www.genmapp.org
	GOEAST	[53]	Public	http://omicslab.genetics.ac.cn/GOEAST/
	GOminer	[48]	Public	http://discover.nci.nih.gov/gominer/index.jsp
	GOStat	[177]	Public	http://gostat.wehi.edu.au
	Ingenuity IPA		Commercial	http://www.ingenuity.com/products/ipa
	WebGestalt	[57]	Public	http://bioinfo.vanderbilt.edu/webgestalt/
	WEGO	[50]	Public	http://wego.genomics.org.cn/cgi-bin/wego/index.pl
Set enrichment	ASSESS	[74]	Public	http://people.genome.duke.edu/assess/
	EnrichR	[139]	Public	http://amp.pharm.mssm.edu/Enrichr
	FatiScan	[81]	Public	http://www.gepas.org
	GAGE	[82]	Public	http://www.bioconductor.org/packages/release/bioc/html/gage.html
	Gazer	[75]	Public	http://expressome.kobic.re.kr/GAzer/document.jsp
	GLOBALTEST	[178]	Public	http://www.bioconductor.org/packages/2.0/bioc/html/globaltest.html
	GSEA	[83]	Public	http://www.broadinstitute.org/gsea/index.jsp
	PAGE	[80]	On request	
	GSVA	[179]	Public	http://www.bioconductor.org/packages/release/bioc/html/GSVA.html
Text mining	Anni	[89]	Public	http://biosemantics.org/anni/
	CoCiter	[100]	Public	http://www.picb.ac.cn/hanlab/cociter/
	CoPub	[15]	Public	http://copub.gatcplatform.nl
	FAUN	[97]	Public	https://grits.eecs.utk.edu/faun/
	Martini	[93]	Public	http://martini.embl.de
	LAITOR	[98]	Public	http://laitor.sourceforge.net
	PESCADOR	[99]	Public	http://cbdm.mdc-berlin.de/tools/pescador/
	TXTGate	[96]	Public	http://tomcat.esat.kuleuven.be/txtgate/
Network enrichment	DEGraph	[128]	Public	http://www.bioconductor.org/packages/release/bioc/html/DEGraph.html
	EnrichNet	[58]	Public	http://www.enrichnet.org
	MetaCore	[180]	Commercial	https://portal.genego.com
	NetGSA	[127]	Public	http://www.biostat.washington.edu/~ashojaie/research.html
	PathNet	[124]	Public	http://www.bioconductor.org/packages/release/bioc/html/PathNet.html
	Pathway-Express	[113]	Registration	http://vortex.cs.wayne.edu/projects.htm
	PWEA	[126]	Public	http://zlab.bu.edu/PWEA/
	SCORE-PAGE	[123]	On request	
	SPIA	[125]	Registration	http://vortex.cs.wayne.edu/ontoexpress/
	THINK-Back-DS	[181]	Public	http://www.eecs.umich.edu/db/think/software.html
	TopoGSA	[181]	Public	http://www.topogsa.org

a) Availability under the assumption of academic, noncommercial use.

Laukens et. al. *Bioinformatics approaches for the functional interpretation of protein lists: From ontology term enrichment to network analysis*. **Proteomics**, 15, 981-996.

NET-GE

- **NET-GE**: NETwork-based Gene Enrichment.

  http://net-ge.biocomp.unibo.it/enrich

- **IDEA (Class B)**: extract **function-specific** modules from a molecular interaction network (STRING).

NET-GE

- **NET-GE**: NETwork-based Gene Enrichment.

  http://net-ge.biocomp.unibo.it/enrich

- **IDEA (Class B)**: extract **function-specific** modules from a molecular interaction network (STRING).
- **PRO**
  - Allows the detection of statistical associations not directly inferable from the annotations of the starting protein set.
  - The general approach does not rely on a specific underlying network.

NET-GE

- **NET-GE**: NETwork-based Gene Enrichment.

  http://net-ge.biocomp.unibo.it/enrich

- **IDEA (Class B)**: extract **function-specific** modules from a molecular interaction network (STRING).
- **PRO**
  - Allows the detection of statistical associations not directly inferable from the annotations of the starting protein set.
  - The general approach does not rely on a specific underlying network.
- **CONS**
  - Computationally hard pre-processing phase (it requires several weeks of calculation).
  - Gene Ontology is updated daily ...

Di Lena, Martelli, Fariselli, Casadio. *NET-GE: a novel NETwork-based Gene Enrichment for detecting biological processes associated to Mendelian diseases*. **BMC Genomics** 2015, 16: S6.

General workflow of the enrichment pipeline

1. **Standard enrichment**.
   - Bonferroni-corrected Fisher's exact test with respect to the selected annotation.
2. **Network-based enrichment**.
   - Bonferroni-corrected Fisher's exact test with respect to the functional modules built for the selected annotation.
3. **Output**.
   - Non-redundant ranking of overrepresented annotations obtained with standard and network-based enrichment.

Module extraction in NET-GE

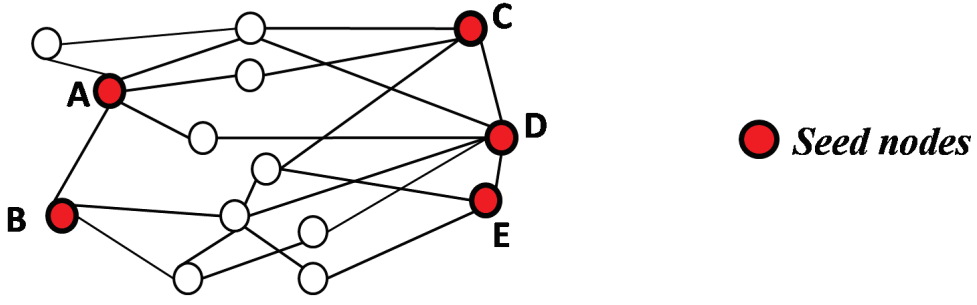
- The (computationally intensive) module extraction phase is aimed at extracting **functional sub-network** from a molecular interaction network.

Module extraction in NET-GE

- The (computationally intensive) module extraction phase is aimed at extracting **functional sub-network** from a molecular interaction network.
- It consists of several phases.
  1. Computation of the **shortest path network** between **seed** proteins.
  2. Ranking of the **connecting nodes** in the shortest path network.
  3. Extraction of the **minimum connecting network** from the shortest path network.
  4. Assessing the quality of the minimal connecting network.

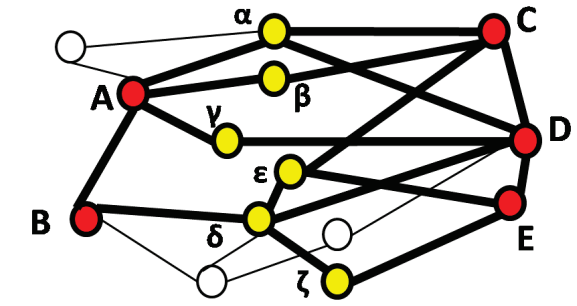
Module extraction in NET-GE

- The (computationally intensive) module extraction phase is aimed at extracting **functional sub-network** from a molecular interaction network.
- It consists of several phases.
  1. Computation of the **shortest path network** between **seed** proteins.
  2. Ranking of the **connecting nodes** in the shortest path network.
  3. Extraction of the **minimum connecting network** from the shortest path network.
  4. Assessing the quality of the minimal connecting network.
- First step: identify in the molecular interaction network all the proteins (seeds) annotated with the same functional-class label.



Module extraction

1. Extract the shortest paths among the seeds and collect the connecting nodes.



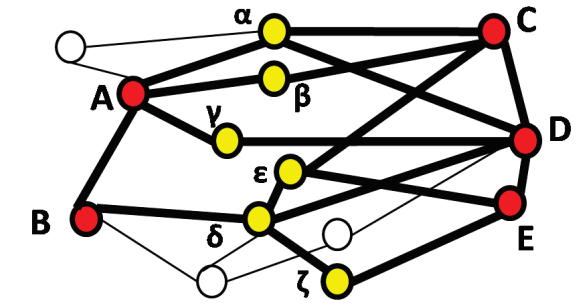
Shortest paths

A	0				
B	1	0			
C	2	3	0		
D	2	2	1	0	
E	3	3	2	1	0
	A	B	C	D	E

Connecting nodes *Links in shortest paths*

Module extraction

1. Extract the shortest paths among the seeds and collect the connecting nodes.



Shortest paths

A	0				
B	1	0			
C	2	3	0		
D	2	2	1	0	
E	3	3	2	1	0
	A	B	C	D	E

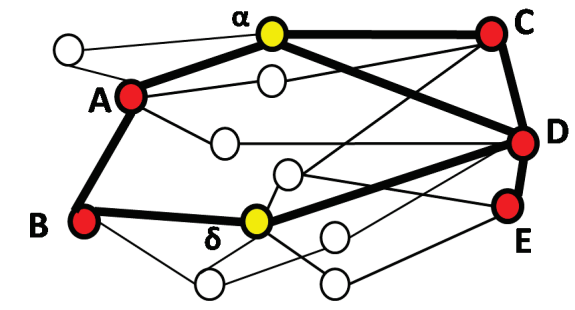
Connecting nodes *Links in shortest paths*

2. Rank the connecting nodes on the basis of: i) number of connected seed pairs (sc), ii) GO semantic similarity (ss), and iii) betweenness centrality (bc).

	sc	ss	bc
α	4	0.9	1.75
δ	3	0.9	2.5
ε	3	0.7	1.08
γ	2	0.8	1
β	2	0.8	0.75
ζ	1	0.7	0.33

Module extraction

3. Iteratively remove nodes with the lowest ranking, while preserving the shortest paths



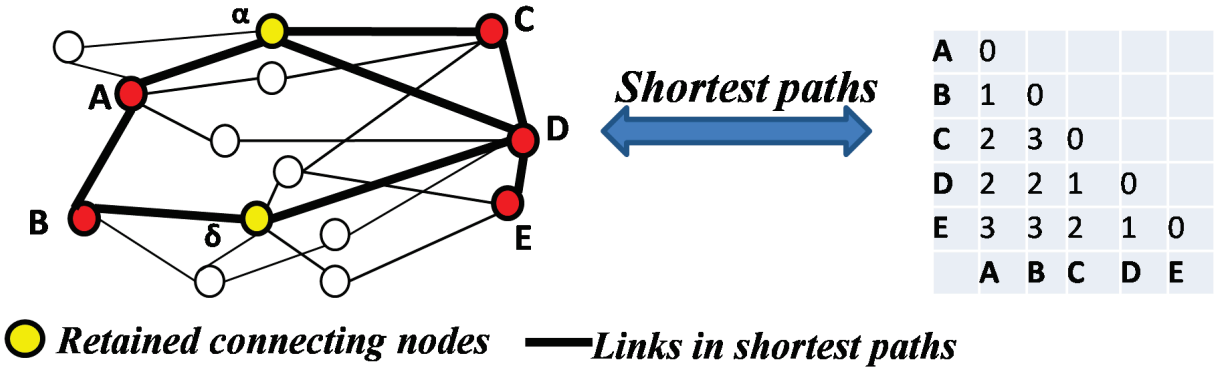
Shortest paths

A	0				
B	1	0			
C	2	3	0		
D	2	2	1	0	
E	3	3	2	1	0
	A	B	C	D	E

Retained connecting nodes *Links in shortest paths*

Module extraction

3. Iteratively remove nodes with the lowest ranking, while preserving the shortest paths



4. Quality filtering: the minimal network is retained if the average **semantic similarity** of the connecting nodes with respect to the reference GO term is significantly high.

Real example of module extraction: GO:0036018

- Consider the BP GO term `GO:0036018`, *cellular response to erythropoietin*.
- UniProtKB accession numbers of proteins annotated with GO:0036018:

  `P02795, P10721, P19235, P80297`

- Genes encoding such proteins:

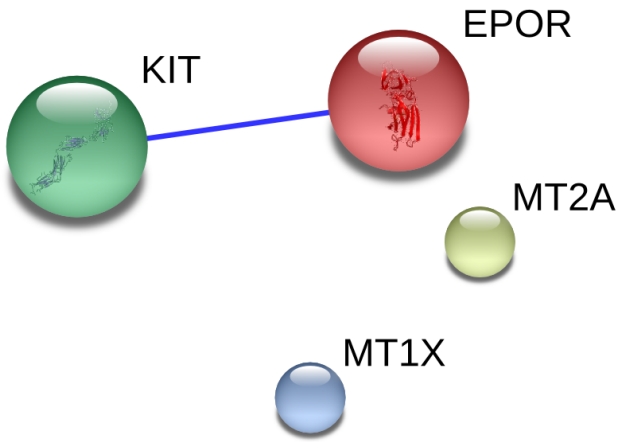
  `MT1X (P80297), MT2A (P02795), KIT (P10721), EPOR (P19235)`

Real example of module extraction: GO:0036018

- Consider the BP GO term `GO:0036018`, *cellular response to erythropoietin*.
- UniProtKB accession numbers of proteins annotated with GO:0036018:

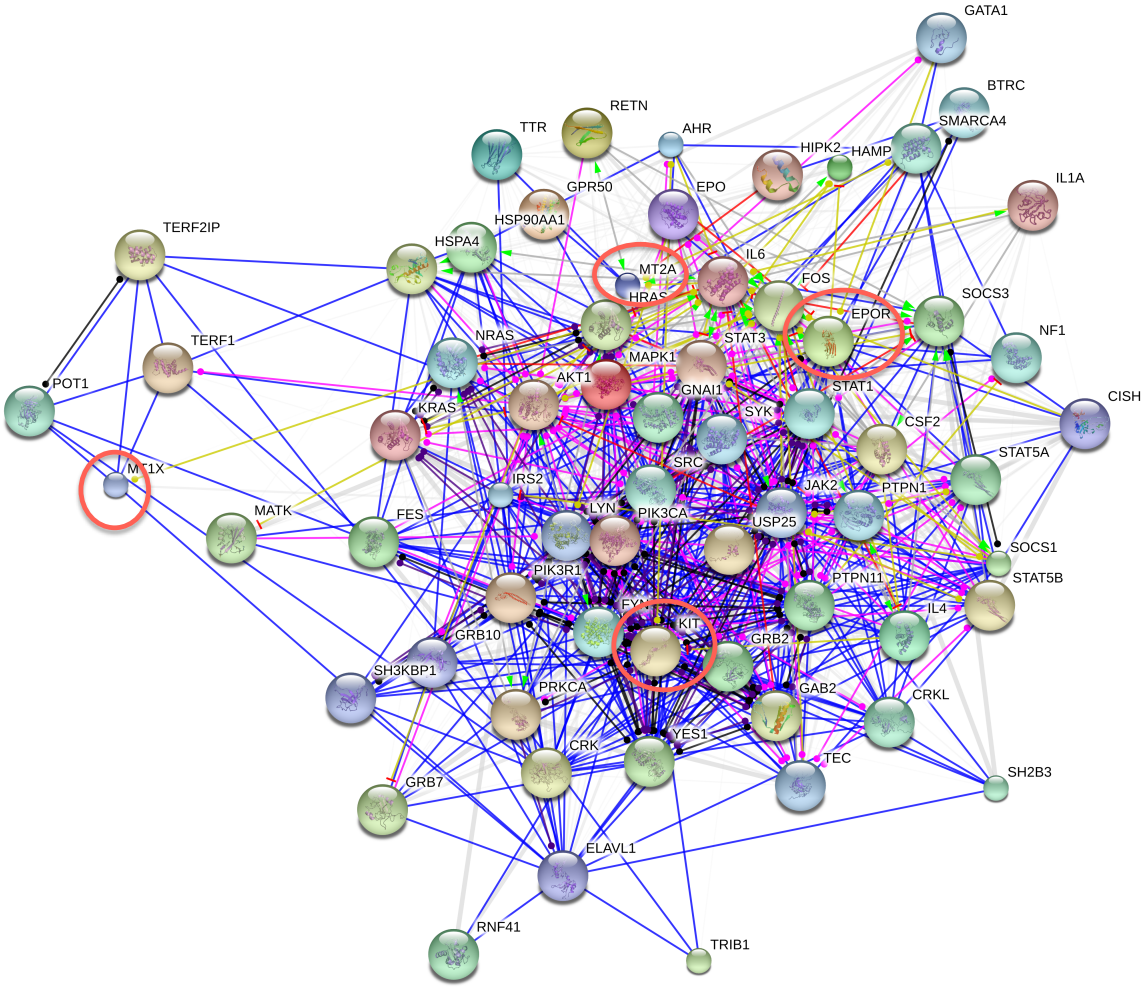
  `P02795, P10721, P19235, P80297`
- Genes encoding such proteins:

  `MT1X (P80297), MT2A (P02795), KIT (P10721), EPOR (P19235)`
- Action links in STRING between genes `MT1X, MT2A, KIT, EPOR`



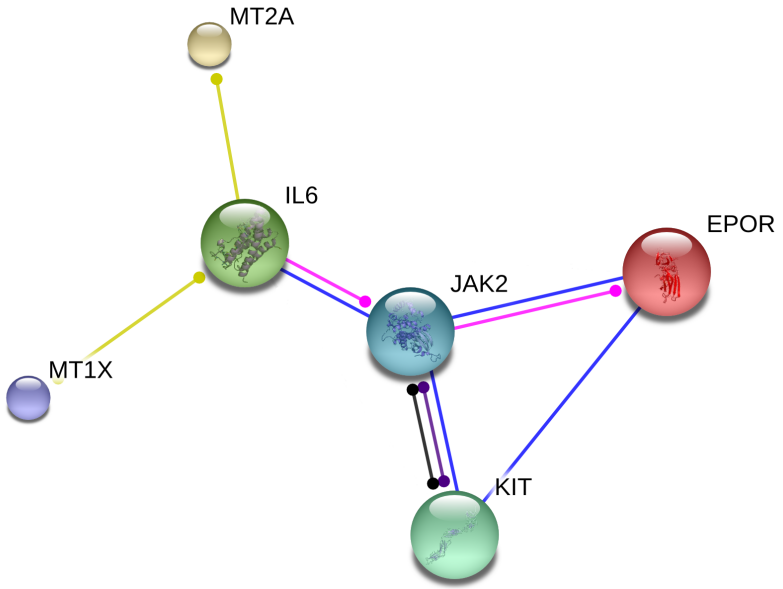
Real example of module extraction: GO:0036018

- Shortest paths network in STRING between genes `MT1X, MT2A, KIT, EPOR`



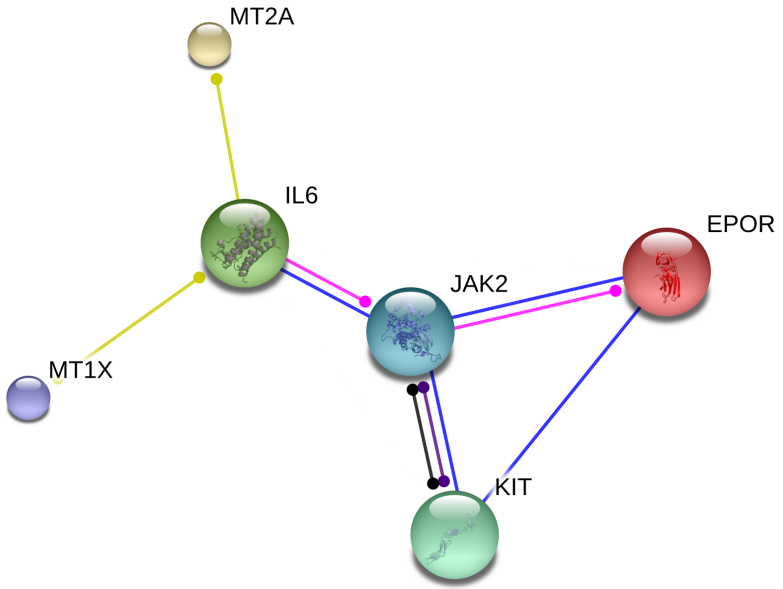
Real example of module extraction: GO:0036018

- Minimum connecting network between genes `MTX1, MT2A, KIT, EPOR`



Real example of module extraction: GO:0036018

- Minimum connecting network between genes `MTX1`, `MT2A`, `KIT`, `EPOR`



- Both *connecting* genes `JAK2` (`O60674`) and `IL6` (`P05231`) have annotation `GO:0019221` *cytokine-mediated signaling pathway*