Gene interaction networks inference and search for complex disease biomarkers by complex networks analysis and data integration

David Correa Martins Jr
david.martins@ufabc.edu.br

Center of Mathematics, Computing and Cognition
Federal University of ABC
Santo André/SP - Brazil

Universidade Federal do ABC

Summary

Universidade Federal do ABC

Summary

Universidade Federal do ABC

Summary

Universidade Federal do ABC

Summary

Universidade Federal do ABC

About UFABC

- Universidade Federal do ABC (Federal University of ABC - UFABC)
- 9 years old university
- Growing fast! ($\sim$ 12,000 undergrad students, $\sim$ 1,500 grad students, $\sim$ 550 professors)
- Interdisciplinarity as key to perform relevant science
- Research quality 60% superior to the world average in terms of impact factor (# 1 in Brazil)
- Strong internationalization (# 1 in Brazil)

Universidade Federal do ABC

Systems Biology

- Systems Biology: interdisciplinary field which studies the complex networks of interactions occurring in biological systems
- Development of models and approaches to reveal emergent properties of cells, tissues and organs, which work as an integrated system
- Tipically involves studies of several types of biological networks (gene regulation, metabolic, protein interactions, cell signaling, etc...)
- Integration and analysis of massive, complex and heterogeneous datasets (Big Data)

Systems Biology

- Systems Biology: interdisciplinary field which studies the complex networks of interactions occurring in biological systems
- Development of models and approaches to reveal emergent properties of cells, tissues and organs, which work as an integrated system
- Tipically involves studies of several types of biological networks (gene regulation, metabolic, protein interactions, cell signaling, etc...)
- Integration and analysis of massive, complex and heterogeneous datasets (Big Data)

Systems Biology

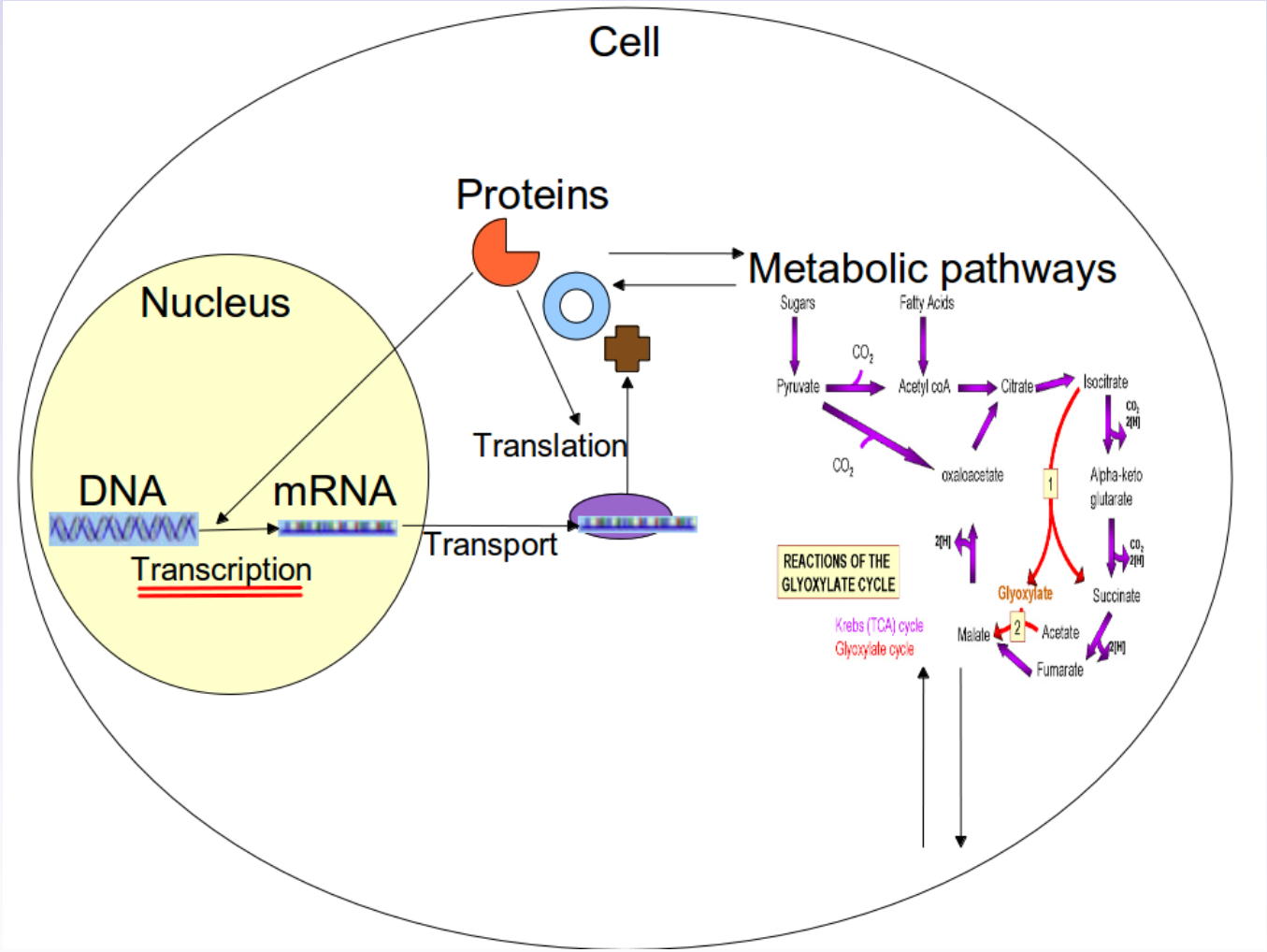
- Systems Biology: interdisciplinary field which studies the complex networks of interactions occurring in biological systems
- Development of models and approaches to reveal emergent properties of cells, tissues and organs, which work as an integrated system
- Tipically involves studies of several types of biological networks (gene regulation, metabolic, protein interactions, cell signaling, etc...)
- Integration and analysis of massive, complex and heterogeneous datasets (Big Data)

Universidade Federal do ABC

Systems Biology

- Systems Biology: interdisciplinary field which studies the complex networks of interactions occurring in biological systems
- Development of models and approaches to reveal emergent properties of cells, tissues and organs, which work as an integrated system
- Tipically involves studies of several types of biological networks (gene regulation, metabolic, protein interactions, cell signaling, etc...)
- Integration and analysis of massive, complex and heterogeneous datasets (Big Data)

Systems Biology



Two main research topics in Systems Biology

- Inference, modeling and simulation of gene regulatory networks (GRN)
- Prioritization of genes associated to complex diseases

Universidade Federal do ABC

GRN inference - Motivation

- Cell control: result of a multivariate activity of genes
- Derivation of general laws on how the cell control works
- Identification of genes associated to certain biochemical features
- Investigation on how to control the dynamics of the biological system and the best way to do it (most practical, least costly, ...)
- Inference of parameters of a GRN from experimental data is one of the greatest challenges of bioinformatics
  - Small number of samples (dozens) with huge dimensionality (thousands of genes)

Universidade Federal do ABC

GRN inference - Motivation

- Cell control: result of a multivariate activity of genes
- Derivation of general laws on how the cell control works
- Identification of genes associated to certain biochemical features
- Investigation on how to control the dynamics of the biological system and the best way to do it (most practical, least costly, ...)
- Inference of parameters of a GRN from experimental data is one of the greatest challenges of bioinformatics
  - Small number of samples (dozens) with huge dimensionality (thousands of genes)

Universidade Federal do ABC

GRN inference - Motivation

- Cell control: result of a multivariate activity of genes
- Derivation of general laws on how the cell control works
- Identification of genes associated to certain biochemical features
- Investigation on how to control the dynamics of the biological system and the best way to do it (most practical, least costly, ...)
- Inference of parameters of a GRN from experimental data is one of the greatest challenges of bioinformatics
    - Small number of samples (dozens) with huge dimensionality (thousands of genes)

Universidade Federal do ABC

GRN inference - Motivation

- Cell control: result of a multivariate activity of genes
- Derivation of general laws on how the cell control works
- Identification of genes associated to certain biochemical features
- Investigation on how to control the dynamics of the biological system and the best way to do it (most practical, least costly, ...)
- Inference of parameters of a GRN from experimental data is one of the greatest challenges of bioinformatics
    - Small number of samples (dozens) with huge dimensionality (thousands of genes)

Universidade Federal do ABC

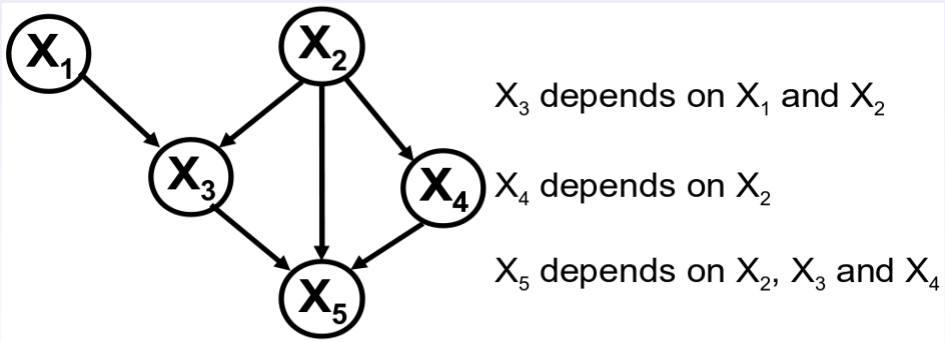
GRN inference - Motivation

- Cell control: result of a multivariate activity of genes
- Derivation of general laws on how the cell control works
- Identification of genes associated to certain biochemical features
- Investigation on how to control the dynamics of the biological system and the best way to do it (most practical, least costly, ...)
- Inference of parameters of a GRN from experimental data is one of the greatest challenges of bioinformatics
  - Small number of samples (dozens) with huge dimensionality (thousands of genes)

Universidade Federal do ABC

GRN inference - Definition

- GRNs are gene interaction networks where the expression level of a gene is controlled by expression levels of other genes
  - Gene expression signal: abundance of transcribed mRNA
  - They can be viewed as graphs where nodes correspond to genes and edges correspond to dependences between genes



Universidade Federal do ABC

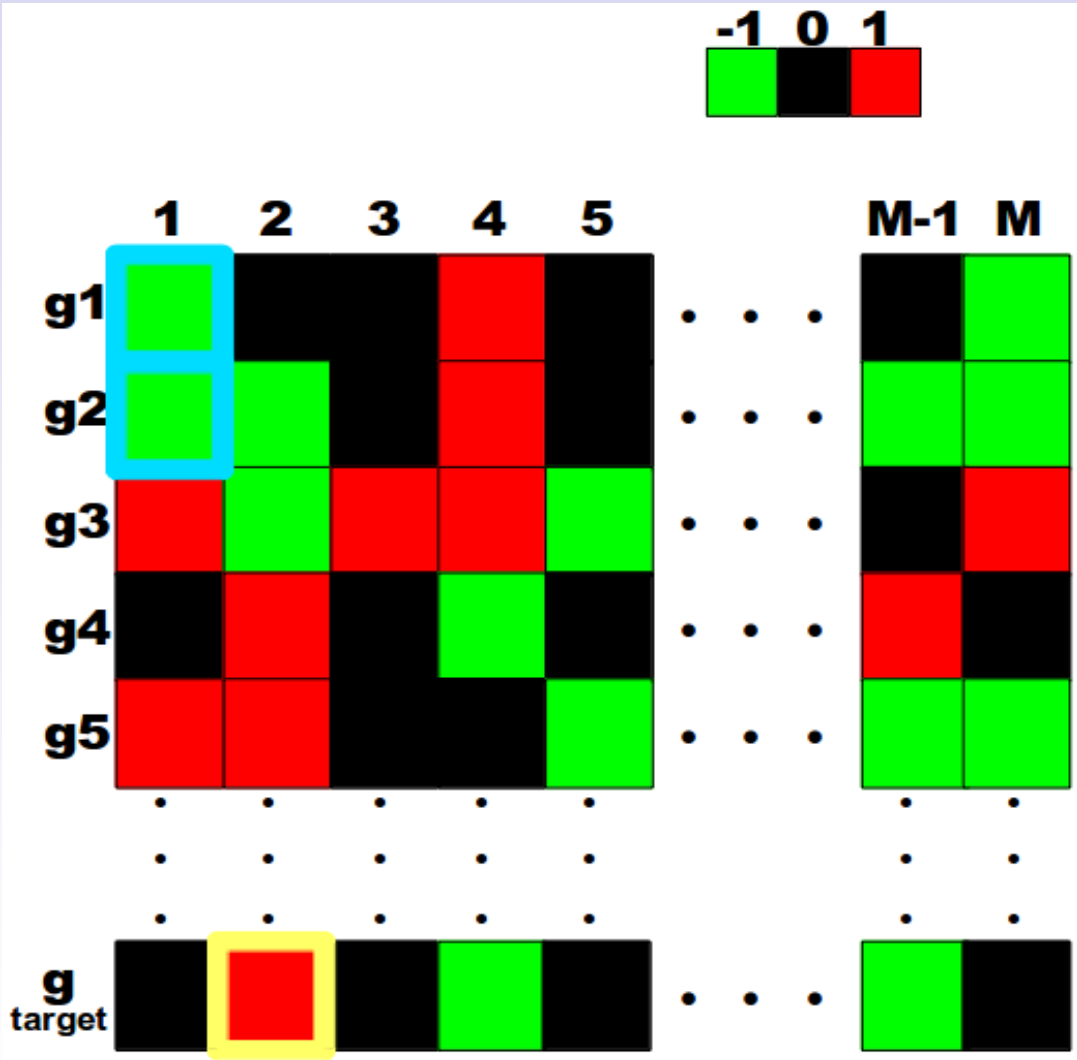
GRN inference - Approach: feature selection

- How to measure the degree of dependence of a gene with regard to other genes?
  - Feature selection
  - Given a target gene, apply a feature selection (search) algorithm which tries to obtain the most relevant genes subset to describe the target behavior
  - Relevance criterion: e.g., mutual information (based on entropy), coefficient of determination (Bayesian error based)

GRN inference - Approach: feature selection

- How to measure the degree of dependence of a gene with regard to other genes?
  - Feature selection
  - Given a target gene, apply a feature selection (search) algorithm which tries to obtain the most relevant genes subset to describe the target behavior
  - Relevance criterion: e.g., mutual information (based on entropy), coefficient of determination (Bayesian error based)

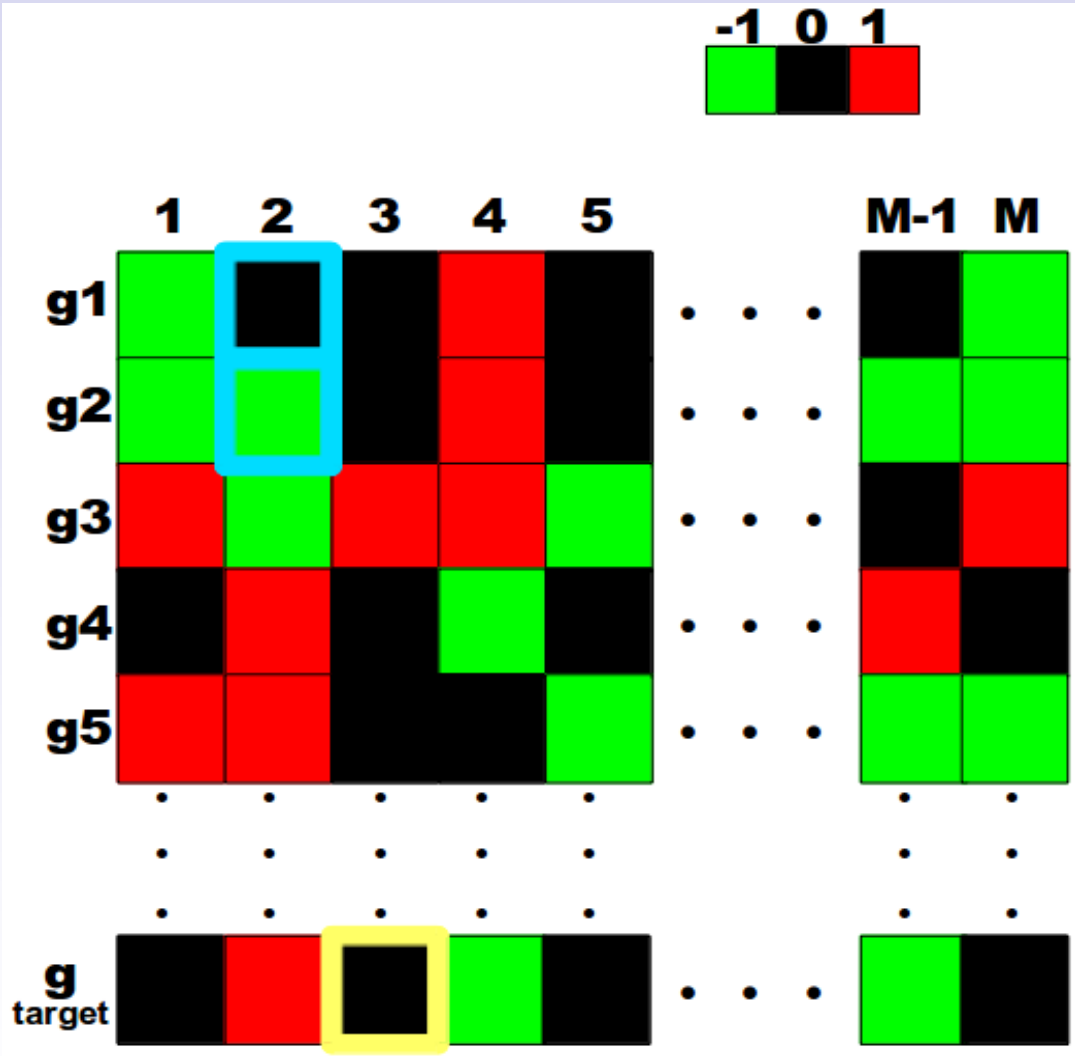
GRN inference - Approach: feature selection



g1	g2	g target −1	0	1
−1	−1	0	0	**1**
−1	0	0	0	0
−1	1	0	0	0
0	−1	0	0	0
0	0	0	0	0
0	1	0	0	0
1	−1	0	0	0
1	0	0	0	0
1	1	0	0	0

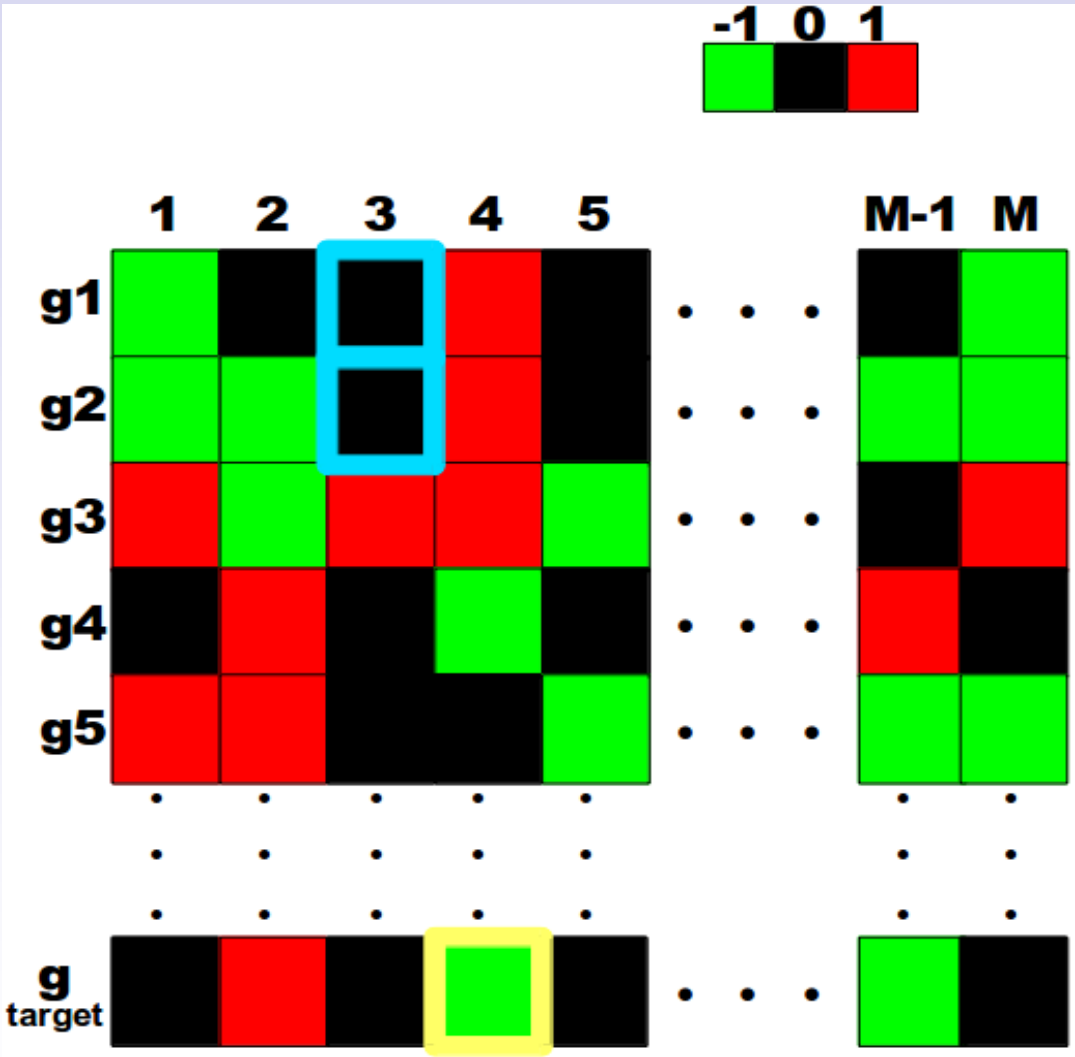
Universidade Federal do ABC

GRN inference - Approach: feature selection



g1	g2	g target −1	0	1
−1	−1	0	0	1
−1	0	0	0	0
−1	1	0	0	0
0	−1	0	**1**	0
0	0	0	0	0
0	1	0	0	0
1	−1	0	0	0
1	0	0	0	0
1	1	0	0	0

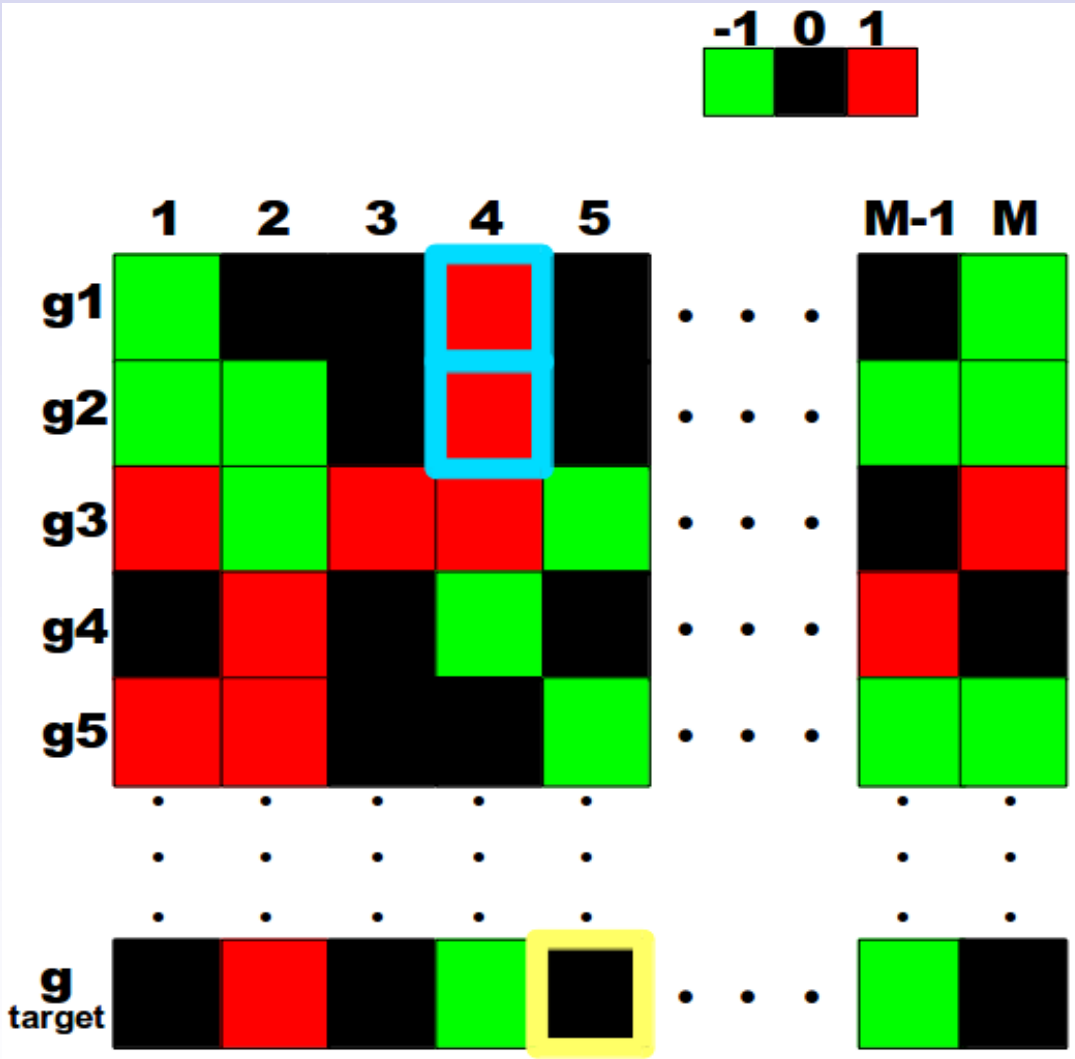
GRN inference - Approach: feature selection



g1	g2	g target −1	0	1
−1	−1	0	0	1
−1	0	0	0	0
−1	1	0	0	0
0	−1	0	1	0
0	0	**1**	0	0
0	1	0	0	0
1	−1	0	0	0
1	0	0	0	0
1	1	0	0	0

Universidade Federal do ABC

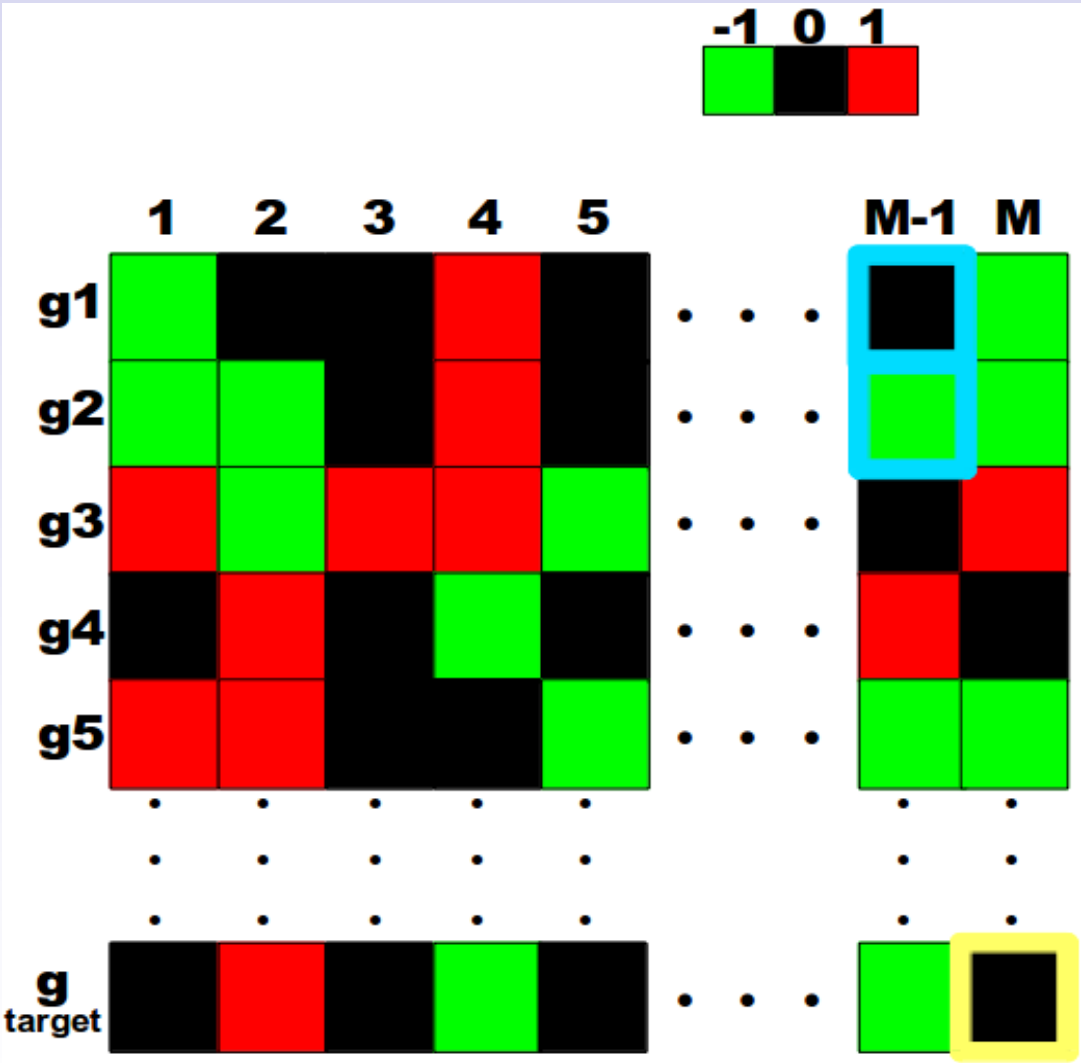
GRN inference - Approach: feature selection



g1	g2	g target −1	0	1
−1	−1	0	0	1
−1	0	0	0	0
−1	1	0	0	0
0	−1	0	1	0
0	0	1	0	0
0	1	0	0	0
1	−1	0	0	0
1	0	0	0	0
1	1	0	**1**	0

Universidade Federal do ABC

GRN inference - Approach: feature selection



g1	g2	g target −1	0	1
−1	−1	0	0	6
−1	0	7	0	0
−1	1	0	4	0
0	−1	0	9	0
0	0	5	0	0
0	1	0	0	1
1	−1	0	0	7
1	0	0	0	0
1	1	8	0	0

Universidade Federal do ABC

GRN inference - Approach: feature selection

g1	g2	g target −1	0	1
−1	−1	0	1	6
−1	0	7	0	0
−1	1	0	4	0
0	−1	0	9	0
0	0	5	0	0
0	1	0	0	1
1	−1	0	0	7
1	0	0	0	0
1	1	8	0	0

Properties of the pair (g1,g2)

- **High** mutual information / CoD
- **Almost perfect** prediction
- **Strong candidate** to be classified between the best pairs (g1 and g2 may be connected to the target gene)

g3	g5	g target −1	0	1
−1	−1	2	2	2
−1	0	3	2	2
−1	1	0	3	1
0	−1	2	4	3
0	0	1	1	2
0	1	1	0	1
1	−1	2	3	1
1	0	1	1	0
1	1	4	2	2

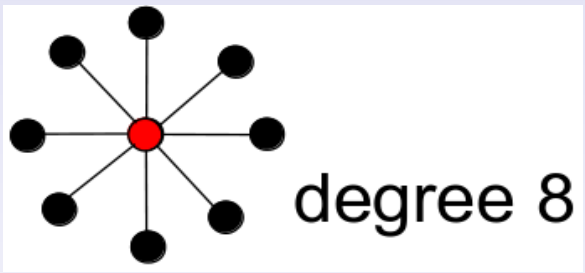
Properties of the pair (g3,g5)

- **Small** mutual information / CoD
- **Very bad** prediction
- **Discarded**

Universidade Federal do ABC

GRN inference - Ongoing research

- How to infer “hubs” from small samples? (and how to decide its input degree?)
  - Hub: gene with large input degree

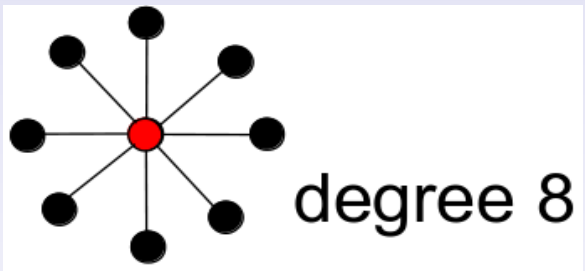


  - In binary systems, a gene with degree 8 has a table with $2^8 = 256$ rows
  - If we have 30 samples, at least 226 rows are not observed (!!!)

Universidade Federal do ABC

GRN inference - Ongoing research

- How to infer “hubs” from small samples? (and how to decide its input degree?)
  - Hub: gene with large input degree

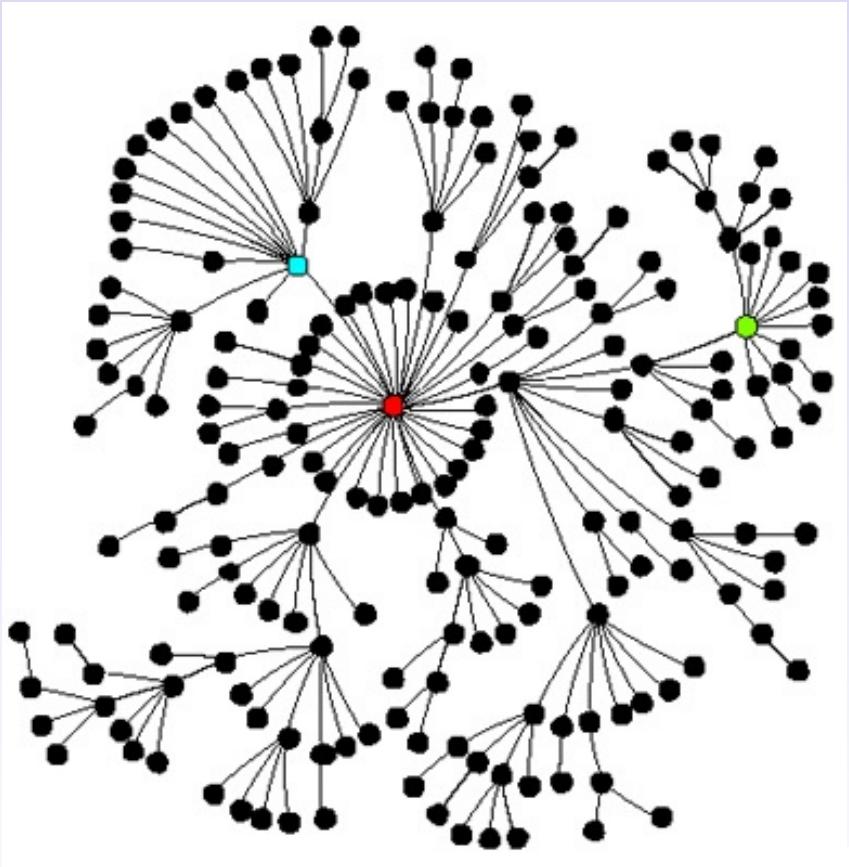


  - In binary systems, a gene with degree 8 has a table with $2^8 = 256$ rows
  - If we have 30 samples, at least 226 rows are not observed (!!!)

Universidade Federal do ABC

GRN inference - Ongoing research

- In particular, inference of hubs is important to infer scale-free networks
  - Small number of nodes with large input degree
  - Large number of nodes with small input degree
- Also known as Barabási-Albert (BA) network model [Barabasi:2004]



Universidade Federal do ABC

GRN inference - SFFS-BA method

- SFFS-BA: GRN inference method guided by topological scale-free properties [Lopes:2014]

Information Sciences 272 (2014) 1–15

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

A feature selection technique for inference of graphs from their known topological properties: Revealing scale-free gene regulatory networks

Fabrício M. Lopes [a,*], David C. Martins Jr. [b], Junior Barrera [c], Roberto M. Cesar Jr. [c,d]

[a] Federal University of Technology - Paraná, Brazil
[b] Federal University of ABC, Brazil
[c] Institute of Mathematics and Statistics, University of São Paulo, Brazil
[d] Brazilian Bioethanol Science and Technology Laboratory (CTBE), Brazil

GRN inference - SFFS-BA (summary and results)

- Adaptation of a classical feature selection method (Sequential Floating Forward Search - SFFS) to look for scale free patterns $\rightarrow$ SFFS-BA
- Comparison involving Sequential Forward Search (SFS), SFFS and SFFS-BA
- Evaluation by simulated artificial gene networks (AGN) generating scale-free topologies and probabilistic Boolean dependences
- Evaluation by real data from *Escherichia coli* microarrays

Universidade Federal do ABC

GRN inference - SFFS-BA (summary and results)

- Adaptation of a classical feature selection method (Sequential Floating Forward Search - SFFS) to look for scale free patterns $\rightarrow$ SFFS-BA
- Comparison involving Sequential Forward Search (SFS), SFFS and SFFS-BA
- Evaluation by simulated artificial gene networks (AGN) generating scale-free topologies and probabilistic Boolean dependences
- Evaluation by real data from *Escherichia coli* microarrays

Universidade Federal do ABC

GRN inference - SFFS-BA (summary and results)

- Adaptation of a classical feature selection method (Sequential Floating Forward Search - SFFS) to look for scale free patterns $\rightarrow$ SFFS-BA
- Comparison involving Sequential Forward Search (SFS), SFFS and SFFS-BA
- Evaluation by simulated artificial gene networks (AGN) generating scale-free topologies and probabilistic Boolean dependences
- Evaluation by real data from *Escherichia coli* microarrays

Universidade Federal do ABC

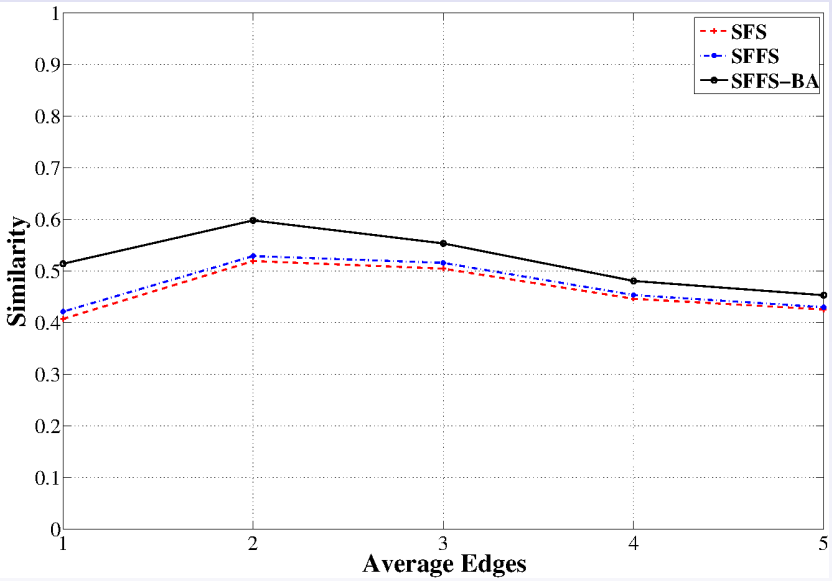
GRN inference - SFFS-BA (summary and results)

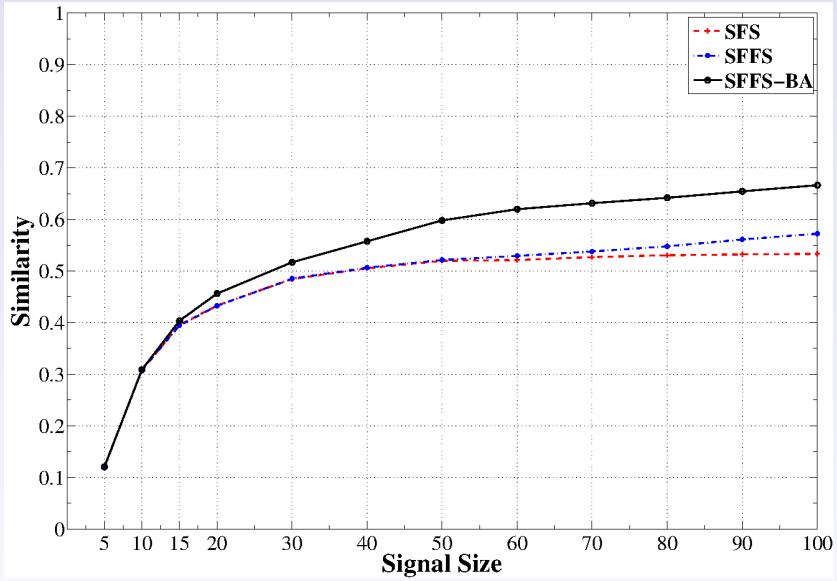
- Adaptation of a classical feature selection method (Sequential Floating Forward Search - SFFS) to look for scale free patterns $\rightarrow$ SFFS-BA
- Comparison involving Sequential Forward Search (SFS), SFFS and SFFS-BA
- Evaluation by simulated artificial gene networks (AGN) generating scale-free topologies and probabilistic Boolean dependences
- Evaluation by real data from *Escherichia coli* microarrays

Universidade Federal do ABC

GRN inference - SFFS-BA (summary and results)

- Artificial gene networks results





Universidade Federal do ABC

GRN inference - SFFS-BA (summary and results)

- *E. coli* results

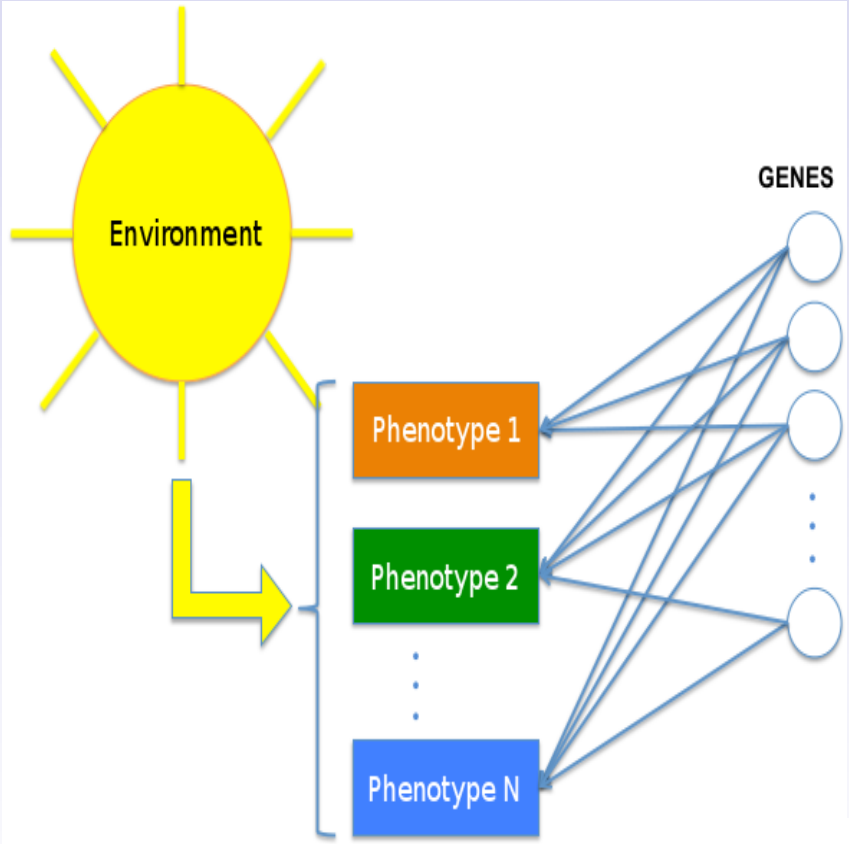
Algorithm	*PPV*	*Sensitivity*	*Similarity*	*AUPR*(%)
SFS	0.1598	0.0169	0.0520	0.0488 (4.88%)
SFFS	0.2416	0.0315	0.0872	0.0629 (6.29%)
SFFS-BA	0.4878	0.0484	0.1537	0.0786 (7.86%)

Universidade Federal do ABC

Universidade Federal do ABC

Complex diseases

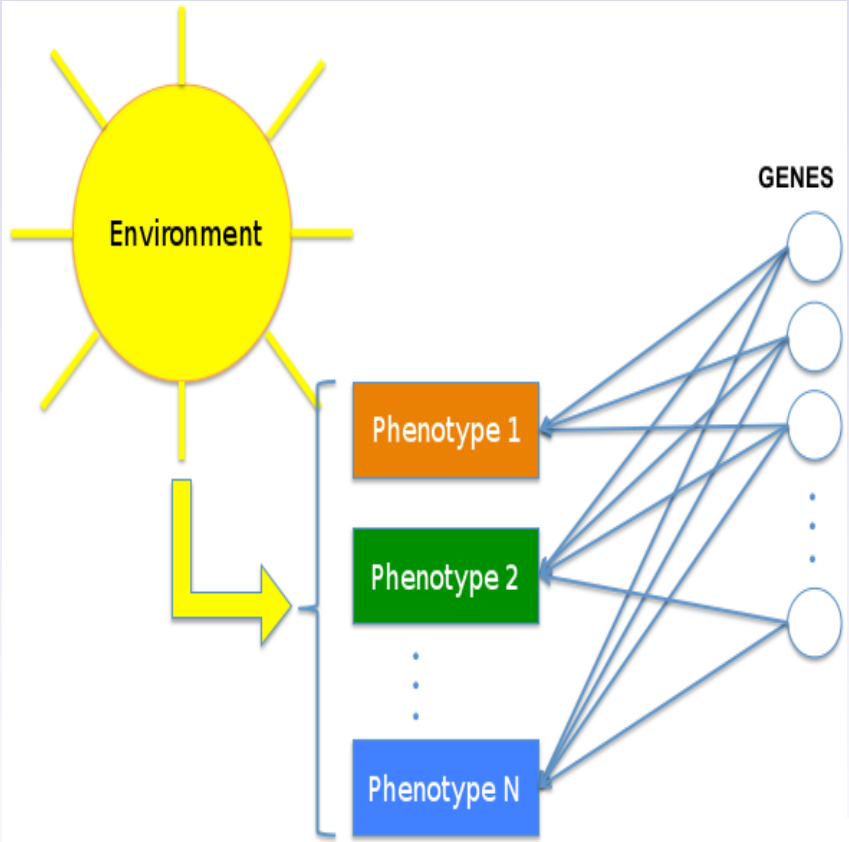
- Complex diseases are polygenic and multifactorial
- Many genes can cause the same phenotype
- A single gene can cause distinct phenotypes
- → studies in complex diseases are challenging
- Distinct studies of a given complex disease usually produce gene lists with very small overlap (small replication)
- Integrative approaches from systems biology, as well as modeling and analysis of complex networks are required



Universidade Federal do ABC

Complex diseases

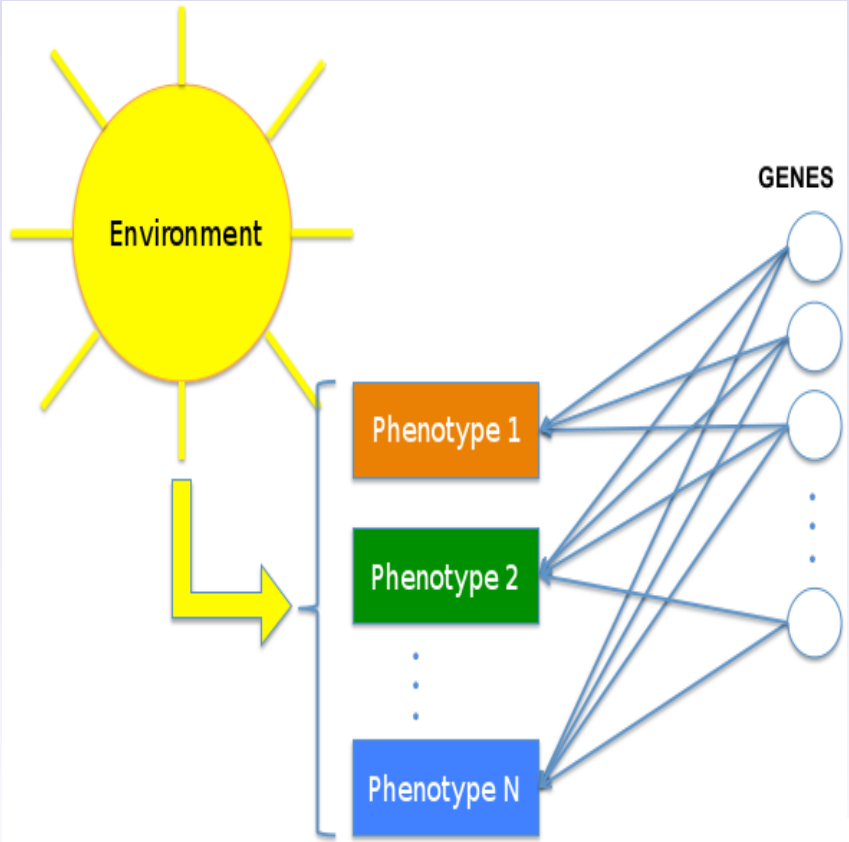
- Complex diseases are polygenic and multifactorial
- Many genes can cause the same phenotype
- A single gene can cause distinct phenotypes
- $\rightarrow$ studies in complex diseases are challenging
- Distinct studies of a given complex disease usually produce gene lists with very small overlap (small replication)
- Integrative approaches from systems biology, as well as modeling and analysis of complex networks are required



Universidade Federal do ABC

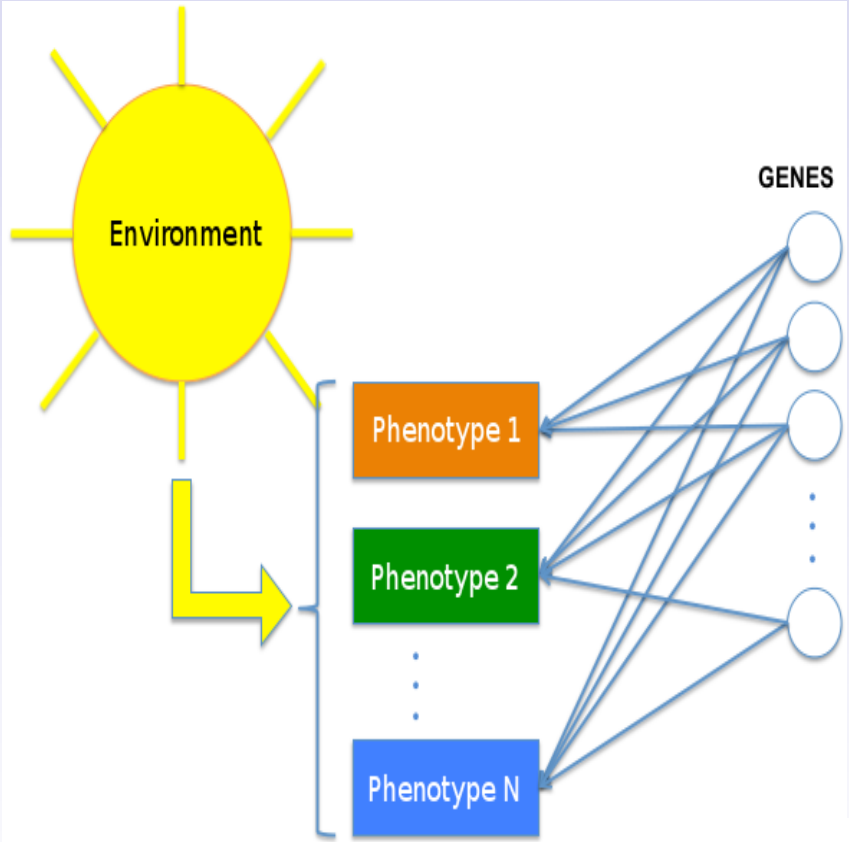
Complex diseases

- Complex diseases are polygenic and multifactorial
- Many genes can cause the same phenotype
- A single gene can cause distinct phenotypes
- $\rightarrow$ studies in complex diseases are challenging
- Distinct studies of a given complex disease usually produce gene lists with very small overlap (small replication)
- Integrative approaches from systems biology, as well as modeling and analysis of complex networks are required



Complex diseases

- Complex diseases are polygenic and multifactorial
- Many genes can cause the same phenotype
- A single gene can cause distinct phenotypes
- $\rightarrow$ studies in complex diseases are challenging
- Distinct studies of a given complex disease usually produce gene lists with very small overlap (small replication)
- Integrative approaches from systems biology, as well as modeling and analysis of complex networks are required



Universidade Federal do ABC

Network Medicine hypotheses [Barabási:2011]

- Modeling of complex network theory properties and Network Medicine hypotheses to prioritize genes
- Hubs: genes/proteins of high degree are considered essential (e.g. TP53)
- Locality Hypothesis: Genes/proteins involved in the same function (or disease phenotype) possess increased tendency to interact with each other
- Modularity Hypothesis: Cellular components associated to a given function (or disease specific phenotype) tend to be in the same cluster
- Parsimony Principle: Molecular pathways usually coincide with the molecular shortest paths between components known to be associated to the disease.

Network Medicine hypotheses [Barabási:2011]

- Modeling of complex network theory properties and Network Medicine hypotheses to prioritize genes
- Hubs: genes/proteins of high degree are considered essential (e.g. TP53)
- Locality Hypothesis: Genes/proteins involved in the same function (or disease phenotype) possess increased tendency to interact with each other
- Modularity Hypothesis: Cellular components associated to a given function (or disease specific phenotype) tend to be in the same cluster
- Parsimony Principle: Molecular pathways usually coincide with the molecular shortest paths between components known to be associated to the disease.

Network Medicine hypotheses [Barabási:2011]

- Modeling of complex network theory properties and Network Medicine hypotheses to prioritize genes
- Hubs: genes/proteins of high degree are considered essential (e.g. TP53)
- Locality Hypothesis: Genes/proteins involved in the same function (or disease phenotype) possess increased tendency to interact with each other
- Modularity Hypothesis: Cellular components associated to a given function (or disease specific phenotype) tend to be in the same cluster
- Parsimony Principle: Molecular pathways usually coincide with the molecular shortest paths between components known to be associated to the disease.

Universidade Federal do ABC

Network Medicine hypotheses [Barabási:2011]

- Modeling of complex network theory properties and Network Medicine hypotheses to prioritize genes
- Hubs: genes/proteins of high degree are considered essential (e.g. TP53)
- Locality Hypothesis: Genes/proteins involved in the same function (or disease phenotype) possess increased tendency to interact with each other
- Modularity Hypothesis: Cellular components associated to a given function (or disease specific phenotype) tend to be in the same cluster
- Parsimony Principle: Molecular pathways usually coincide with the molecular shortest paths between components known to be associated to the disease.

Universidade Federal do ABC

Network Medicine hypotheses [Barabási:2011]

- Modeling of complex network theory properties and Network Medicine hypotheses to prioritize genes
- Hubs: genes/proteins of high degree are considered essential (e.g. TP53)
- Locality Hypothesis: Genes/proteins involved in the same function (or disease phenotype) possess increased tendency to interact with each other
- Modularity Hypothesis: Cellular components associated to a given function (or disease specific phenotype) tend to be in the same cluster
- Parsimony Principle: Molecular pathways usually coincide with the molecular shortest paths between components known to be associated to the disease.

NERI method

- NERI: NEtwork Medicine Relative Importance [Simões:2015]

Simões *et al. BMC Bioinformatics* 2015, **16**(Suppl 19):S9
http://www.biomedcentral.com/1471-2105/16/S19/S9

BMC Bioinformatics

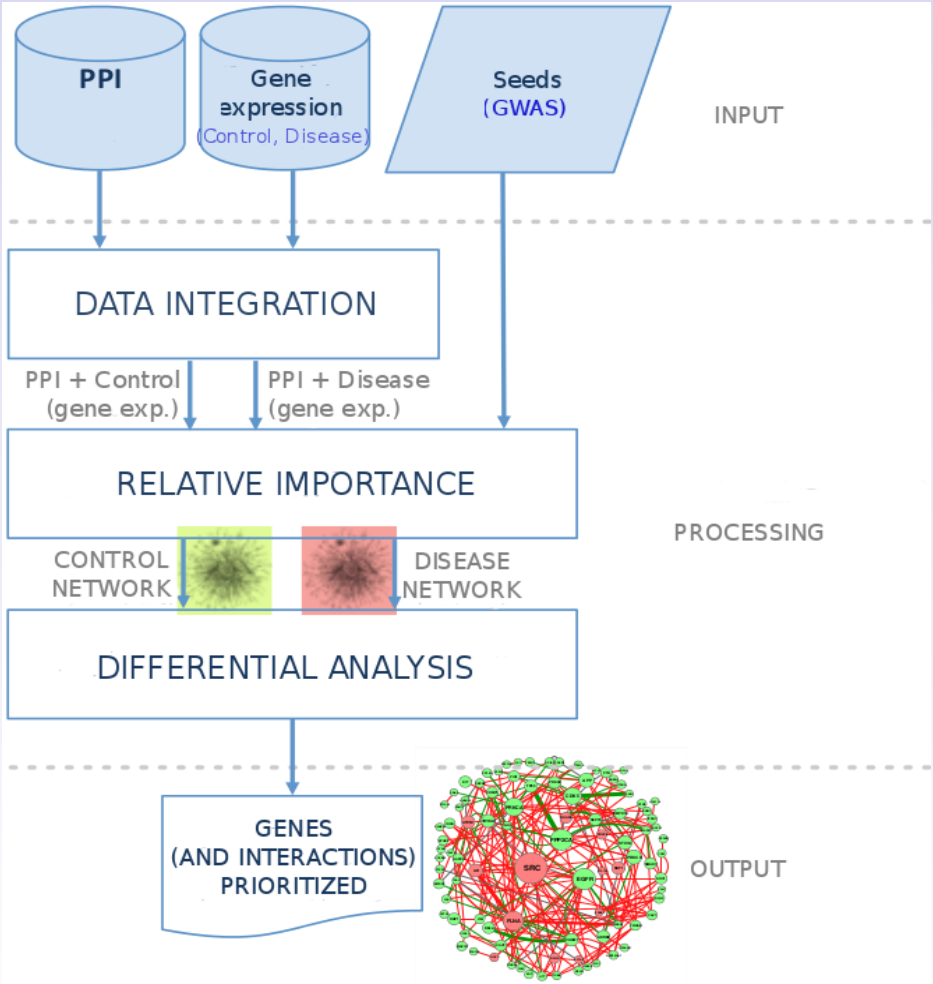
RESEARCH Open Access

NERI: network-medicine based integrative approach for disease gene prioritization by relative importance

Sérgio N Simões[1,2*], David C Martins Jr[3], Carlos AB Pereira[1], Ronaldo F Hashimoto[1], Helena Brentani[4,5,6]

Universidade Federal do ABC

NERI method - Overview



Universidade Federal do ABC

NERI method - Overview

- Modeling the locality, modularity and parsimony hypotheses
- Integration of seeds (genome-wide association studies - GWAS), gene expression and protein-protein interaction networks (PPI) data
- Obtainment of two PPI network cuttings around the neighborhood of the seeds: one for control and one for disease conditions
- Cutting: best shortest paths (according to gene expression concordance) connecting the seeds
- Two relative importances are assigned to each gene: one for control and another for disease condition
- Relative importance: mix of fequency along the shortest paths between seeds, concordance of expression along these paths and proximity to the seeds
- Genes with the largest difference in their two relative importances (control/disease) are prioritized

Universidade Federal do ABC

NERI method - Overview

- Modeling the locality, modularity and parsimony hypotheses
- Integration of seeds (genome-wide association studies - GWAS), gene expression and protein-protein interaction networks (PPI) data
- Obtainment of two PPI network cuttings around the neighborhood of the seeds: one for control and one for disease conditions
- Cutting: best shortest paths (according to gene expression concordance) connecting the seeds
- Two relative importances are assigned to each gene: one for control and another for disease condition
- Relative importance: mix of fequency along the shortest paths between seeds, concordance of expression along these paths and proximity to the seeds
- Genes with the largest difference in their two relative importances (control/disease) are prioritized

Universidade Federal do ABC

NERI method - Overview

- Modeling the locality, modularity and parsimony hypotheses
- Integration of seeds (genome-wide association studies - GWAS), gene expression and protein-protein interaction networks (PPI) data
- Obtainment of two PPI network cuttings around the neighborhood of the seeds: one for control and one for disease conditions
- Cutting: best shortest paths (according to gene expression concordance) connecting the seeds
- Two relative importances are assigned to each gene: one for control and another for disease condition
- Relative importance: mix of fequency along the shortest paths between seeds, concordance of expression along these paths and proximity to the seeds
- Genes with the largest difference in their two relative importances (control/disease) are prioritized

Universidade Federal do ABC

NERI method - Overview

- Modeling the locality, modularity and parsimony hypotheses
- Integration of seeds (genome-wide association studies - GWAS), gene expression and protein-protein interaction networks (PPI) data
- Obtainment of two PPI network cuttings around the neighborhood of the seeds: one for control and one for disease conditions
- Cutting: best shortest paths (according to gene expression concordance) connecting the seeds
- Two relative importances are assigned to each gene: one for control and another for disease condition
- Relative importance: mix of fequency along the shortest paths between seeds, concordance of expression along these paths and proximity to the seeds
- Genes with the largest difference in their two relative importances (control/disease) are prioritized

Universidade Federal do ABC

NERI method - Overview

- Modeling the locality, modularity and parsimony hypotheses
- Integration of seeds (genome-wide association studies - GWAS), gene expression and protein-protein interaction networks (PPI) data
- Obtainment of two PPI network cuttings around the neighborhood of the seeds: one for control and one for disease conditions
- Cutting: best shortest paths (according to gene expression concordance) connecting the seeds
- Two relative importances are assigned to each gene: one for control and another for disease condition
- Relative importance: mix of fequency along the shortest paths between seeds, concordance of expression along these paths and proximity to the seeds
- Genes with the largest difference in their two relative importances (control/disease) are prioritized

Universidade Federal do ABC

NERI method - Overview

- Modeling the locality, modularity and parsimony hypotheses
- Integration of seeds (genome-wide association studies - GWAS), gene expression and protein-protein interaction networks (PPI) data
- Obtainment of two PPI network cuttings around the neighborhood of the seeds: one for control and one for disease conditions
- Cutting: best shortest paths (according to gene expression concordance) connecting the seeds
- Two relative importances are assigned to each gene: one for control and another for disease condition
- Relative importance: mix of fequency along the shortest paths between seeds, concordance of expression along these paths and proximity to the seeds
- Genes with the largest difference in their two relative importances (control/disease) are prioritized

Universidade Federal do ABC

NERI method - Overview

- Modeling the locality, modularity and parsimony hypotheses
- Integration of seeds (genome-wide association studies - GWAS), gene expression and protein-protein interaction networks (PPI) data
- Obtainment of two PPI network cuttings around the neighborhood of the seeds: one for control and one for disease conditions
- Cutting: best shortest paths (according to gene expression concordance) connecting the seeds
- Two relative importances are assigned to each gene: one for control and another for disease condition
- Relative importance: mix of fequency along the shortest paths between seeds, concordance of expression along these paths and proximity to the seeds
- Genes with the largest difference in their two relative importances (control/disease) are prioritized

Universidade Federal do ABC

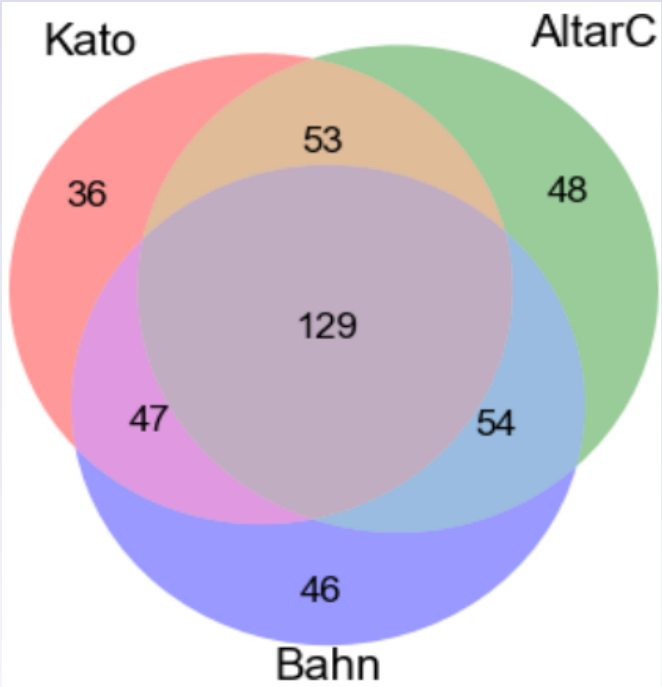
NERI method - Results

- Case study: Schizophrenia
- 30 seed genes obtained from association studies (also known as “core genes”)
- KATO (33 C, 34 D), ALTAR (29 C, 21 D) and BAHN (33 C, 34 D) gene expression databases

Universidade Federal do ABC

NERI method - Results

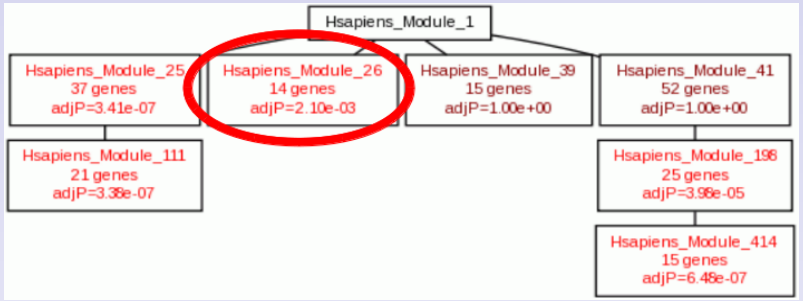
- Intersection of the top 10% genes ranked by the method in each study KATO, ALTAR and BAHN
- Intersection p-value $< 10^{-58}$ (hipergeometric test)
- Large replication among the gene expression studies

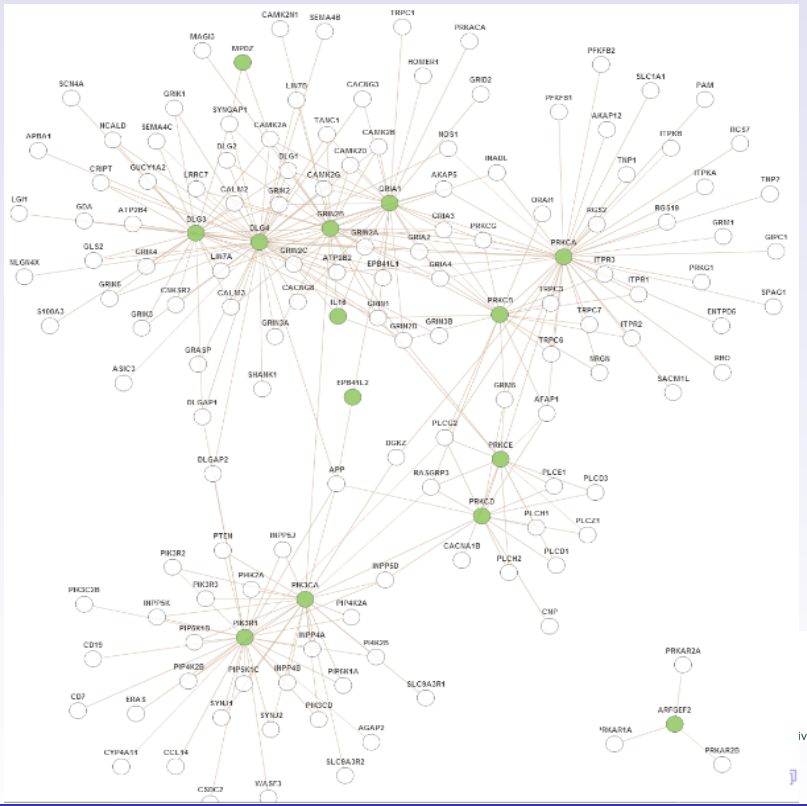


Universidade Federal do ABC

NERI method - Results

- WebGestalt protein interaction enrichment of the overlap genes list (129 genes)
- Module 26 (glutamate receptor signaling pathway) with 14 genes (green nodes)
  - Such function is known to be associated to schizophrenia





Universidade Federal do ABC

Conclusion

Data integration and complex networks analyses are keys to improve GRN inference and gene prioritization processes

Acknowledgements

- FAPESP (procs. 11/50761-2, 13/26644-1, 14/10488-3, 15/01587-0)
- CNPq (procs. 559955/2010-3, 306991/2011-0, 304955/2014-0)
- CAPES
- Collaborators from UFABC, USP, UTFPR, IFES, Hospital A. C. Carmargo, Indian Institute of Technology and Texas A&M University

References

- [Barabási:2004] A. L. Barabási, Z. Oltvai (2004). Network biology: Understanding the cell's functional organization. Nature Reviews Genetics, 5:101-113.
- [Barabási:2011] A. L. Barabasi, N. Gulbahce, and J. Loscalzo (2011). Network medicine: A network-based approach to human disease. Nat Rev Genet, 12(1):56-68.
- [Lopes:2014] F. M. Lopes, D. C. Martins-Jr, J. Barrera, and R. M. Cesar-Jr (2014). A feature selection technique for inference of graphs from their known topological properties: Revealing scale-free gene regulatory networks. Information Sciences, 272:1-15.
- [Simoes:2015] S. N. Simoes, D. C. Martins-Jr, C. A. B. Pereira, R. F. Hashimoto, and H. Brentani. (2015). NERI: network-medicine based integrative approach for disease gene prioritization by relative importance. BMC Bioinformatics, 16(19):1-14.

Thank You!

Universidade Federal do ABC