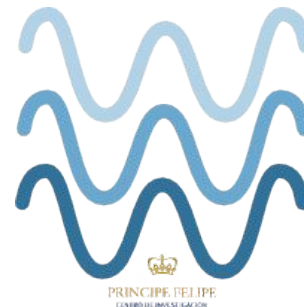


Functional Profiling

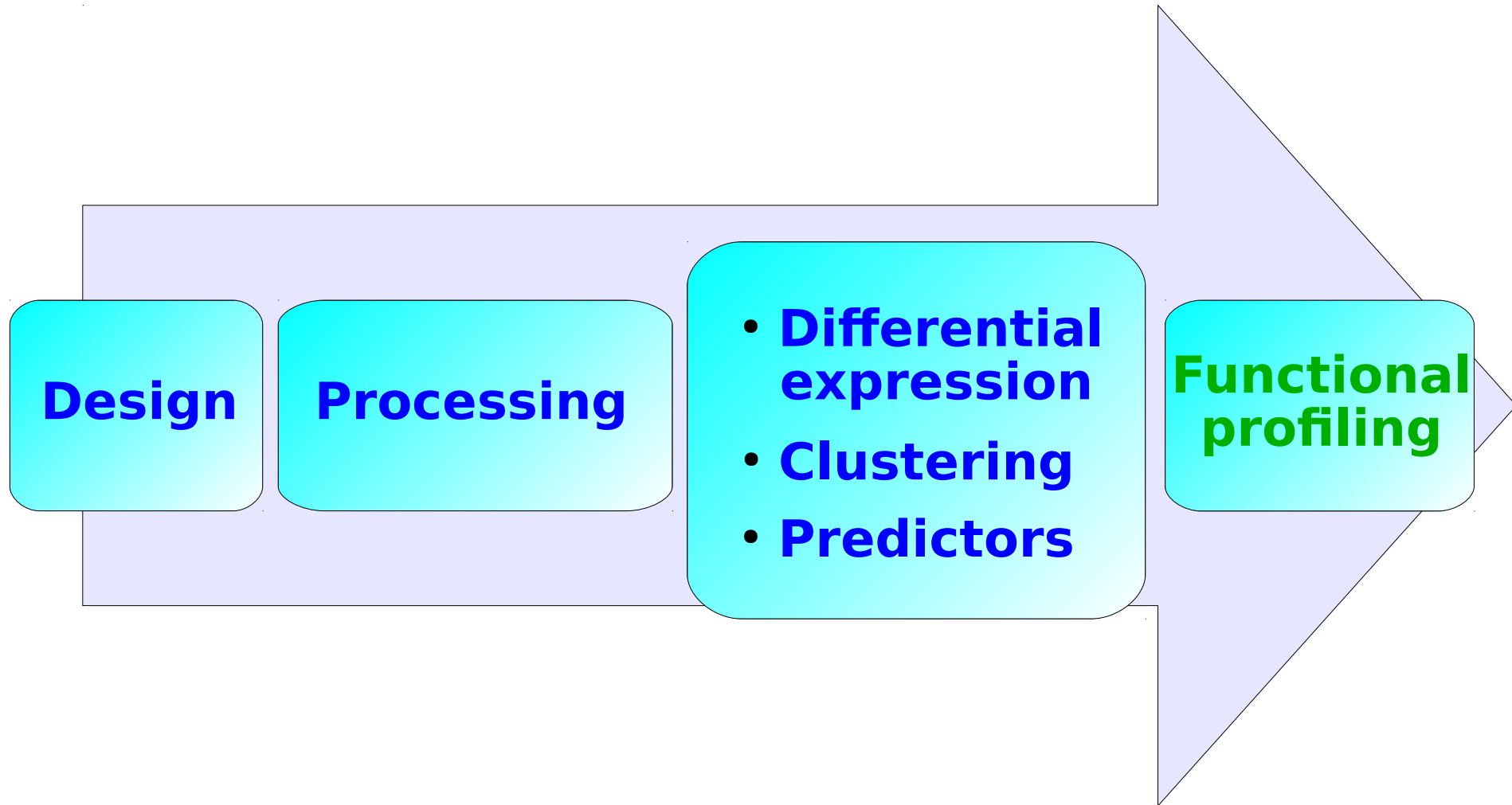
Francisco García García, fgarcia@cipf.es



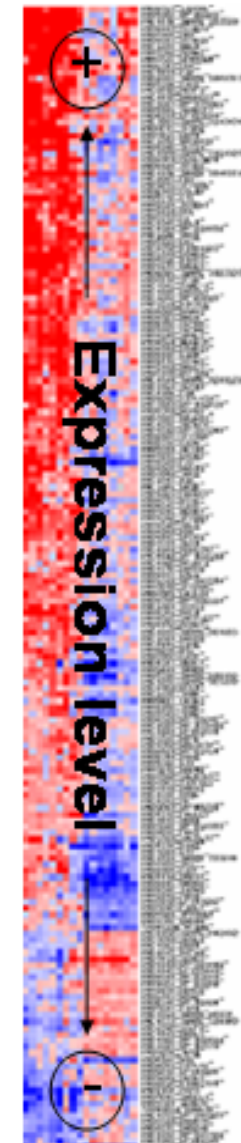
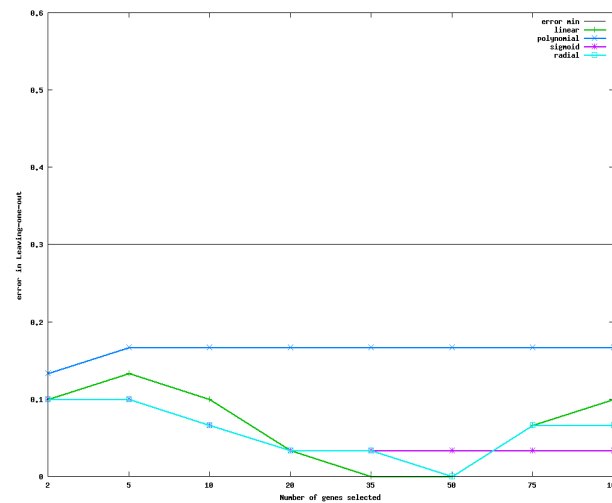
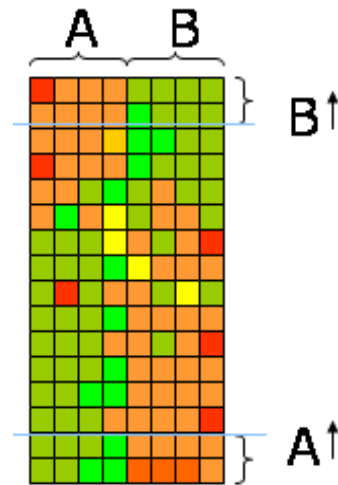
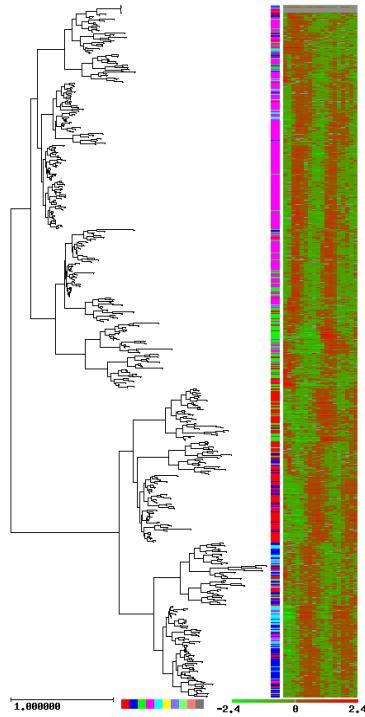
Outline

- Introduction
- Over-Representation Analysis (ORA)
- Gene Set Analysis (GSA)
- Network Analysis (NA)

Expression Data Analysis Pipeline



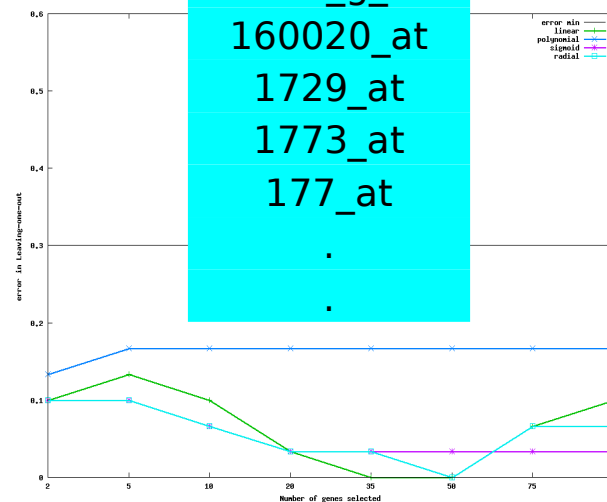
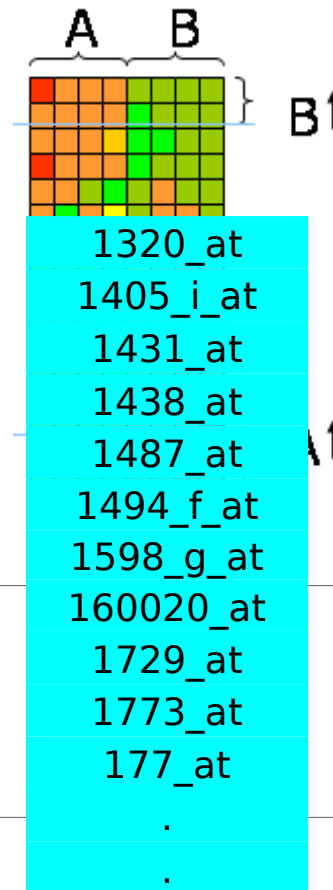
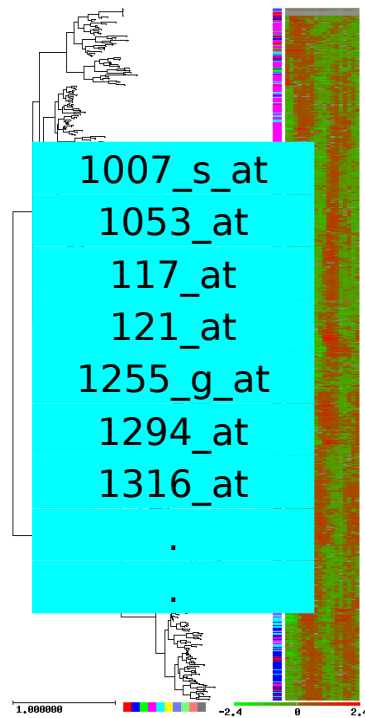
Genome-scale experiment output



Introduction

Functional Profiling

Genome-scale experiment output



Introduction

Functional Profiling

Questions we try to answer

- Is there any significant functional enrichment in my gene list / gene sets?
- Are these genes involved in common pathways?
- Do they share specific regulation?
- Are they involved in the same disease?

Functional databases



Homo sapiens



Mus musculus



Rattus norvegicus



Gallus gallus



Danio rerio



Drosophila melanogaster



C. elegans



Saccharomyces cerevisiae



Arabidopsis thaliana

UniProt/Swiss-Prot

UniProtKB/TrEMBL

Ensembl IDs

EntrezGene

Affymetrix

Agilent

**Genes
IDs**

HGNC symbol

EMBL acc

RefSeq

PDB

Protein Id

IPI....

Biological databases

**KEGG
pathways**

**Biocarta
pathways**

**Keywords
Swissprot**

**Gene
Ontology**

Biological Process
Molecular
Function Cellular
Component

**Gene
Expression
in tissues**

**Regulatory
elements**

MiRNA, CisRed

Transcription Factor
Binding Sites

**Bioentities from
literature:**

**Diseases terms
Chemical terms**

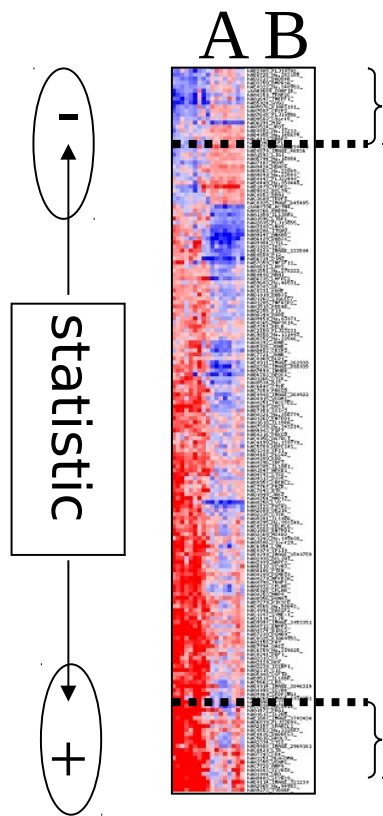
Introduction

Functional Profiling

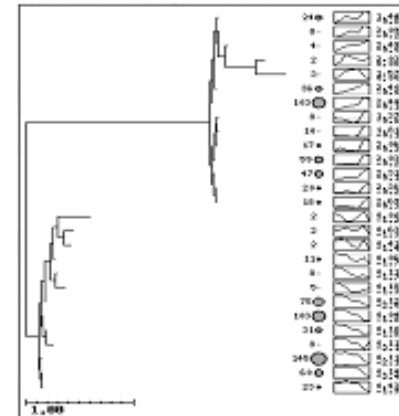
Outline

- Introduction
- Over-Representation Analysis (ORA)
- Gene Set Analysis (GSA)
- Network Analysis (NA)

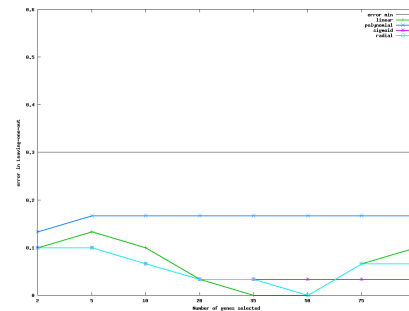
Over-Representation Analysis



1007_s_at
1053_at
117_at
121_at
1255_g_at
1294_at
1316_at
.
.



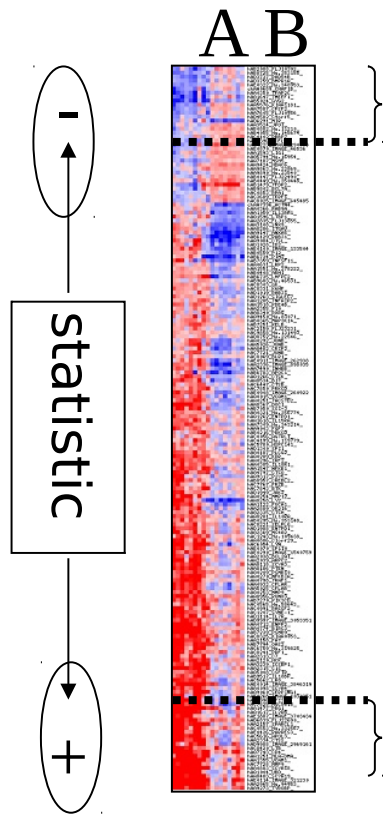
1320_at
1405_i_at
1431_at
1438_at
1487_at
1494_f_at
1598_g_at
160020_at
1729_at
1773_at
177_at
.
.



ORA

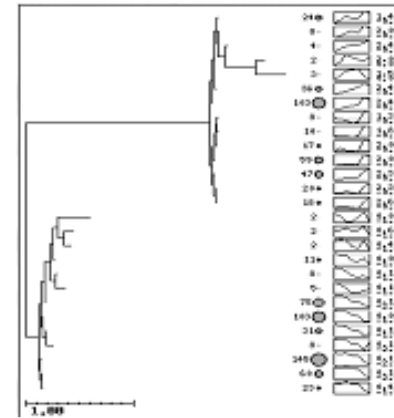
Functional Profiling

Over-Representation Analysis

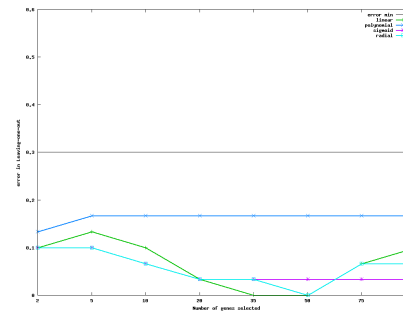


1007_s_at
1053_at
117_at
121_at
1255_g_at
1294_at
1316_at
.
.

Function
4/7



1320_at
1405_i_at
1431_at
1438_at
1487_at
1494_f_at
1598_g_at
160020_at
1729_at
1773_at
177_at
.
.

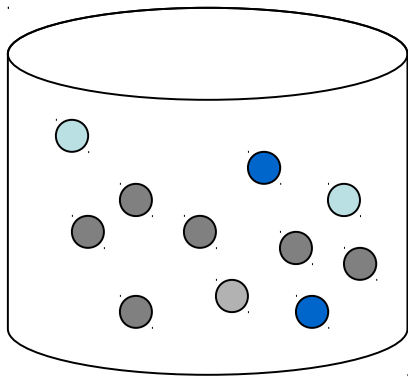


Function
2/11

Over-Representation Analysis

FatiGO test

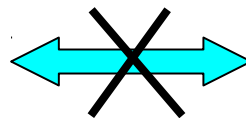
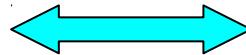
One Gene List (A)



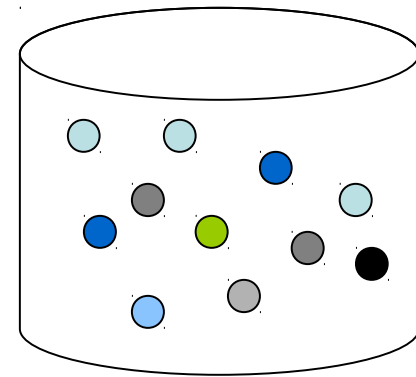
Biosynthesis 60% ●

Sporulation 20% ●

Are these two
groups of genes
carrying out
different
biological roles?



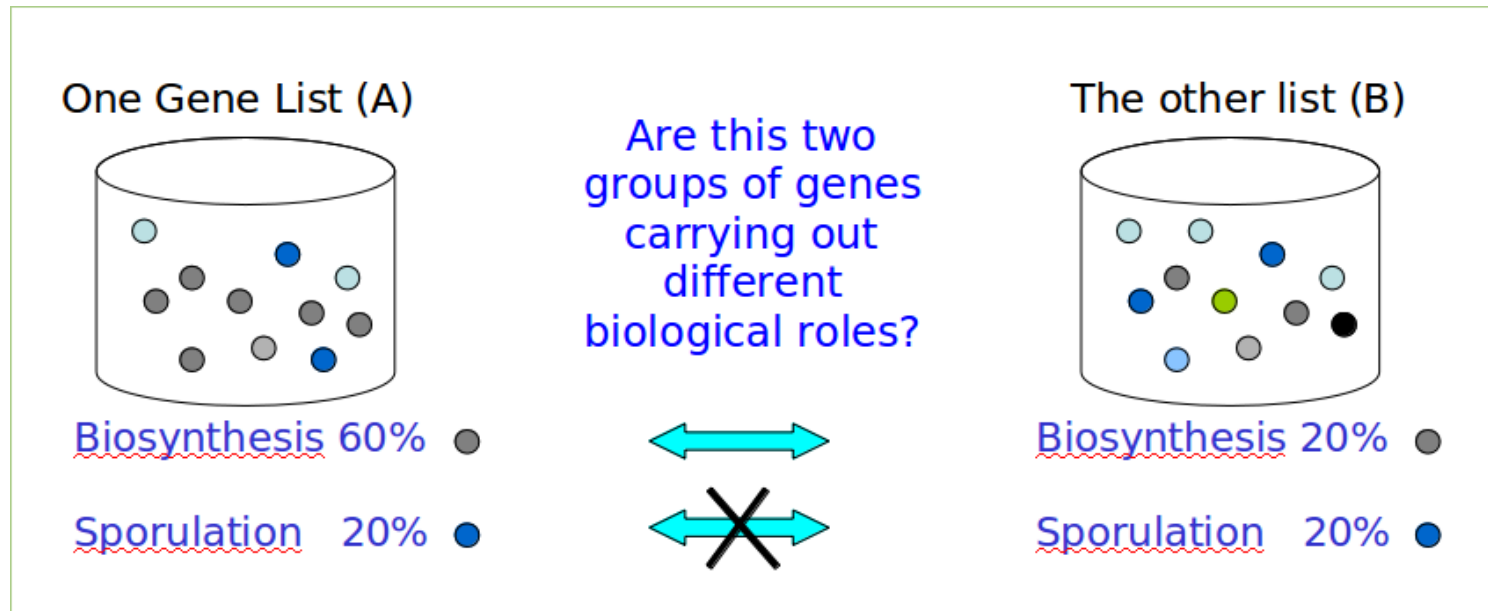
The other list (B)



Biosynthesis 20% ●

Sporulation 20% ●

Over-Representation Analysis



Genes in group A have significantly to do with biosynthesis, but not with sporulation.

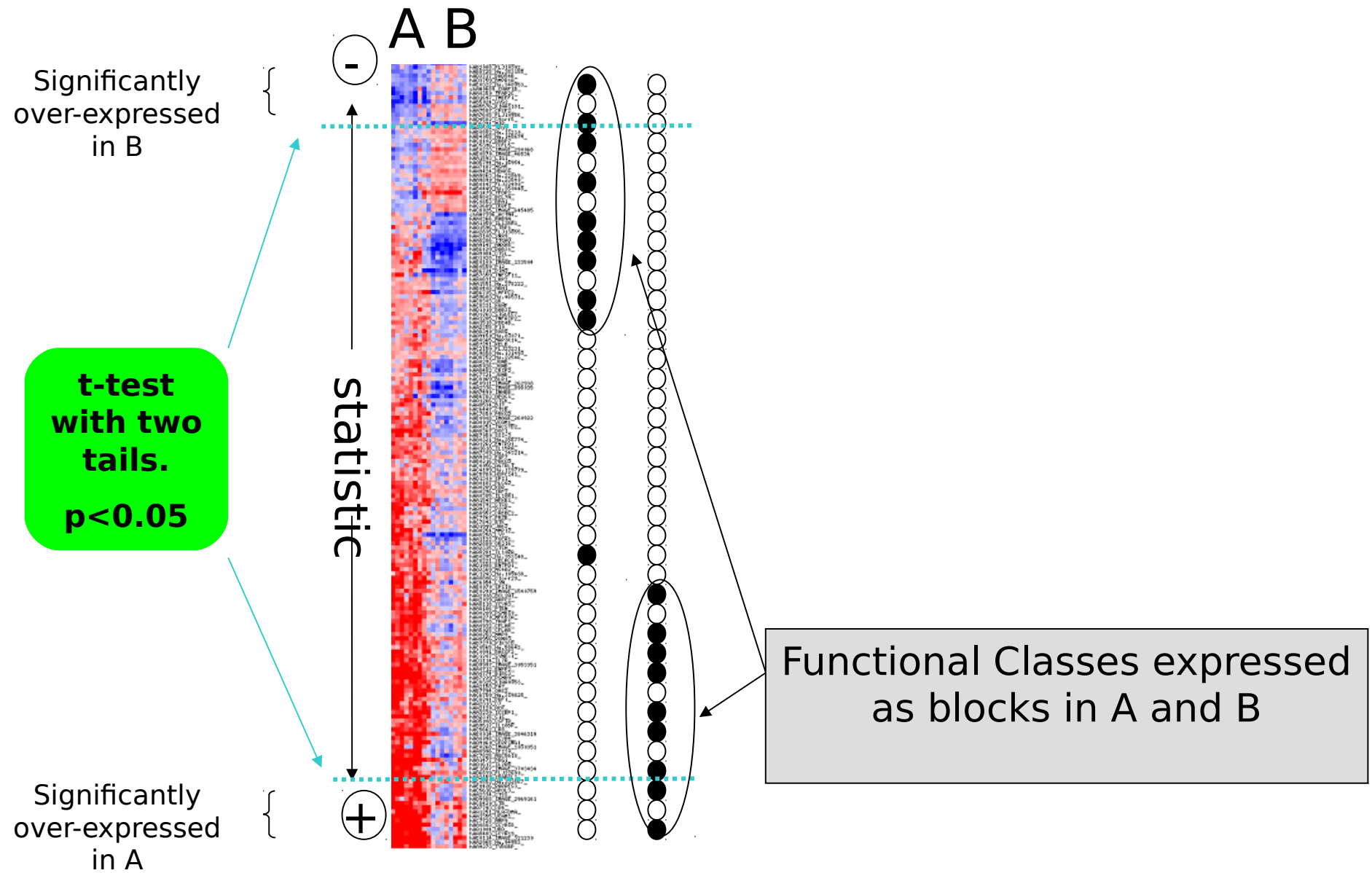
	A	B
Biosynthesis	6	2
No biosynthesis	4	8

We do this for each term (GO, miRNA, Interpro , ...)
Thousand of terms, so Multiple Test Correction is needed!!!

Outline

- Introduction
- Over-Representation Analysis (ORA)
- **Gene Set Analysis (GSA)**
- Network Analysis (NA)

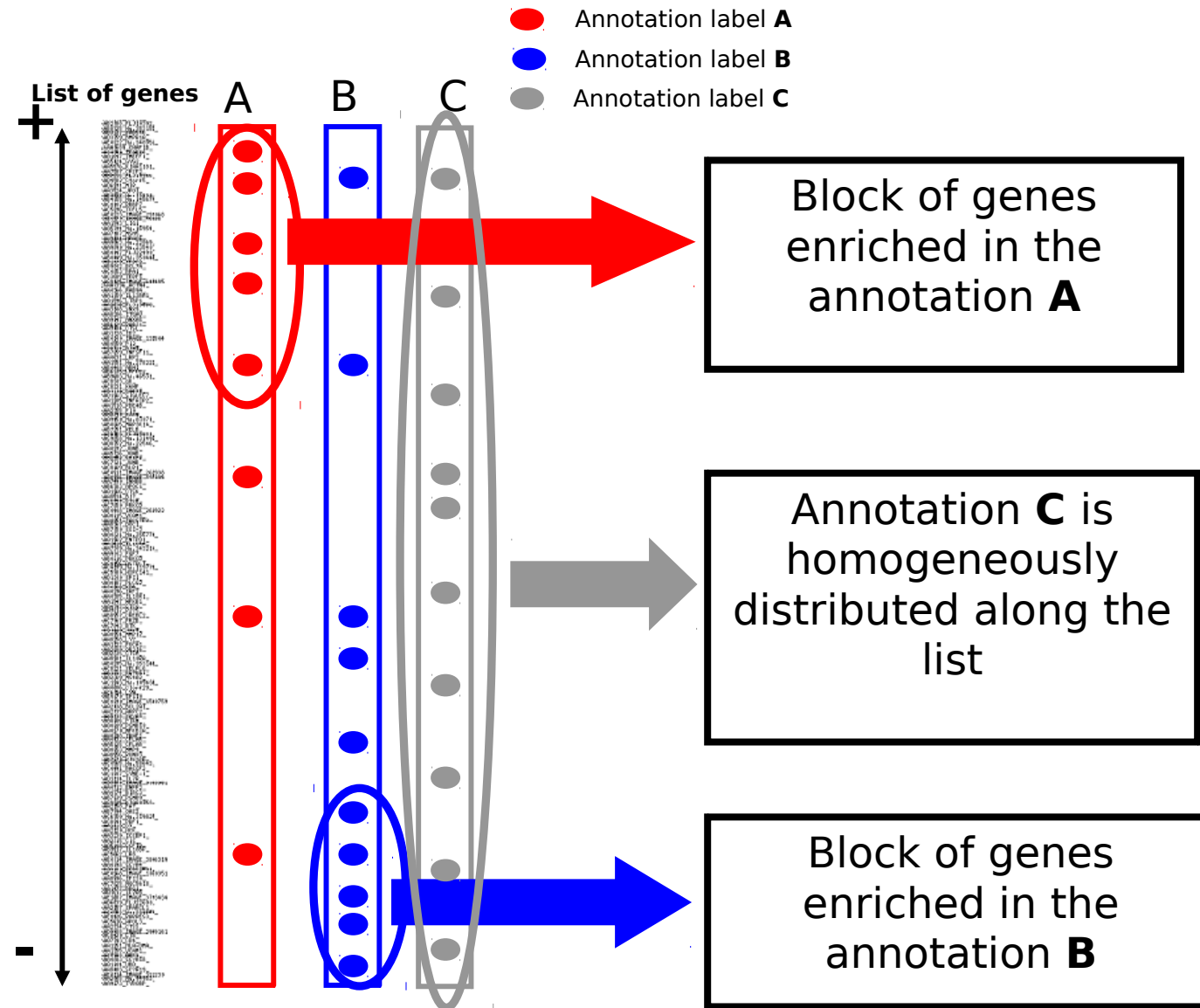
Gene Set Analysis



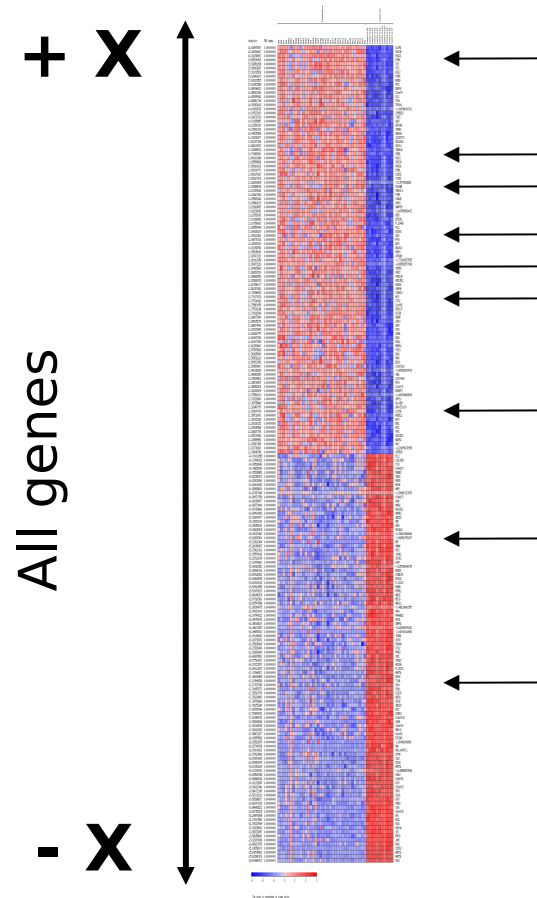
GSA

Functional Profiling

Gene Set Analysis



Gene Set Analysis



$$\ln \left(\frac{P(g \in F)}{P(g \notin F)} \right) = K + \alpha X$$

alpha > 0 : increasing X increases the probability of the gene to be annotated

alpha < 0 : decreasing X increases the probability of the gene to be annotated

Outline

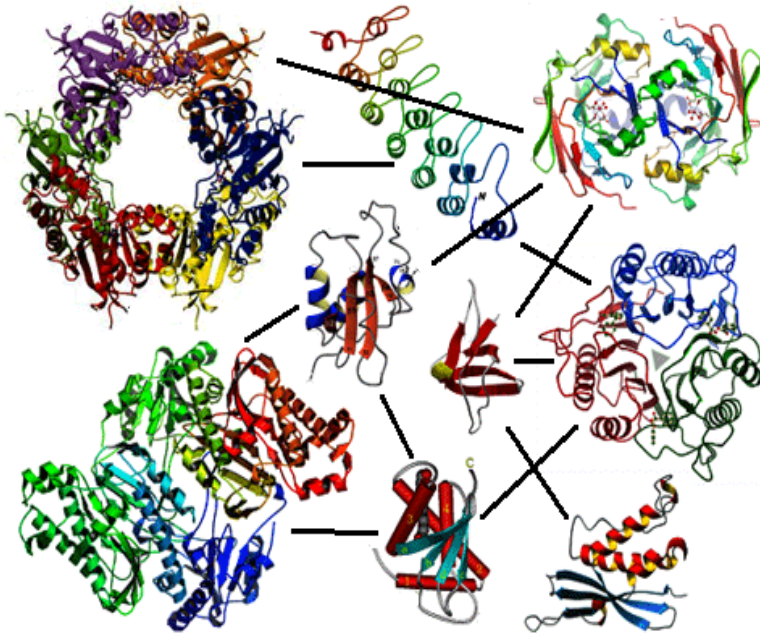
- Introduction
- Over-Representation Analysis (ORA)
- Gene Set Analysis (GSA)
- Network Analysis (NA)

Protein-Protein Interactions (PPI)

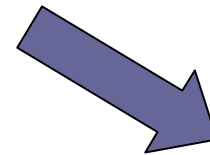
- How to extract information about **sets** of genes?
- How to perform **functional enrichment analysis** using protein-protein interactions as annotation source?
- How to **prioritize candidate genes**?
- How to get **new functional candidate genes**?

Graph Theory

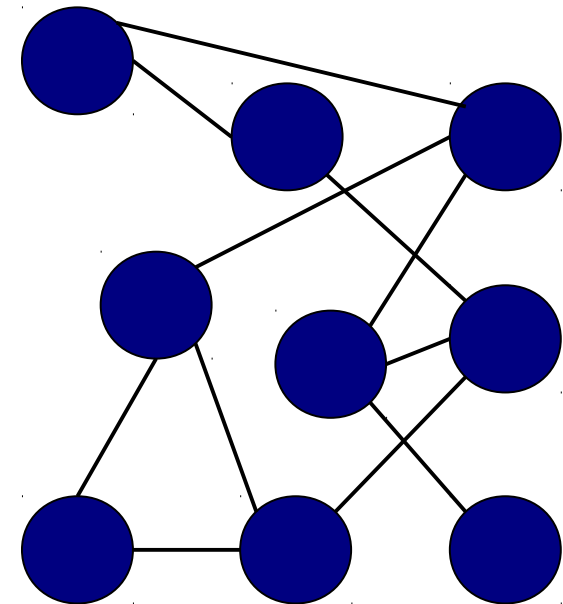
Set of proteins interacting



Nodes = proteins
Edges = interaction events



Undirected graph



structured data

Graph Theory

Graph theory may help us to study protein networks.
Some interesting parameters:

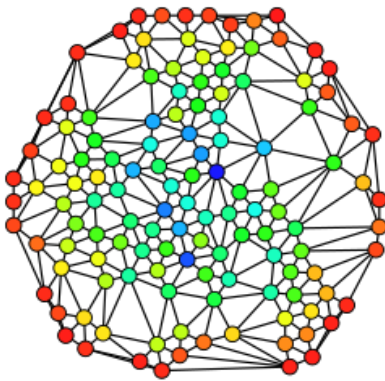
- **Degree (connectivity or connections)**: number of edges connected to a node. Nodes with high degree are called **hubs**.

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

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

σ_{st} is total number of shortest paths in the graph.

$\sigma_{st}(V)$ is the number of shortest paths that pass through node V

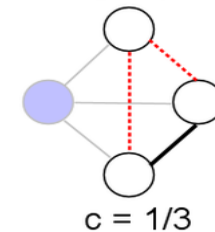


Graph Theory

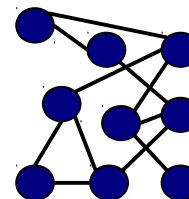
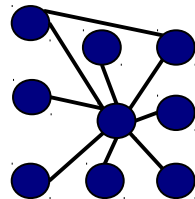
- **Clustering coefficient (of a node)**: A measure of how interconnected the neighbours of that node are. Proportion of links between the nodes within its neighbourhood divided by the number of links that could possibly exist between them.

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

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:

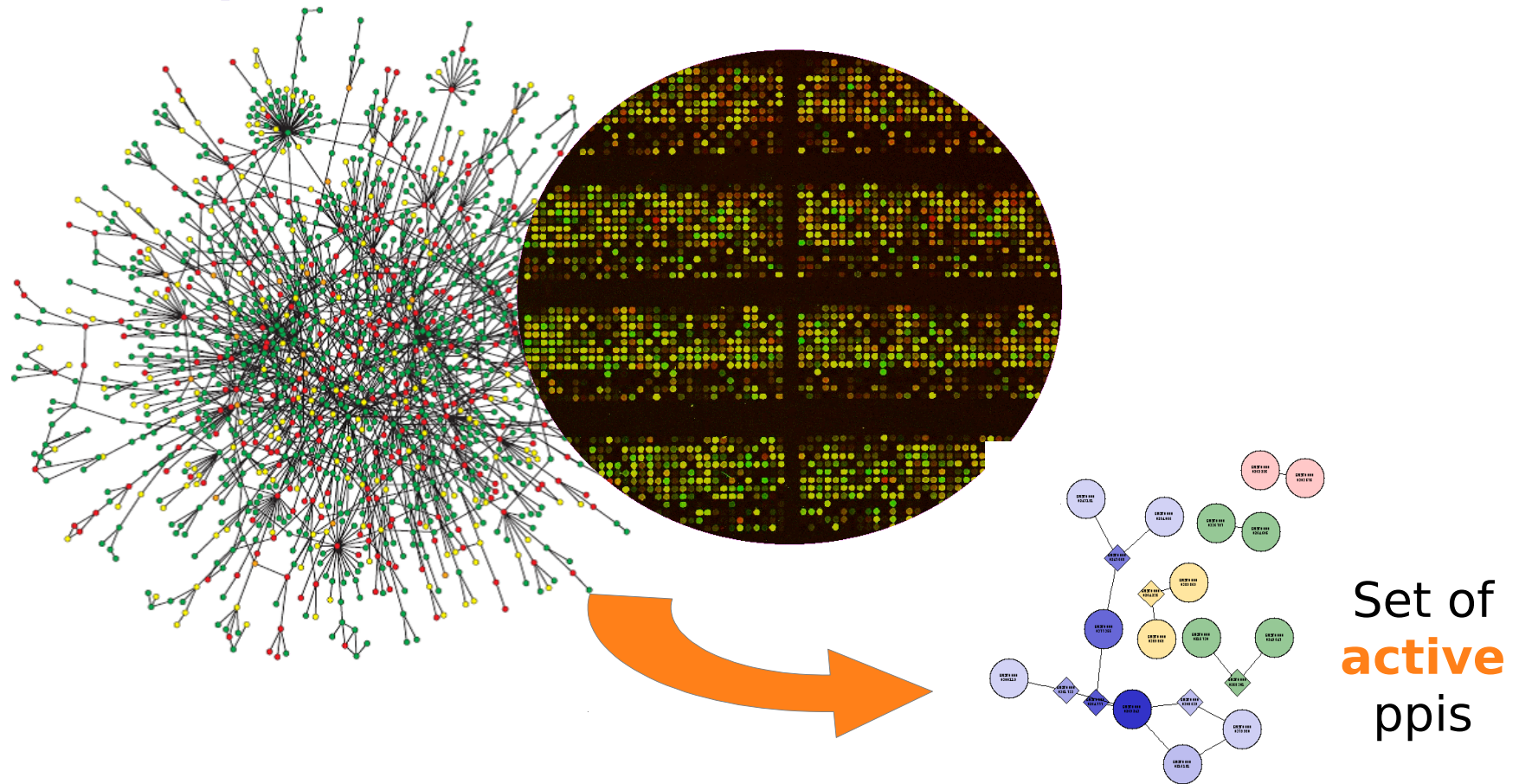
Shortest path. The path with less edges that connects two nodes.

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**.

Interactome & Transcriptome

- **Interactome.** Complete collection of protein-protein interactions in the cell.
- **Transcriptome** determines the real interactome.



Interactome & Transcriptome

Goal

To develop a methodology that may **extract from lists of proteins/genes** the ppi networks acting and evaluates whether they have importance in the **cooperative behaviour** of the list.

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**