

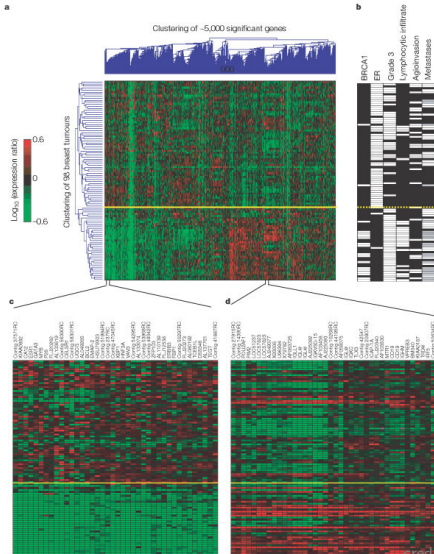
PPI Networks and Gene Expression

Adrin Jalali

July 6, 2013

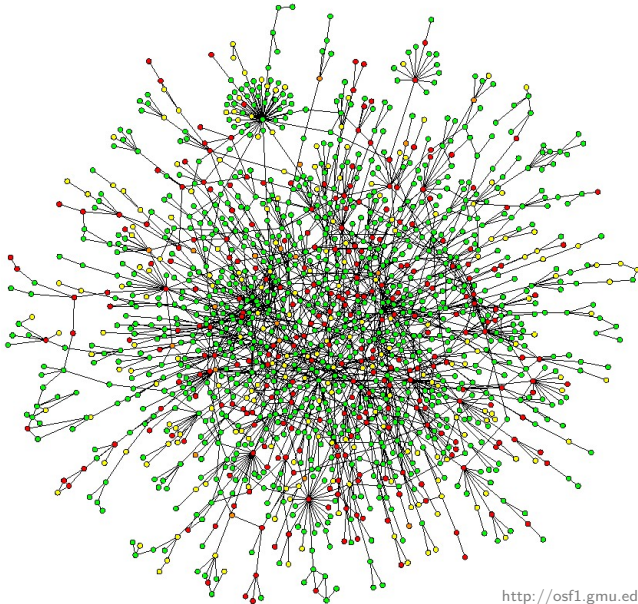


Van't Veer breast-cancer data



credit: 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
© Mary Ann Liebert, Inc.
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 SHAMIR¹

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 (Sæfkle 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)

¹Iliriyah School of Computer Science, Tel Aviv University, Tel Aviv, Israel.

²Yahoo! Research, Haifa, Israel.

³IBM 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.:

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

3. Dual to Primal

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

2. Dual problem

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

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

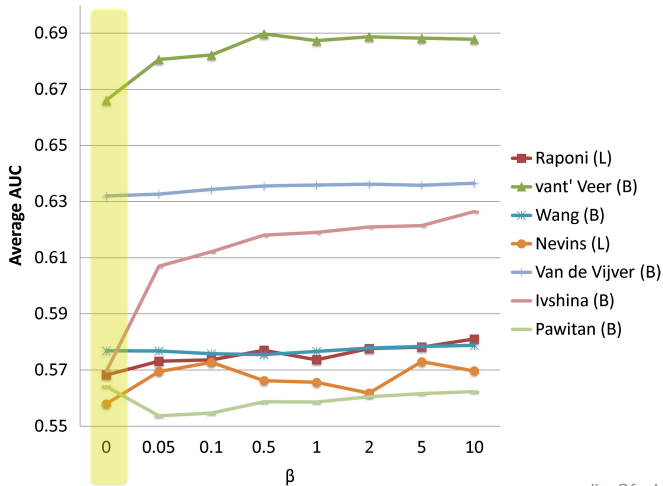
s.t.:

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

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

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

NICK Performance Summary



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

Synthesize data

- 1 A random graph
- 2 Signal nodes:

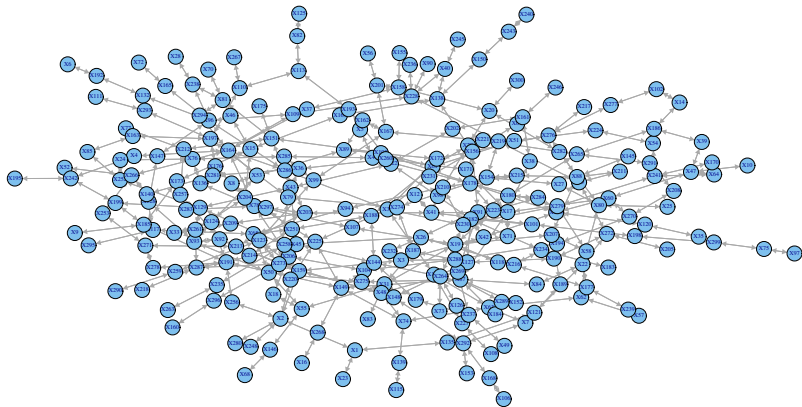
$$f(n) = \begin{cases} N(-\mu, 1) & \text{if } n \text{ is in class 1} \\ N(\mu, 1) & \text{if } n \text{ is in class 2} \end{cases}$$

- 3 Random nodes:

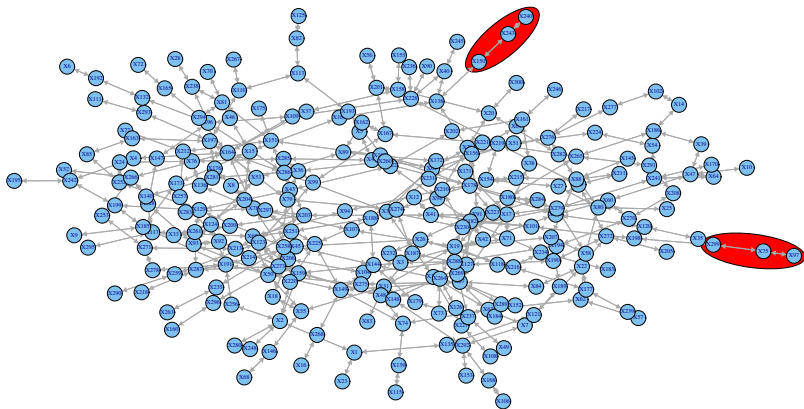
$$f(n) = N(0, 1)$$

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

Synthesized data



Synthesized data easy scenario



Results

- 1 Extract pairs of genes with mutual absolute large w
- 2 Synthesized easy: all implanted pathways come on top of the list
- 3 Synthesized hard: they are vanished

Results

- ① Extract pairs of genes with mutual absolute large w
- ② Synthesized easy: all implanted pathways come on top of the list
- ③ Synthesized hard: they are vanished
- ④ Van't veer:
 - ① Slightly better performance, although not necessarily as reported.
 - ② You find even better genes in w/o network scenario.
 - ③ Well known genes are of very high degree in the network.

Idea

- 1 Estimate density distribution of each gene for class A.

Idea

- ① Estimate density distribution of each gene for class A.
- ② For each sample:
 - ① Extract abnormal genes according to above estimated distributions.
 - ② Extract the part of PPI network induced by extracted genes (almost)

Idea

- ① Estimate density distribution of each gene for class A.
- ② For each sample:
 - 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 density distribution of each gene for class A.
- ② For each sample:
 - 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.

Thank You!