Using Graph Theory to Analyze Gene Network Coherence

Francisco A. Gómez-Vela
fgomez@upo.es

Norberto Díaz-Díaz
ndiaz@upo.es

José A. Lagares

José A. Sánchez

Jesús S. Aguilar

BIGS BioInformatics Group Seville

Outlines

- Introduction
- Proposed Methodology
- Experiments
- Conclusions

Outlines

- **Introduction**
- Proposed Methodology
- Experiments
- Conclusions

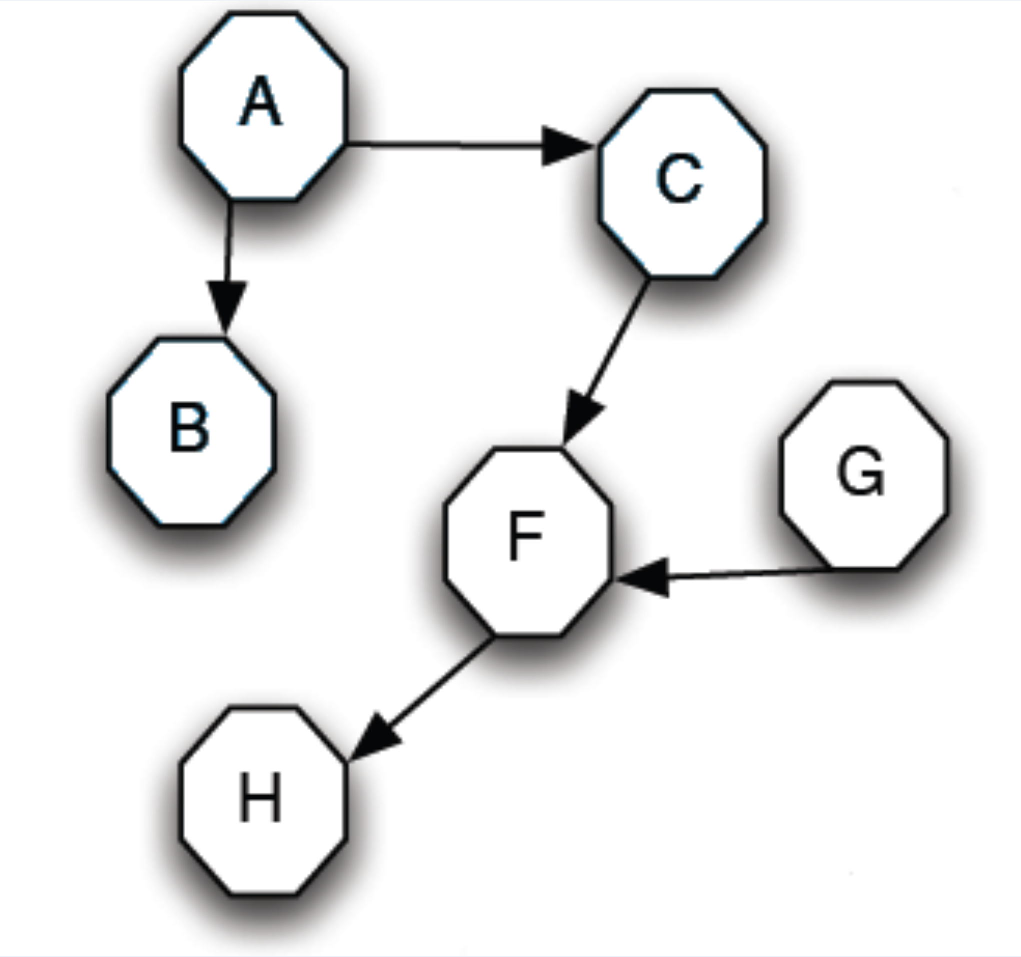
Introduction

Gene Network

- There is a need to generate patterns of expression, and behavioral influences between genes from microarray.
- GNs arise as a visual and intuitive solution for gene-gene interaction.
- They are presented as a graph:
    - Nodes: are made up of genes.
    - Edges: relationships among these genes.

Introduction

Gene Network



Introduction
Gene Network

- Many GN inference algorithms have been developed as techniques for extracting biological knowledge
  - Ponzoni et al., 2007.
  - Gallo et al., 2011.

- They can be broadly classified as (Hecker M, 2009):
  - Boolean Network
  - Information Theory Model
  - Bayesian Networks

Introduction

Gene Network Validation in Bioinformatics

- Once the network has been generated, it is very important to assure network reliability in order to illustrate the quality of the generated model.

  - **Synthetic data based validation**
    - This approach is normally used to validate new *methodologies or algorithms*.
  - **Well-Known data based validation**
    - The literature prior knowledge is used to validate *gene networks*.

Introduction

Well-Known Biological data based Validation

- The quality of a GN can be measured by a direct comparison between the obtained GN and prior biological knowledge (Wei and Li, 2007; Zhou and Wong, 2011).

- However, these approaches are not entirely accurate as they only take direct gene–gene interactions into account for the validation task, leaving aside the weak (indirect) relationships (Poyatos, 2011).

Outlines

- Introduction
- **Proposed Methodology**
- Experiments
- Conclusions

Proposed Methodology

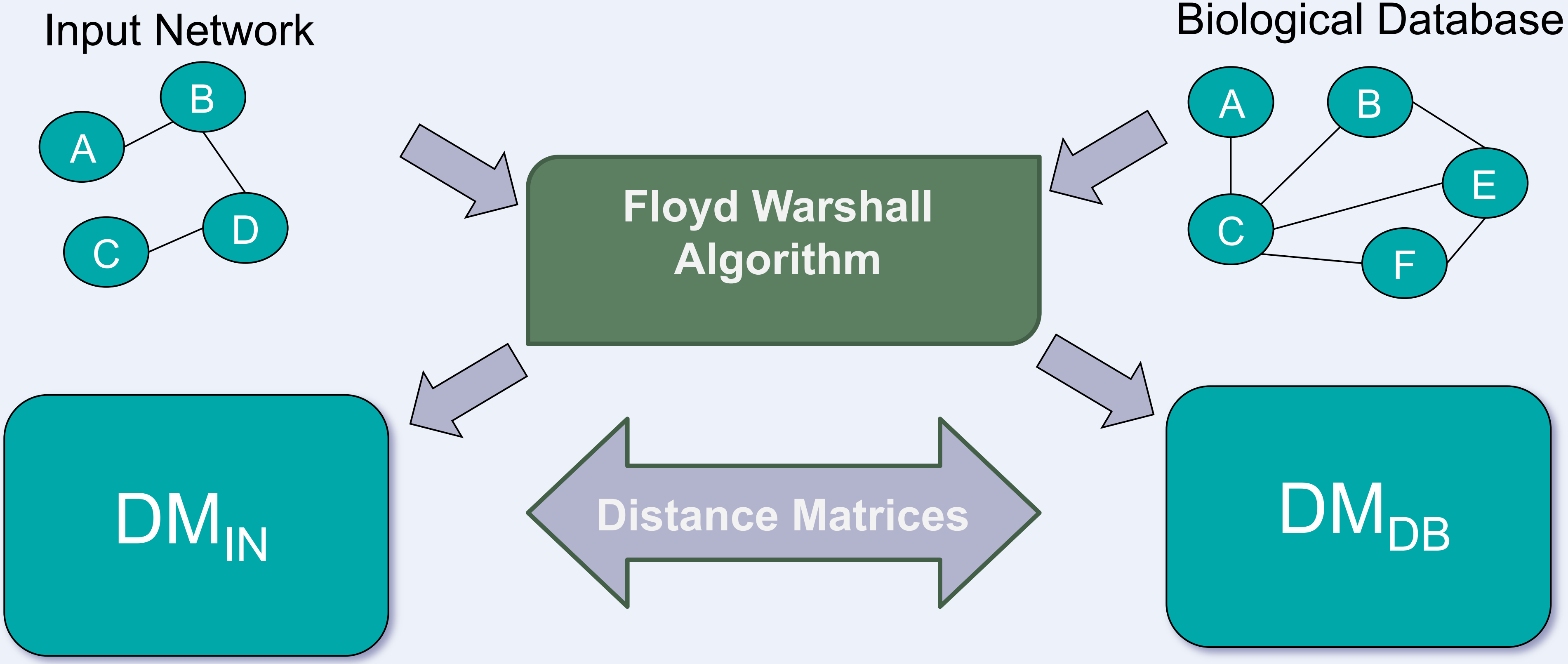
- The main features of our method:

  - Evaluate the similarities and differences between gene networks and biological database.

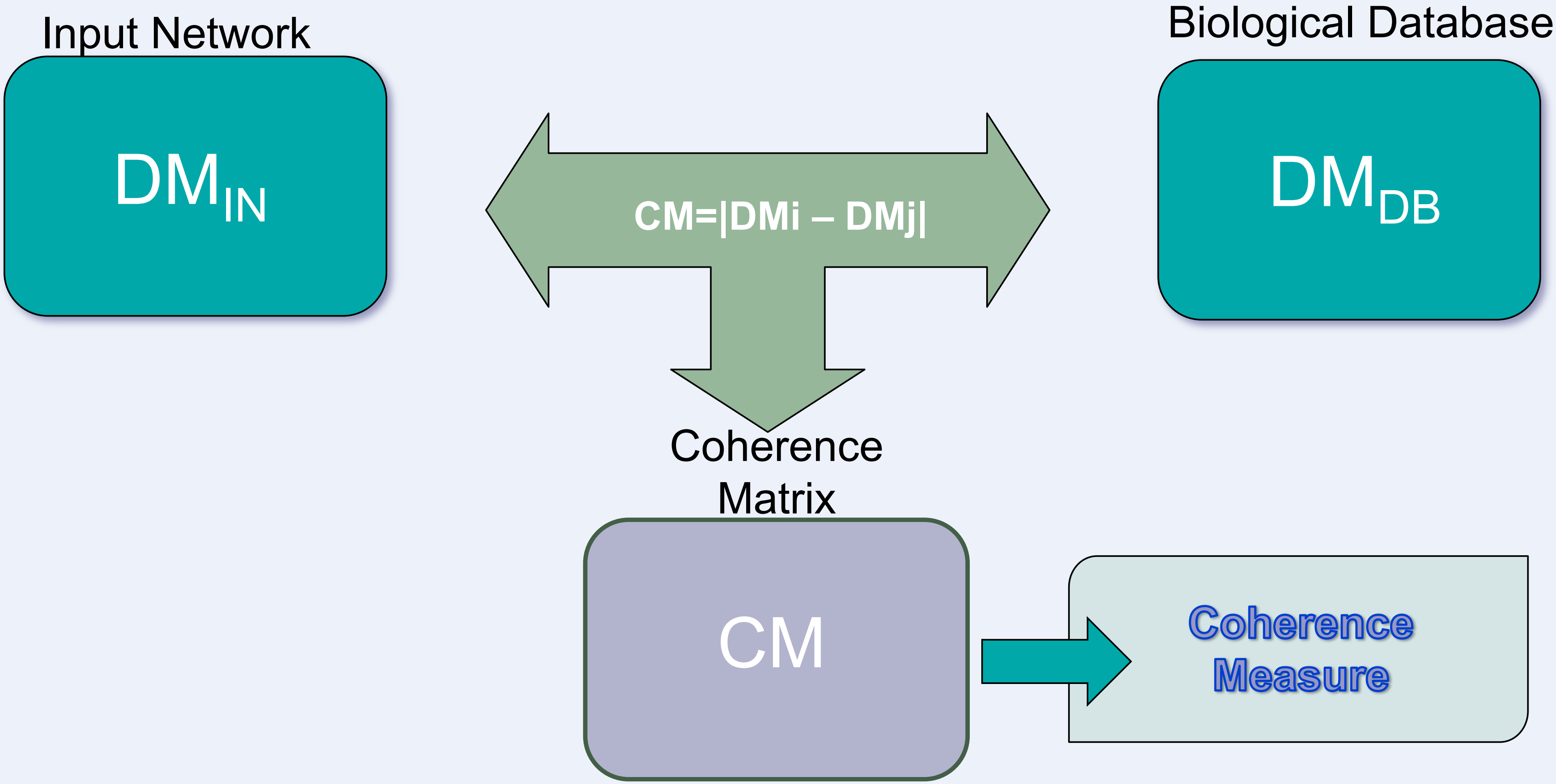
  - Take into account the indirect gene-gene relationships for the validation process.

  - Using Graph Theory to evaluate with gene networks and obtain different measures.

Proposed Methodology



Proposed Methodology



Input Network

DM_{IN}

Biological Database

DM_{DB}

CM=|DMi – DMj|

Coherence Matrix

CM

Coherence Measure

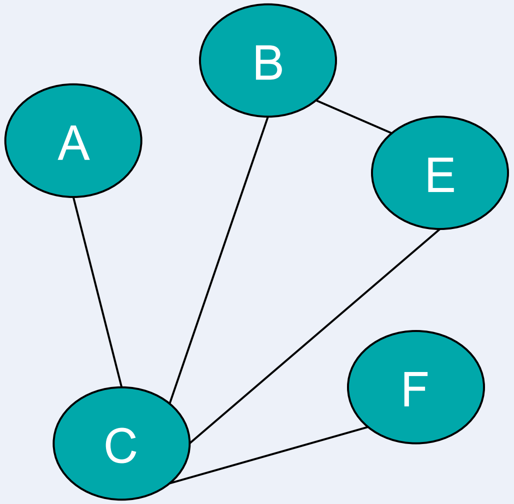
$CM = |DM_{IN} - DM_{DB}|$

Proposed Methodology

Floyd-Warshall Algorithm

- This approach is a graph analysis method that solves the shortest path between nodes.

Network



Distance Matrix

	A	B	C	E	F
A	0	2	1	1	2
B	2	0	1	1	2
C	1	1	0	2	1
E	1	1	2	0	1
F	2	2	1	1	0

Proposed Methodology
Distance Threshold

- Distance threshold (δ)

  - It is used to exclude relationships that lack biological meaning.

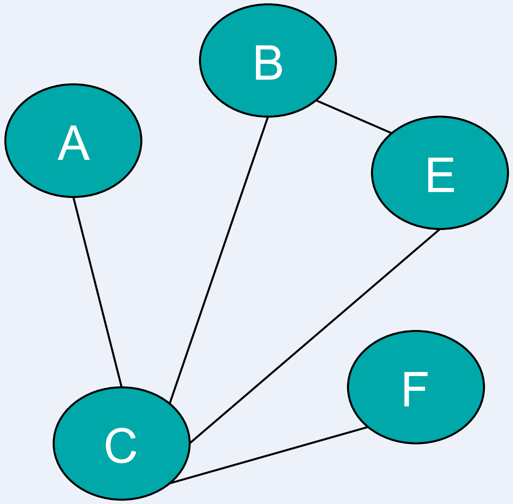
  - This threshold denotes the maximum distance to be considered as relevant in the Distance Matrix generation process.

  - If the minimum distance between two genes is greater than δ, then no path between the genes will be assumed.

Proposed Methodology

Distance Threshold

Network



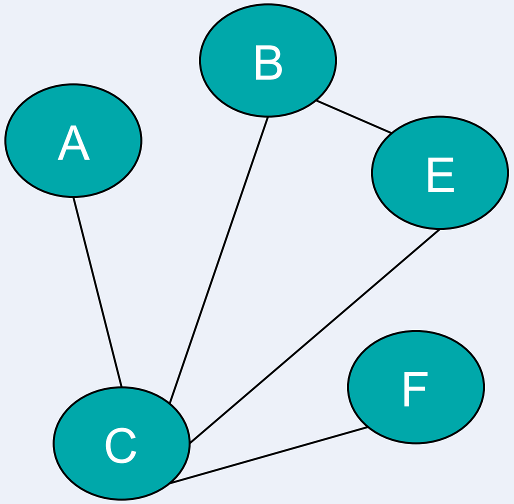
$\delta = 1$

Distance Matrix

	A	B	C	E	F
A	0	~~2~~	1	1	~~2~~
B	~~2~~	0	1	1	~~2~~
C	1	1	0	~~2~~	1
E	1	1	~~2~~	0	1
F	~~2~~	~~2~~	1	1	0

Proposed Methodology
Distance Threshold

Network



$\delta = 1$

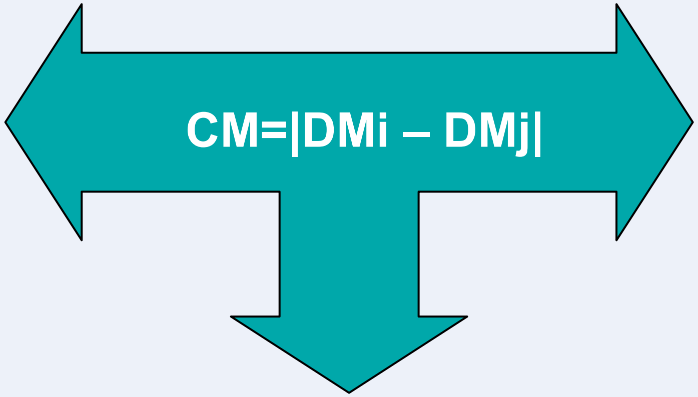
Distance Matrix

	A	B	C	E	F
A	0	∞	1	1	∞
B	∞	0	1	1	∞
C	1	1	0	∞	1
E	1	1	∞	0	1
F	∞	∞	1	1	0

Proposed Methodology

DM_{IN}

	A	B	C	D
A	0	1	∞	2
B	1	0	2	1
C	∞	2	0	1
D	2	1	1	0



CM=|DMi – DMj|

DM_{DB}

	A	B	C	E	F
A	0	2	1	2	2
B	2	0	1	1	2
C	1	1	0	1	1
E	2	1	1	0	1
F	2	2	1	1	0

Coherence Matrix (CM)

	A	B	C
A	0	1	∞
B	1	0	1
C	∞	1	0

Proposed Methodology
Obtaining Measures

- Coherence Level threshold (θ)
  - This threshold denotes the maximum coherence level to be considered as relevant in the Coherence Matrix.
  - It is used to obtain well-Known indices by using the elements of the coherence matrix:

$$CM_{i,j}\begin{cases} |v-y|<=\theta \longrightarrow \mathrm{TP} \\ |v-y|>\theta \longrightarrow \mathrm{FP} \\ |\infty-y|\ (\alpha) \longrightarrow \mathrm{FN} \\ |\infty-\infty|(\beta) \longrightarrow \mathrm{TN} \end{cases} \qquad 0< v,y \ <\infty$$

Proposed Methodology

$$\theta = 3$$

Coherence Matrix

	A	B	C	D	E
A	-	1	α	4	7
B	1	-	β	2	5
C	α	β	-	1	8
D	4	2	1	-	1
E	7	5	8	1	-

Proposed Methodology

θ = 3

Coherence Matrix

	A	B	C	D	E
A	-	TP	FN	FP	FP
B	TP	-	TN	TP	FP
C	FN	TN	-	TP	FP
D	FP	TP	TP	-	TP
E	FP	FP	FP	TP	-

Outlines

- Introduction
- Proposed Methodology
- **Experiments**
- Conclusions

Results

Real data experiment

- Input networks were obtained by applying four inference network techniques on the well-known yeast cell cycle expression data set (Spellman et al., 1998).
  - Soinov et al., 2003.
  - Bulashevska et al., 2005.
  - Ponzoni (GRNCOP) et al., 2007
- Comparison with Well-Known data:
  - BioGrid
  - KEGG
  - SGD
  - YeastNet

Results

Real data experiment

- Several studies were carried out using different threshold value combinations:
    - Distance threshold (δ) and Coherence level threshold (θ) have been modified from one to five, generating 25 different combinations.
- The results show that the higher δ and θ values, the greater is the noise introduced.
    - The most representative result, was obtained for δ=4 and θ=1.

Results

	Soinov		Bulashevska		Ponzoni	
	Accuracy	F-measure	Accuracy	F-measure	Accuracy	F-measure
Biogrid	0,27	0,42	0,65	0,79	**0,82**	**0,90**
KEGG	**0,58**	0,48	0,34	**0,50**	0,28	0,43
SGD	0,31	0,47	0,53	0,69	**1**	**1**
YeastNet	0,29	0,45	0,50	0,66	**1**	**1**

Results

- These results are consistent with the experiment carried out in Ponzoni et al., 2007.
- Ponzoni was successfully compared with Soinov and Bulashevska approaches.

Outlines

- Introduction
- Proposed Methodology
- Experiments
- **Conclusions**

Conclusions

- A new approach of a gene network validation framework is presented:
    - The methodology not only takes into account the direct relationships, but also the indirect ones.
    - Graph theory has been used to perform validation task.

Conclusions

- Experiments with Real Data.
  - These results are consistent with the experiment carried out in Ponzoni et al., 2007.
  - Ponzoni was successfully compared with Soinov and Bulashevska approaches.
  - These behaviours are also found in the obtained results. Ponzoni presents better coherence values than Soinov and Bulashevska in BioGrid, SGD and YeastNet.

Future Works

- The methodology has been improved:
  - The elements in coherence matrix will be weighted based on the gene-gene relationships distance.
  - A new measure, based on different databases will be generated.
- Moreover, a Cytoscape plugin will be implemented.

Some References

Pavlopoulos GA, et al. (2011): **Using graph theory to analyze biological networks.** *BioData Mining*, **4:**10.

Asghar A, et al (2012) **Speeding up the Floyd–Warshall algorithm for the cycled shortest path problem**. AppliedMathematics Letters 25(1): 1

Bulashevska S and Eils R (2005) **Inferring genetic regulatory logic from expression data.** Bioinformatics 21(11):2706.

Ponzoni I, et al (2007) **Inferring adaptive regulationthresh-olds and association rules from gene expressiondata through combinatorial optimization learning**.IEEE/ACM Transaction on Computation Biology andBioinformatics 4(4):624.

Poyatos JF (2011). **The balance of weak and strong interactions in genetic networks**. PloS One 6(2):e14598._

UNIVERSIDAD PABLO DE OLAVIDE SEVILLA

BioInformatics Research Group

Using Graph Theory to Analyze Gene Network Coherence

Thanks for your attention

NETTAB