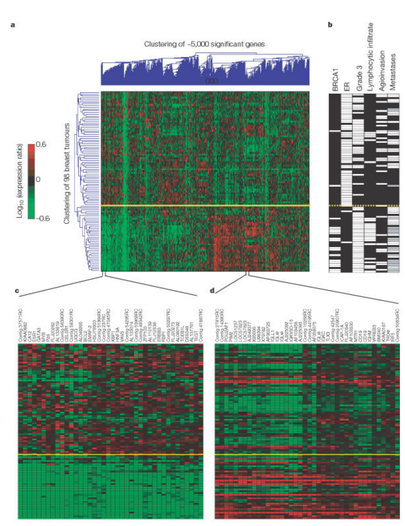
Analyzing How Protein Interaction Networks Improve Classification Performance in Gene Expression Data Analysis

Adrin Jalali and Nico Pfeifer
Department of Computational Biology and Applied Algorithmics
Max Planck Institute for Informatics
Saarbrücken, Germany

MAX-PLANCK-GESELLSCHAFT

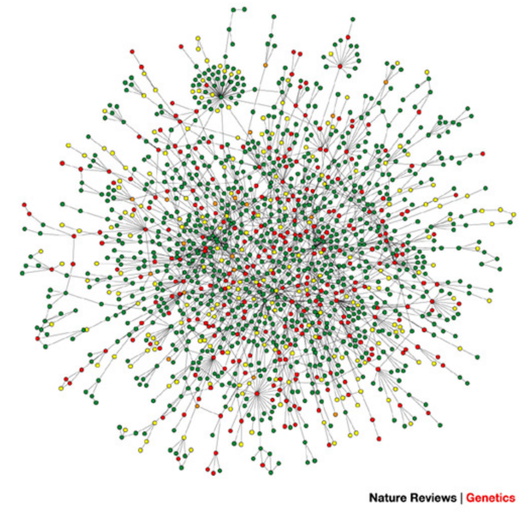
Motivation



Unsupervised two-dimensional cluster analysis of 98 breast tumours [1]

Problem statement:

- Input: Gene expression data
- Output: Prognosis (Poor vs. Good), Metastases
- Goal: Classify samples and find important genes
- Issue: Hard to classify due to large number of features (genes) compared to number of samples ($\sim \mathbf{22000} \gg \mathbf{98}$)



Yeast protein interaction network. The colour of a node indicates the phenotypic effect of removing the corresponding protein (red = lethal, green = non-lethal, orange = slow growth, yellow = unknown) [2]

Method

1. SVM modified objective function [3]

$$\min_{\mathbf{w},w_0}\left\{\frac{1}{2}\|\mathbf{w}\|^2+\frac{1}{2}\beta\sum_{(j,k)\in E}(w_j-w_k)^2\right\}$$

s.t.:

$$\forall i\in\{1,\cdots,n\}:(\mathbf{w}\mathbf{x}_i+w_0)y_i\geq 1$$

2. Dual Problem

$$\max_{\alpha}\left\{\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j(\mathbf{x}_i^T\mathbf{L})(\mathbf{L}^T\mathbf{x}_j)\right\}$$

$$\mathbf{L}\mathbf{L}^T=(\mathbf{I}+\beta\mathbf{B})^{-1}$$

s.t.:

$$\forall i\in\{1,\cdots,n\}:\sum_{i=1}^{n}\alpha_i y_i=0$$

$$\forall i\in\{1,\cdots,n\}:\alpha_i\geq 0$$

Laplacian matrix:

$$\mathbf{B}=\mathbf{D}-\mathbf{A}$$

3. Dual to Primal

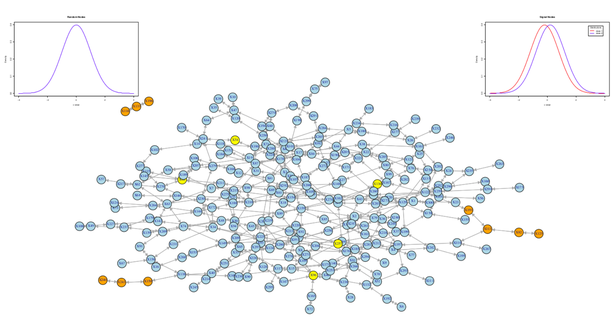
$$\mathbf{w}=(\mathbf{I}+\beta\mathbf{B})^{-1}\sum_{i=1}^{n}\alpha_i y_i\mathbf{x}_i$$

What we do:

- Reverse engineer the learned machine to extract important genes after using the network information.
- Solve SVM problem for original and transformed data.
- Calculate $\mathbf{w}$ for both models.
- Compute for each pair of nodes, for each model:

$$Score(i,j)=\frac{|w_i|+|w_j|}{2}\times e^{-max(d_G(i,j),1)}$$

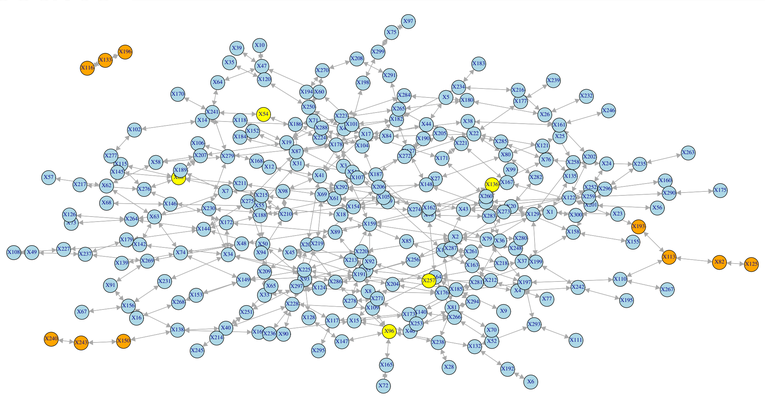
- Report pairs with highest scores for both trained models.



Blue: random gene, Orange: Signal node being a member of a pathway of signal nodes, Yellow: A lonely signal node

- Signal nodes (genes): $f(n)=\begin{cases}N(-\mu,1) & \text{if } n \text{ is in class } 1\\ N(\mu,1) & \text{if } n \text{ is in class } 2\end{cases}$
- Random nodes (non-informative genes): $f(n)=N(0,1)$

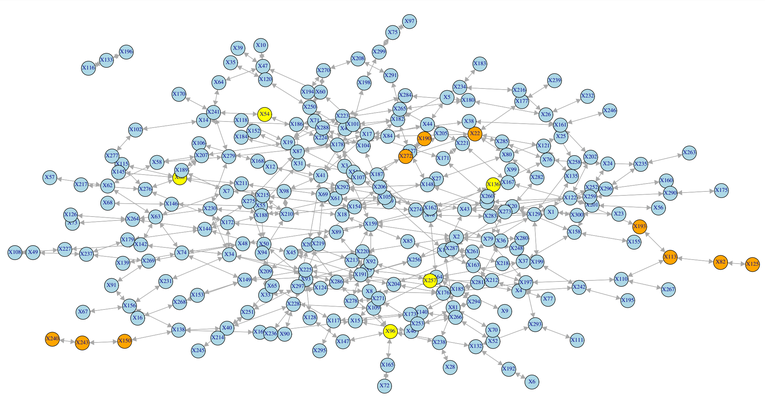
Results



Easy

Original		Transformed	
X196	X196	X196	X196
X233	X233	X196	X133
X88	X88	X133	X116
X116	X116	X95	X95
X197	X197	X39	X39
X148	X148	X59	X59
X148	X273	X243	X243
X160	X160	X114	X114
X95	X95	X243	X150
X88	X115	X39	X47
X53	X8	X150	X150
X195	X195	X125	X125

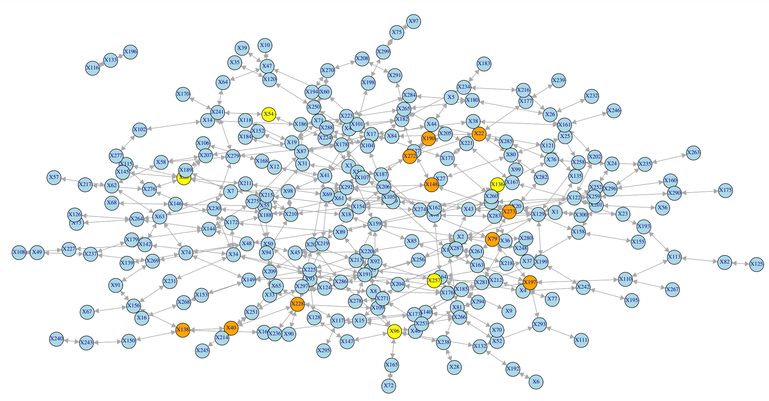
AUC (Original):	60.6
AUC (Transformed):	62.4
wc p-value (paired):	5.669e-09



Medium

Original		Transformed	
X190	X190	X233	X233
X233	X233	X112	X112
X277	X277	X190	X272
X190	X127	X86	X86
X272	X272	X243	X150
X106	X106	X150	X150
X150	X150	X246	X246
X88	X215	X106	X106
X51	X51	X35	X35
X73	X73	X247	X247
X162	X162	X272	X22
X112	X112	X100	X100

AUC (Original):	60.1
AUC (Transformed):	61.5
wc p-value (paired):	1.383e-06



Hard

Original		Transformed	
X190	X190	X233	X233
X233	X233	X112	X112
X88	X88	X86	X86
X190	X127	X272	X272
X26	X26	X205	X205
X272	X272	X146	X68
X101	X41	X298	X298
X22	X22	X90	X90
X146	X146	X100	X100
X278	X278	X297	X297
X88	X115	X127	X148
X148	X148	X247	X247

AUC (Original):	60.2
AUC (Transformed):	62.5
wc p-value (paired):	8.151e-13

References and Acknowledgment

Acknowledgment:
Prof. Dr. Dr. Thomas Lengauer, Sarvesh Nikumbh, Nora Speicher, Anna Feldmann.

References:
1. van't Veer, Laura J., et al. "Gene expression profiling predicts clinical outcome of breast cancer." nature 415.6871 (2002): 530-536.
2. Barabási, Albert-László, and Zoltan N. Oltvai. "Network biology: understanding the cell's functional organization." Nature Reviews Genetics 5.2 (2004): 101-113.
3. Lavi, Ofer, Gideon Dror, and Ron Shamir. "Network-induced classification kernels for gene expression profile analysis." Journal of Computational Biology 19.6 (2012): 694-709.