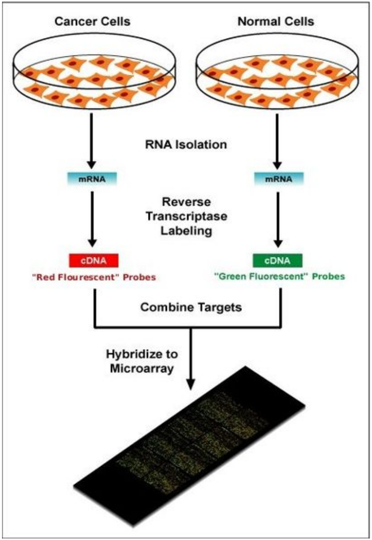
PPI Networks and Gene Expression

Adrin Jalali

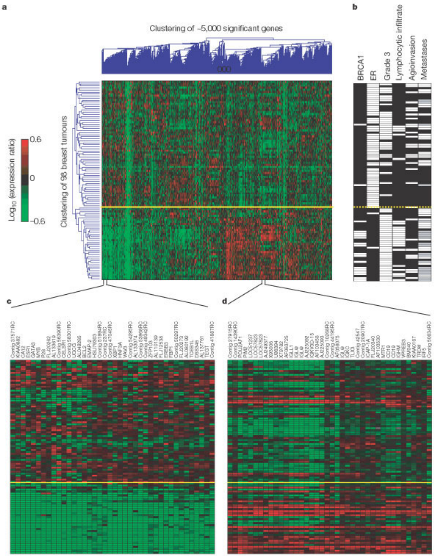
July 8, 2013

Microarray Gene Expression



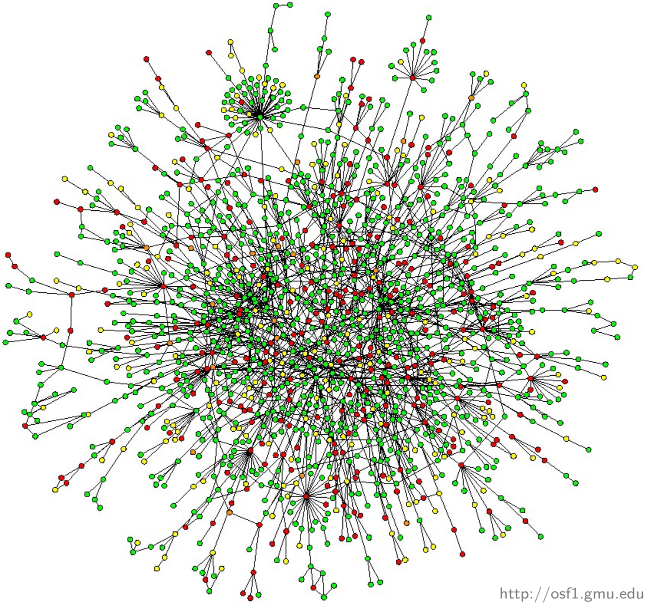
credit: en.wikipedia.org

Van't Veer breast-cancer data



Laura J. van 't Veer et.al. Nature, (2002)

Yeast Protein Interaction Network



credit:
http://osf1.gmu.edu/ rcouch/chem665.htm

JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 19, Number 6, 2012

Pp. 694–709
DOI: 10.1089/cmb.2012.0065

Network-Induced Classification Kernels for Gene Expression Profile Analysis

OFER LAVI,[1,3] GIDEON DROR,[2] and RON SHAMIR[1]

ABSTRACT

Computational classification of gene expression profiles into distinct disease phenotypes has been highly successful to date. Still, robustness, accuracy, and biological interpretation of the results have been limited, and it was suggested that use of protein interaction information jointly with the expression profiles can improve the results. Here, we study three aspects of this problem. First, we show that interactions are indeed relevant by showing that co-expressed genes tend to be closer in the network of interactions. Second, we show that the improved performance of one extant method utilizing expression and interactions is not really due to the biological information in the network, while in another method this is not the case. Finally, we develop a new kernel method—called NICK—that integrates network and expression data for SVM classification, and demonstrate that overall it achieves better results than extant methods while running two orders of magnitude faster.

Key word: algorithms.

1. INTRODUCTION

In the past decade, gene expression profiles based on DNA microarrays have been widely used to detect disease biomarkers. These profiles, measuring thousands of gene expression levels simultaneously, served as the basis for feature selection and classification methods and have been shown to provide better prognosis than prior models (Paik et al., 2006). However, the biomarker sets created by such methods have several drawbacks: Analysis often results in hundreds of genes, biological interpretation of the selected genes is difficult, and the overlap between the sets of genes selected as features in similar studies is very poor (Ein-Dor et al., 2005). In addition, genes selected in one dataset often do not perform well on other datasets (Chuang et al., 2007). This lack of robustness of biomarker selection was decisively demonstrated by Ein-Dor et al. (2005). To overcome this problem, Ein-Dor et al. suggested enlarging the sample size, or dividing the sample in advance into known homogeneous subsets based on some prior knowledge, and analyzing each subset separately (Sørlie et al., 2003).

We would like then to develop methods for detecting sets of biomarkers that (1) are more meaningful biologically and (2) are more stable across different studies. Such sets would be more useful for downstream biological research. The two goals do not always go hand in hand; for example, Hwang et al. (2008)

[1]Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv, Israel.
[2]Yahoo! Research, Haifa, Israel.
[3]IBM Haifa Research Lab, Haifa, Israel.

NICK

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.:

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

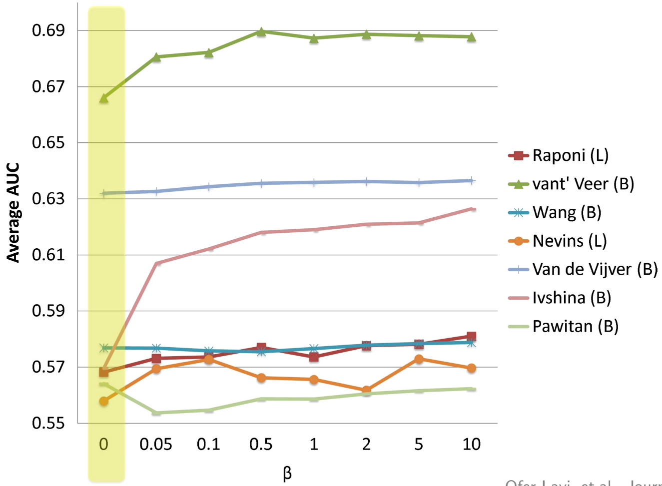
$$\forall i \in \{1,\cdots,n\} : \alpha_i \geq 0$$

3. Dual to Primal

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

Ofer Lavi, et.al., Journal of Computational Biology, (2012)

NICK Performance Summary



Ofer Lavi, et.al., Journal of Computational Biology, (2012)

Synthesize data

1. A random graph
2. Signal nodes:

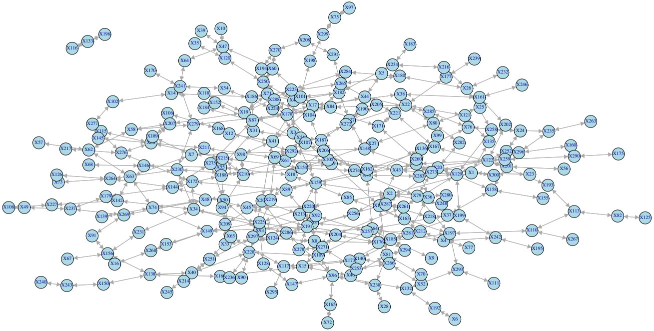
$$f(n) = \begin{cases} N(-\mu, 1) & \text{if } n \text{ is in class 1} \\ N(\mu, 1) & \text{if } n \text{ is in class 2} \end{cases}$$

3. Random nodes:

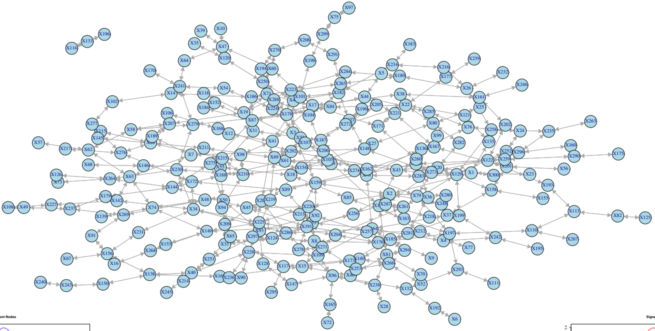
$$f(n) = N(0, 1)$$

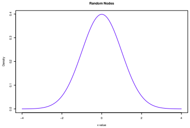
4. Pathway: 2, 3, or 4 connected signal nodes.

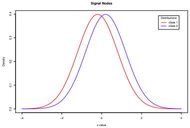
Synthesized data



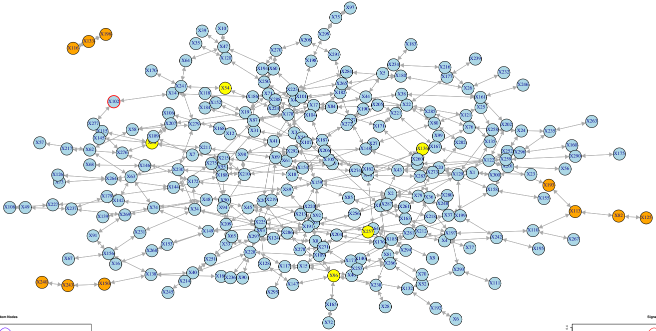
Synthesized data

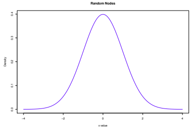


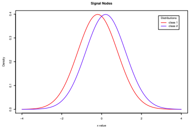




Synthesized data

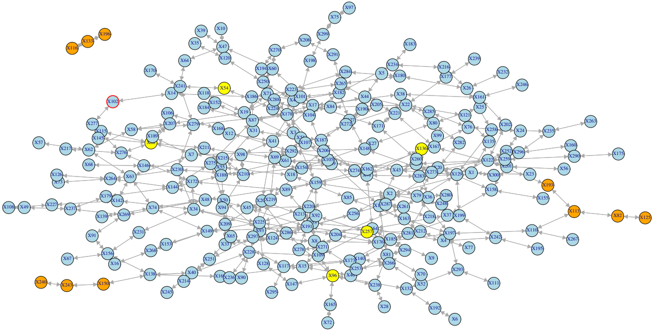






Synthesized data easy scenario

AUC:	60.6
X196	X196
X53	X53
X233	X233
X39	X39
X88	X88
X196	X133
X116	X116
X127	X127
X197	X197
X127	X148
X148	X148
X150	X150
X148	X273
X116	X133
X160	X160
X96	X96
X95	X95
X273	X273
X88	X115
X40	X40
X53	X8
X53	X164
X195	X195
X56	X56
X113	X113
X179	X179
X96	X147
X240	X240
X197	X110
X197	X293

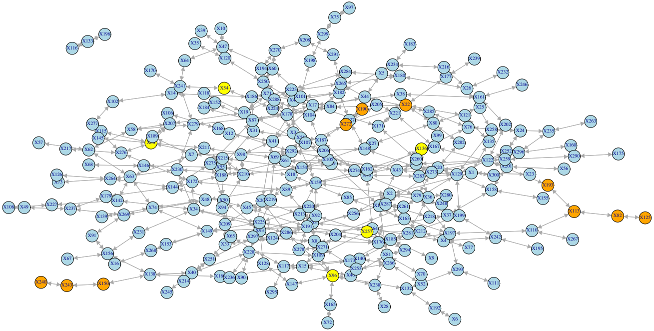


wc p-value (paired): 5.669e-09

AUC:	62.4
X196	X196
X233	X233
X196	X133
X133	X133
X133	X116
X116	X116
X95	X95
X240	X240
X39	X39
X240	X243
X59	X59
X106	X106
X243	X243
X106	X168
X114	X114
X168	X168
X243	X150
X56	X56
X39	X47
X298	X298
X150	X150
X247	X247
X125	X125
X83	X83
X125	X82
X82	X82
X160	X160
X195	X195
X196	X116
X82	X113

Synthesized data medium scenario

AUC:	60.1
X190	X190
X104	X104
X233	X233
X190	X272
X277	X277
X88	X88
X190	X127
X165	X165
X272	X272
X272	X22
X106	X106
X165	X96
X150	X150
X250	X250
X88	X215
X22	X22
X51	X51
X28	X28
X73	X73
X35	X35
X162	X162
X113	X113
X112	X112
X277	X102
X96	X96
X104	X206
X88	X207
X104	X182
X96	X132
X277	X276

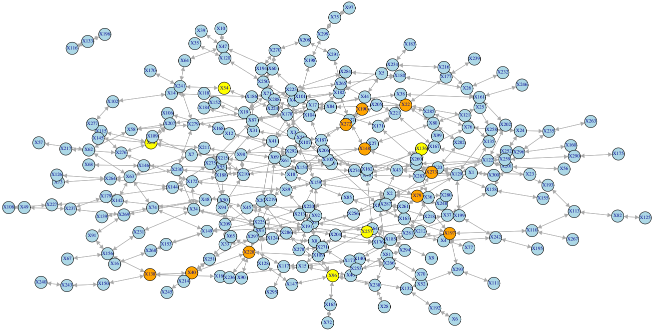


wc p-value (paired): 1.383e-06

AUC:	61.5
X233	X233
X190	X190
X112	X112
X240	X240
X190	X272
X240	X243
X86	X86
X243	X243
X243	X150
X190	X127
X150	X150
X272	X272
X246	X246
X298	X298
X106	X106
X125	X125
X35	X35
X125	X82
X247	X247
X272	X69
X272	X22
X82	X82
X100	X100
X257	X257
X82	X113
X28	X28
X272	X205
X131	X131
X113	X113
X246	X161

Synthesized data hard scenario

AUC:	60.2
X190	X190
X101	X101
X233	X233
X190	X272
X88	X88
X297	X297
X190	X127
X93	X93
X26	X26
X138	X138
X272	X272
X272	X22
X101	X41
X123	X123
X22	X22
X101	X198
X146	X146
X228	X228
X278	X278
X72	X72
X88	X115
X96	X96
X148	X148
X112	X112
X35	X35
X96	X132
X161	X161
X155	X155
X297	X251
X138	X169



wc p-value (paired): 8.151e-13

AUC:	62.5
X233	X233
X190	X190
X112	X112
X190	X272
X86	X86
X190	X127
X272	X272
X272	X205
X205	X205
X146	X146
X146	X68
X68	X68
X298	X298
X272	X22
X90	X90
X127	X127
X100	X100
X272	X69
X297	X297
X72	X72
X127	X148
X155	X155
X247	X247
X196	X196
X148	X148
X101	X101
X21	X21
X106	X106
X90	X228
X297	X251

Van't Veer

AUC:	72.9
X85453	X85453
X85453	X92140
X6605	X6605
X56886	X56886
X10640	X10640
X8817	X8817
X56894	X56894
X6605	X332
X5733	X5733
X57758	X57758
X7532	X7532
X51	X51
X7566	X7566
X3267	X3267
X89953	X89953
X5713	X5713
X5193	X5193
X5365	X5365
X10874	X10874
X5982	X5982

AUC:	73.6
X9917	X9917
X84279	X84279
X197370	X197370
X51143	X51143
X58475	X58475
X55585	X55585
X25949	X25949
X54892	X54892
X126695	X126695
X57168	X57168
X10456	X10456
X148223	X148223
X9742	X9742
X253558	X253558
X342527	X342527
X10175	X10175
X83930	X83930
X57035	X57035
X145482	X145482
X57465	X57465

wc p-value (paired):	0.006

Van't Veer

AUC:	72.9
X85453	X85453
X85453	X92140
X6605	X6605
X56886	X56886
X10640	X10640
X8817	X8817
X56894	X56894
X6605	X332
X5733	X5733
X57758	X57758
X7532	X7532
X51	X51
X7566	X7566
X3267	X3267
X89953	X89953
X5713	X5713
X5193	X5193
X5365	X5365
X10874	X10874
X5982	X5982

Node	Degree
X85453	12
X6605	98
X56886	26
X10640	16
X8817	152
X56894	28
X5733	150
X57758	8
X7532	86
X51	172
X7566	16
X3267	56
X89953	4
X5713	126
X5193	32
X5365	70
X10874	132
X5982	172
X92140	20
X332	328

Node	Degree
X9917	0
X84279	0
X197370	0
X51143	0
X58475	0
X55585	0
X25949	0
X54892	0
X126695	0
X57168	0
X10456	0
X148223	0
X9742	0
X253558	0
X342527	0
X10175	0
X83930	0
X57035	0
X145482	0
X57465	0

AUC:	73.6
X9917	X9917
X84279	X84279
X197370	X197370
X51143	X51143
X58475	X58475
X55585	X55585
X25949	X25949
X54892	X54892
X126695	X126695
X57168	X57168
X10456	X10456
X148223	X148223
X9742	X9742
X253558	X253558
X342527	X342527
X10175	X10175
X83930	X83930
X57035	X57035
X145482	X145482
X57465	X57465

wc p-value (paired):	0.006

Idea

1. Estimate gene expression value probability distributions for samples of class A.

Idea

1. Estimate gene expression value probability distributions for samples of class A.
2. For each sample in class A and B:
    1. Extract abnormal genes according to above estimated distributions.
    2. Extract the part of PPI network induced by extracted genes (almost)

Idea

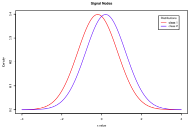
1. Estimate gene expression value probability distributions for samples of class A.
2. For each sample in class A and B:
   1. Extract abnormal genes according to above estimated distributions.
   2. Extract the part of PPI network induced by extracted genes (almost)
3. Use a graph kernel for labeled graphs to classify extracted graphs.

Idea

1. Estimate gene expression value probability distributions for samples of class A.
2. For each sample in class A and B:
    1. Extract abnormal genes according to above estimated distributions.
    2. Extract the part of PPI network induced by extracted genes (almost)
3. Use a graph kernel for labeled graphs to classify extracted graphs.
4. Extract common sub-graphs from individual graphs that seem to be helping the classification.

Idea

1. Estimate gene expression value probability distributions for samples of class A.
2. For each sample in class A and B:
    1. Extract abnormal genes according to above estimated distributions.
    2. Extract the part of PPI network induced by extracted genes (almost)
3. Use a graph kernel for labeled graphs to classify extracted graphs.
4. Extract common sub-graphs from individual graphs that seem to be helping the classification.



Weighted idea

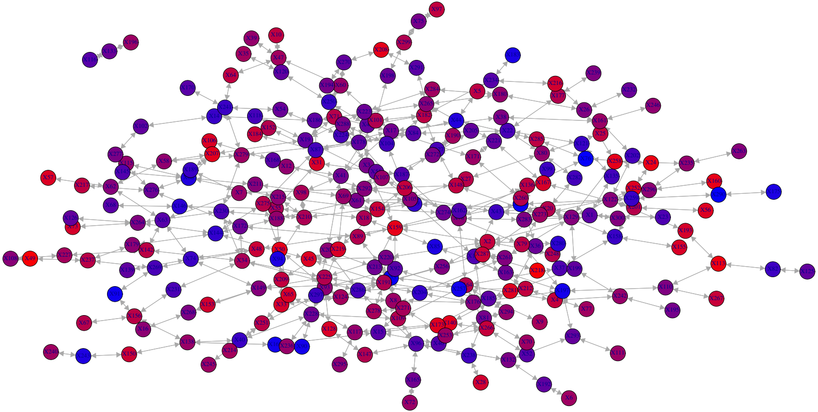


Figure : Class A member

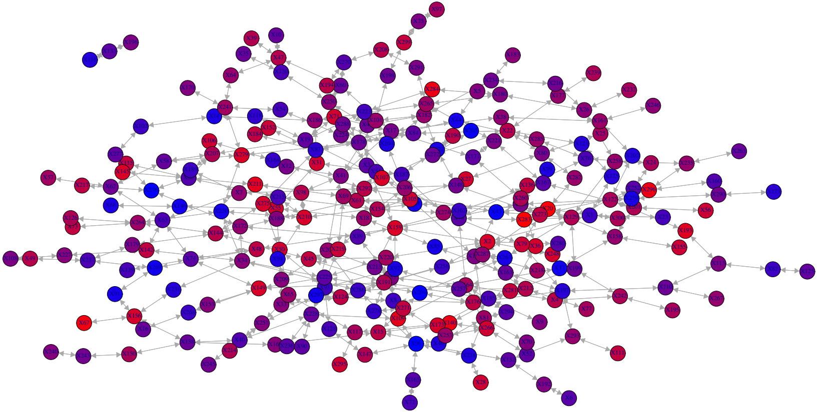


Figure : Class B member

Finished!

Thank You!
Questions?