

Network topology information-based prediction of human disease genes

Marcio L. Acencio

Pedro R. Costa

Daniel Nolli

Ney Lemke

Department of Physics and Biophysics
Institute of Biosciences - São Paulo State University
Botucatu - São Paulo - Brazil

2008

Outline

- 1 Introduction
- 2 Proposal
- 3 Methodology
- 4 Results
- 5 Discussion
- 6 Acknowledgements

Outline

- 1 Introduction**
- 2 Proposal
- 3 Methodology
- 4 Results
- 5 Discussion
- 6 Acknowledgements

Disease Genes: definition

Genes at which mutations are known to cause heritable genetic disease in humans

Disease Genes Identification: what we could learn from it?

- Better understanding of the underlying molecular mechanisms of the disease in question.
- Serve as direct targets for better treatments:
 - Pharmacogenetics
 - Interventions
- Predictions of susceptibility to and course of the disease
- Knowledge for treatment or prevention

Disease Genes Identification: experimental approaches

- Main strategy - genetic linkage analysis followed by positional cloning:
 - Time-consuming and labor-intensive processes.
 - Result: dozens to hundreds of candidate genes (usually 20 to 200 genes).

Disease Genes Identification: computational approaches

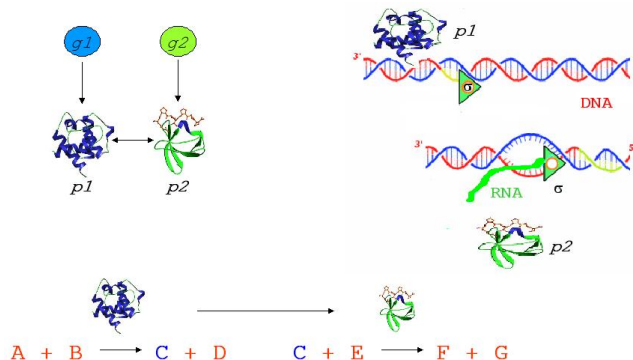
- Algorithms for prediction of disease genes based on shared functional annotation to known disease genes;
- Machine learning-based approaches based on sequence features (protein size, evolutionary conservation rates, number of exons) on known disease genes;
- Machine learning-based approaches based on topological features (e.g. degree, average distance to disease genes) of human protein-protein interactions.

Integrated Biological Networks

Gene interaction networks containing simultaneously:

- Protein-protein physical interactions
- Metabolic interactions
- Transcriptional regulation interactions

Integrated Biological Networks



Outline

- 1 Introduction
- 2 Proposal**
- 3 Methodology
- 4 Results
- 5 Discussion
- 6 Acknowledgements

Motivation

- Integration generates knowledge.
- Human IBN might provide us with new opportunity for predicting disease genes.
- Similar approach successfully applied by our group (Silva et al., 2008*) to predict essential genes in *Escherichia coli*

*J. P. M. Silva, M. L. Acencio, J. C. M. Mombach, R. Vieira, J. C. Silva, M. Sinigaglia, and N. Lemke (2008): *In silico* network topology-based prediction of gene essentiality. *Physica A*. 387, 1049-1055

Proposal

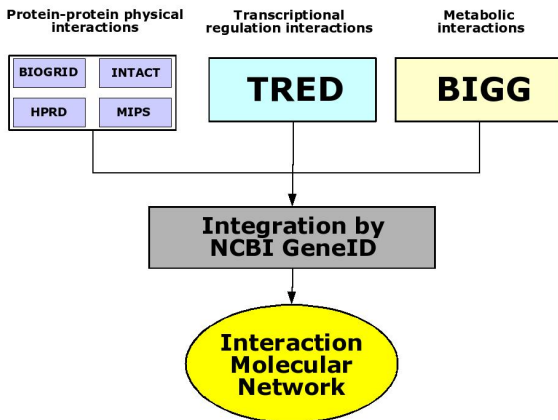
Prediction of human disease genes based on topological features of the human integrated biological network (IBN)

Outline

- 1 Introduction
- 2 Proposal
- 3 Methodology**
- 4 Results
- 5 Discussion
- 6 Acknowledgements

Human IBN construction

Only literature-curated, experimentally verified interactions



Human IBN construction

Metabolic Interactions:

- Nodes = genes coding for enzymes
- Interactions = metabolites (reactants or products)
- Interactions via currency metabolites (e.g. H^+ , H_2O , ATP, ADP, pyrophosphate, orthophosphate, NAD^+ , NADH, $NADP^+$, NADPH) were removed.

Software Infrastructure

- Mathematica[®] (Wolfram Research) - Integration and Statistical Analysis
- Weka (Wakaito Environment for Knowledge Analysis) - Machine Learning
- NetworkX (Python package for graph theory) - Graph properties

Machine Learning Algorithms

- J48 classifier:
 - WEKA's implementation of C4.5 algorithm;
 - Builds decision trees from a set of attributes and training data using the concept of information entropy;

Machine Learning Algorithms

- LMT (*Logistic Model Tree*):
 - Is a tree model where each leaf is a logistic regression model.
 - Generates probabilities for the instances.

Decision-tree construction: attributes

For each gene (centrality measures):

- Clustering coefficient (c);
- Betweenness centrality:
 - Through protein-protein interactions (*inbet*);
 - Through metabolic interactions (*inbetmet*);
 - Through transcriptional regulation interactions (*inbetreg*).
- Closeness centrality (*closeness*);
- Number of "twin genes", i.e. genes with identical number and types of interactions.

Decision-tree construction: attributes

For each gene (type of interaction):

- Number of protein-protein physical interactions (*PPI*);
- Number of metabolic interactions:
 - Number of reactants (incoming edges; *Metin*)
 - Number of products (outgoing edges; *Metout*)
- Number of transcriptional regulation interactions:
 - Number of regulating genes (incoming edges; *Regin*)
 - Number of regulated genes (outgoing edges; *Regout*)

Decision-tree construction: training set

- Disease genes: 1,893 genes from The OMIM Morbid Map
 - Replicated 4 × to solve the data imbalance problem (total: 8,330 disease genes)
- Non-disease genes: Remaining 8,400 genes in the constructed human IBN

Decision-tree: performance evaluation

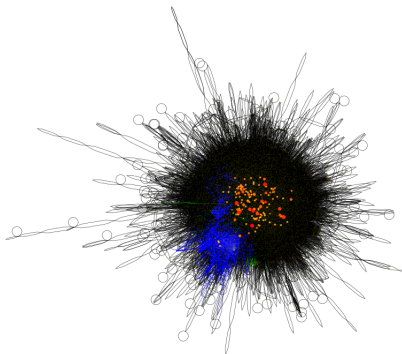
- Generated decision-tree applied to training data itself;
- Calculation of recall, precision and accuracy;
- Construction of ROC curve.

Outline

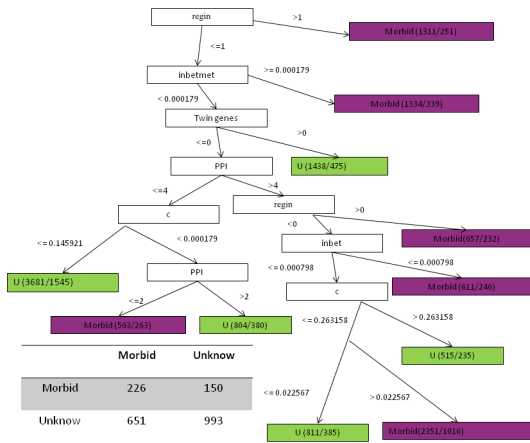
- 1 Introduction
- 2 Proposal
- 3 Methodology
- 4 Results**
- 5 Discussion
- 6 Acknowledgements

Human IBN

- $\approx 10,200$ genes.
- $\approx 64,000$ experimentally verified interactions:
 - $\approx 36,600$ protein-protein physical interactions (57%)
 - $\approx 3,000$ transcriptional regulation interactions (5%)
 - $\approx 24,400$ metabolic interactions (38%)



Analysis of Generated Decision-Tree - J48



Decision-tree: interpretation

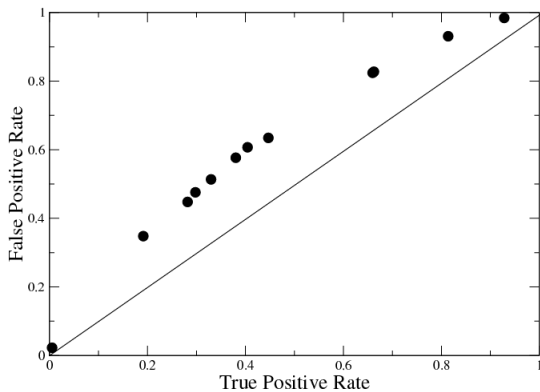
- *Regin*, i.e. number of regulating genes (number of regulating transcription factors), is the most important feature for morbidity.
- Betweenness centrality through metabolic interactions (*inbetmet*) is the second most important feature for morbidity.
- Genes that have a "twin gene" are not morbid.

Decision-tree: performance evaluation - J48

Class	Morbid	Unknown
Morbid	226	150
Unknown	651	993

- Correct Classified Instances: 60%
- Recall: 0.6
- Area under ROC: 0.63

ROC curve J48

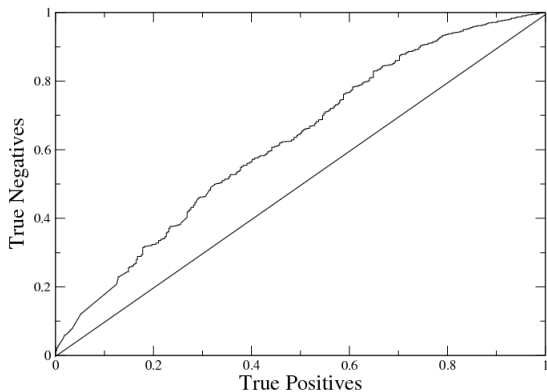


Decision-tree: performance evaluation - LMT

Class	Morbid	Unknown
Morbid	238 \pm 13	137 \pm 13
Unknown	675 \pm 29	968 \pm 29

- Correct Classified Instances: 60% \pm 2
- Recall: 0.63 \pm 0.02

ROC curve LMT



Classification of genes in disease loci - LMT

Classifier applied to the cystic fibrosis locus as determined by linkage analysis.

Known disease gene: **CFTR**

Gene	Probability of morbidity
MET	0.88
CFTR	0.86
CAV1	0.82
WNT2	0.64

Classification of genes in disease loci - LMT

Classifier applied to the infantile hypertrophic pyloric stenosis locus as determined by linkage analysis.

Disease gene(s) not known; only candidates

Gene	Probability of morbidity
MMP20	0.88
MMP12	0.80
MMP3	0.78
ATM	0.78
MMP10	0.76
ACAT1	0.74
JOSD3	0.72
TRPC6	0.69

Outline

- 1 Introduction
- 2 Proposal
- 3 Methodology
- 4 Results
- 5 Discussion**
- 6 Acknowledgements

Drawbacks

- The regulatory network is too incomplete - only 5% of known transcription factors present in human IBN.
 - Currently, there is only one freely accessible human transcriptional regulation network database biased to disease-related transcription factors.
- There is no known method to classify a gene as nondisease gene: classifiers trained on genes not known to be involved in disease.

Perspectives

- Inclusion of more transcriptional regulation interactions.
- Integration with biological function of gene and disease phenotype information to predict disease-specific genes.
- Generalization of this method can be a very useful tool for detection of targets for new drugs.
- We can use information about target drugs to improve the selection of good targets.

Outline

- 1 Introduction
- 2 Proposal
- 3 Methodology
- 4 Results
- 5 Discussion
- 6 Acknowledgements**

Acknowledgements

We wish to thank FAPESP (research grants 2007/02827-9 and 2007/01213-7) and CNPq (research grant 474278/2006-9) for supporting this work.