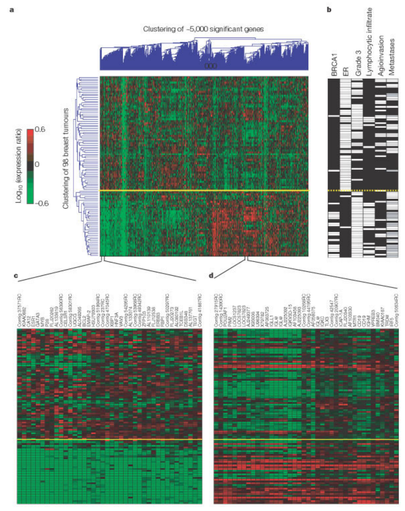
Analyzing How Protein Interaction Networks Improve Classification Performance in Gene Expression Data Analysis

Adrin Jalali

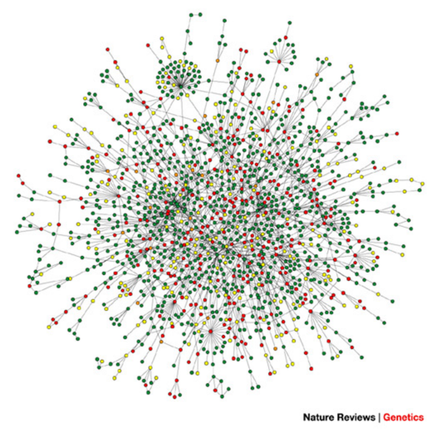
MPI for Informatics

MAX-PLANCK-GESELLSCHAFT

Data



1. Input: Gene expression data 2. Output: Prognosis (Poor vs. Good), Metastases 3. Goal: Classify and find important genes 4. Issue: Hard to classify due to huge number of features (genes) compared to number of samples ($\sim \mathbf{22000} \gg \mathbf{98}$)
- Laura J. van 't Veer et.al. Nature, (2002)



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), A. BarabÃ¡si, Z. Oltvai, Nature Reviews Genetics, (2004)

Method

1. SVM modified objective function

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

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.:

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

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

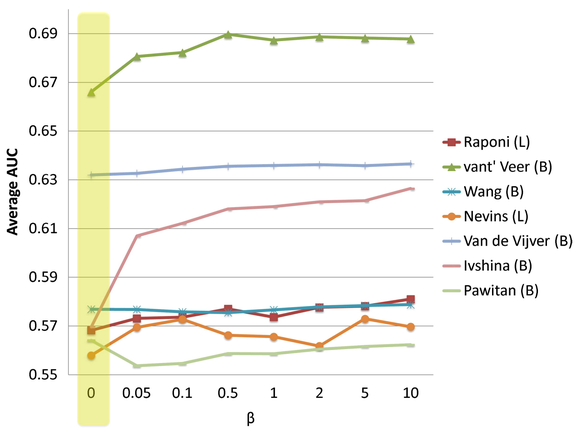
Laplacian matrix:

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

3. Dual to Primal

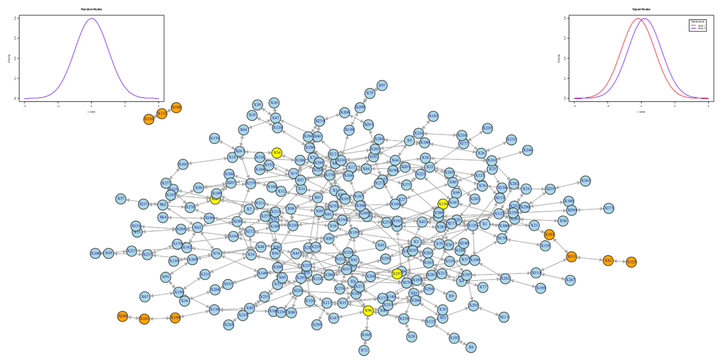
$$\mathbf{w}=(\mathbf{I}+\beta\mathbf{B})^{-1}\sum_{i=1}^{n}\alpha_i y_i\mathbf{x}_i$$

Method (cntd.)



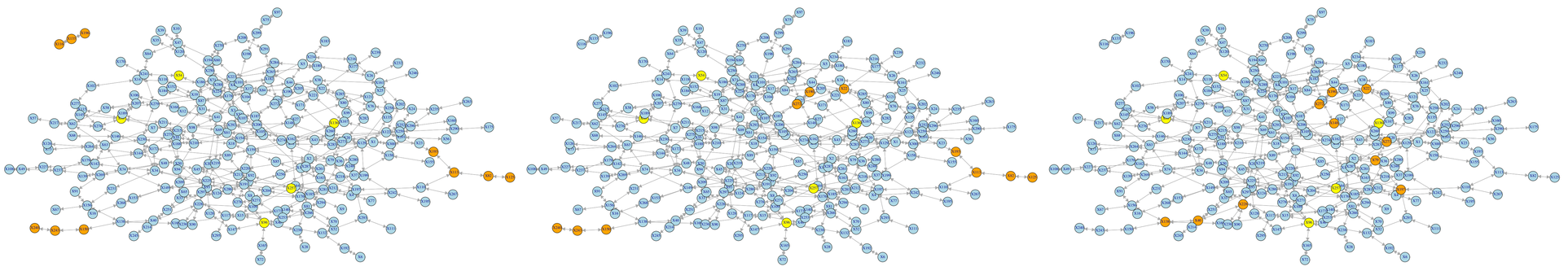
1. Co-expressed genes tend to be close in the PPI-Network. 2. Exploit this fact to modify the SVM objective function - called NICK. Ofer Lavi, et.al., Journal of Computational Biology, (2012)

- 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: $\boldsymbol{Score(i,j)=\frac{|w_i|+|w_j|}{2}\times e^{-max(d_G(i,j),1)}}$
- 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): $\boldsymbol{f(n)=\begin{cases}N(-\mu,1) & \text{if } n \text{ is in class } 1\\ N(\mu,1) & \text{if } n \text{ is in class } 2\end{cases}}$ - Random nodes (non-informative genes): $\boldsymbol{f(n)=N(0,1)}$

Results



Easy Medium Hard

Original				Transformed				Original				Transformed			
X196	X196	X53	X53	X196	X196	X233	X233	X190	X190	X104	X104	X233	X233	X190	X190
X233	X233	X39	X39	X196	X133	X133	X133	X233	X233	X190	X272	X112	X112	X240	X240
X88	X88	X196	X133	X133	X116	X116	X116	X277	X277	X88	X88	X190	X272	X240	X243
X116	X116	X127	X127	X95	X95	X240	X240	X190	X127	X165	X165	X86	X86	X243	X243
X197	X197	X127	X148	X39	X39	X240	X243	X272	X272	X272	X22	X243	X150	X190	X127
X148	X148	X150	X150	X59	X59	X106	X106	X106	X106	X165	X96	X150	X150	X272	X272
X148	X273	X116	X133	X243	X243	X106	X168	X150	X150	X250	X250	X246	X246	X298	X298
X160	X160	X96	X96	X114	X114	X168	X168	X88	X215	X22	X22	X106	X106	X125	X125
X95	X95	X273	X273	X243	X150	X56	X56	X51	X51	X28	X28	X35	X35	X125	X82
X88	X115	X40	X40	X39	X47	X298	X298	X73	X73	X35	X35	X247	X247	X272	X69
X53	X8	X53	X164	X150	X150	X247	X247	X162	X162	X113	X113	X272	X22	X82	X82
X195	X195	X56	X56	X125	X125	X83	X83	X112	X112	X277	X102	X100	X100	X257	X257

Original				Transformed			
X190	X190	X101	X101	X233	X233	X190	X190
X233	X233	X190	X272	X112	X112	X190	X272
X88	X88	X297	X297	X86	X86	X190	X127
X190	X127	X93	X93	X272	X272	X272	X205
X26	X26	X138	X138	X205	X205	X146	X146
X272	X272	X272	X22	X146	X68	X68	X68
X101	X41	X123	X123	X298	X298	X272	X22
X22	X22	X101	X198	X90	X90	X127	X127
X146	X146	X228	X228	X100	X100	X272	X69
X278	X278	X72	X72	X297	X297	X72	X72
X88	X115	X96	X96	X127	X148	X155	X155
X148	X148	X112	X112	X247	X247	X196	X196

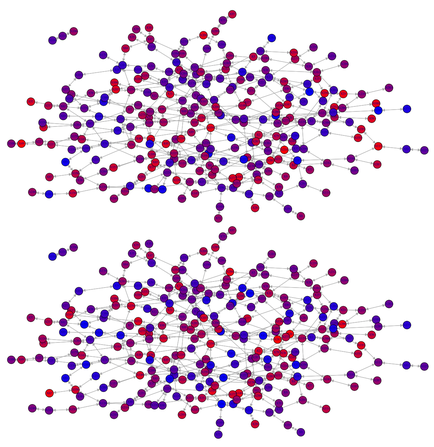
Easy	
AUC (Original):	60.6
AUC (Transformed):	62.4
wc p-value (paired):	5.669e-09
Medium	
AUC (Original):	60.1
AUC (Transformed):	61.5
wc p-value (paired):	1.383e-06
Hard	
AUC (Original):	60.6
AUC (Transformed):	62.4
wc p-value (paired):	5.669e-09

Performance summary of three cases

Next...

- Estimate gene expression value probability distributions for samples of class A.
- For each sample in class A and B:
  - Weight genes according to their distance from means of estimated distributions.
  - Extract higher weighted parts of generated network.
- Use a graph kernel for labeled graphs to classify extracted graphs.

- Random Walks ($\boldsymbol{A^k}$: number of walks of length $\boldsymbol{k}$)
- Sub-Tree Kernels
- Shortest Paths Kernels (1-shortest path, k-shortest paths)
- Graphlet Kernels (Isomorphism proved for $\boldsymbol{n=k+1, n\leq 11}$)
- Laplacian Matrix - Eigenvalues (Isomorph$(\boldsymbol{G_1,G_2})\iff\exists\boldsymbol{P}:\boldsymbol{B(G_1)=P^tB(G_2)P}$), number of connected components = number of $\boldsymbol{\lambda_i=0}$



Sample graphs of two classes

References and Acknowledgment

Acknowledgment:
Prof. Dr. Dr. Thomas Lengauer, Dr. Nico Pfeifer, Sarvesh Nimkubh, 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. 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.
3. Barabási, Albert-László, and Zoltan N. Oltvai. "Network biology: understanding the cell's functional organization." Nature Reviews Genetics 5.2 (2004): 101-113.