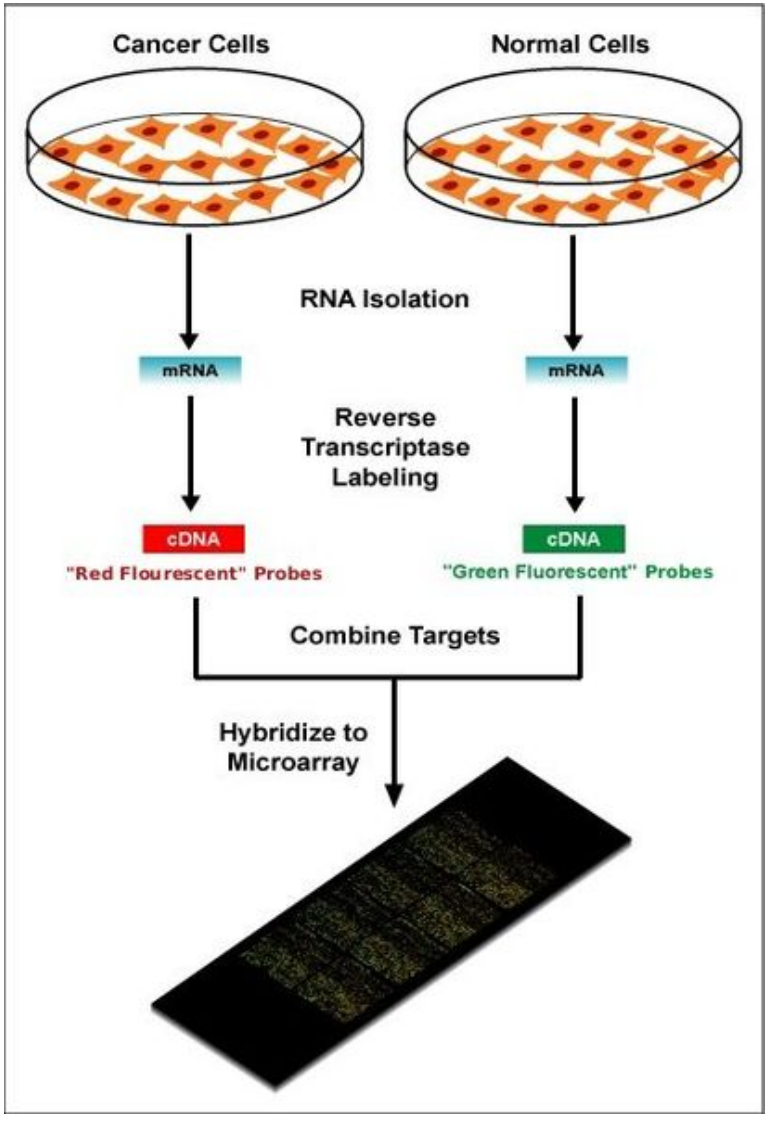
PPI Networks and Gene Expression

Adrin Jalali

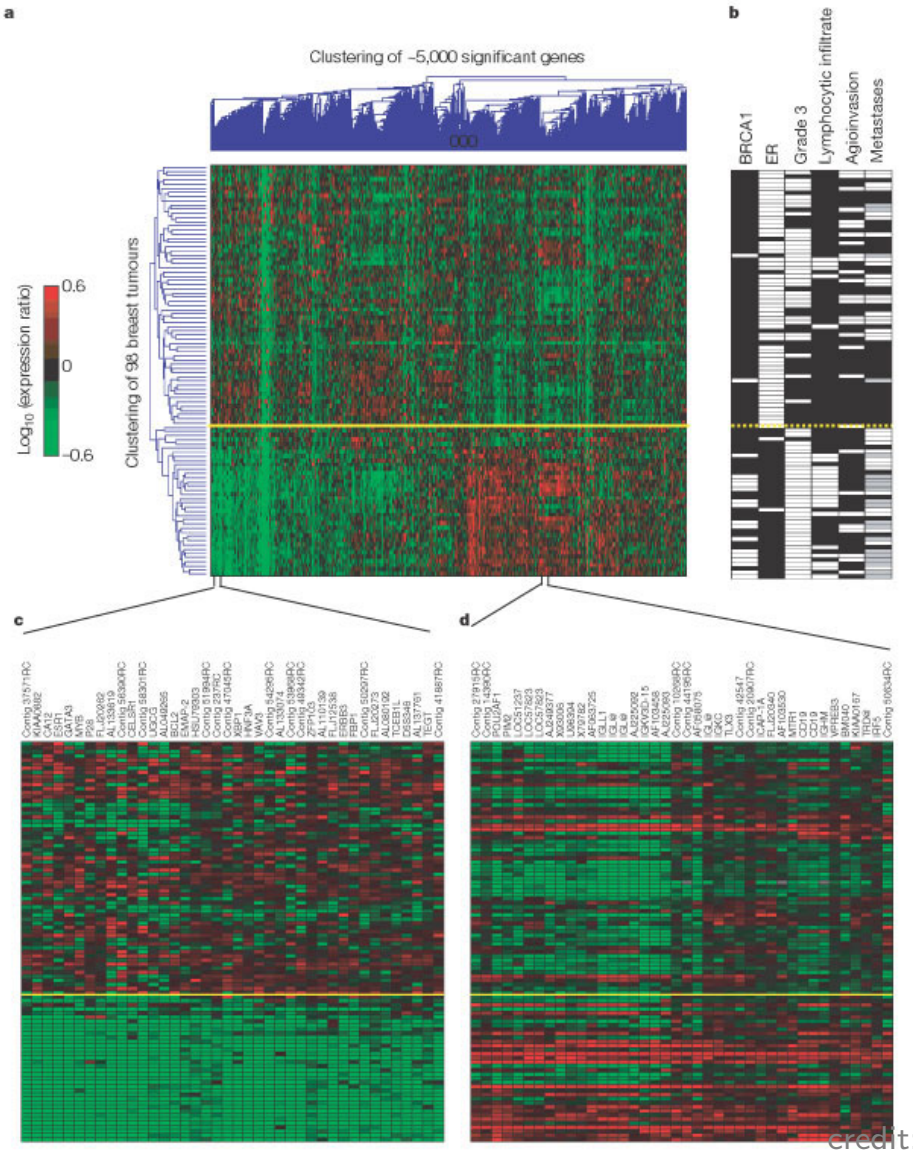
July 7, 2013

Microarray Gene Expression



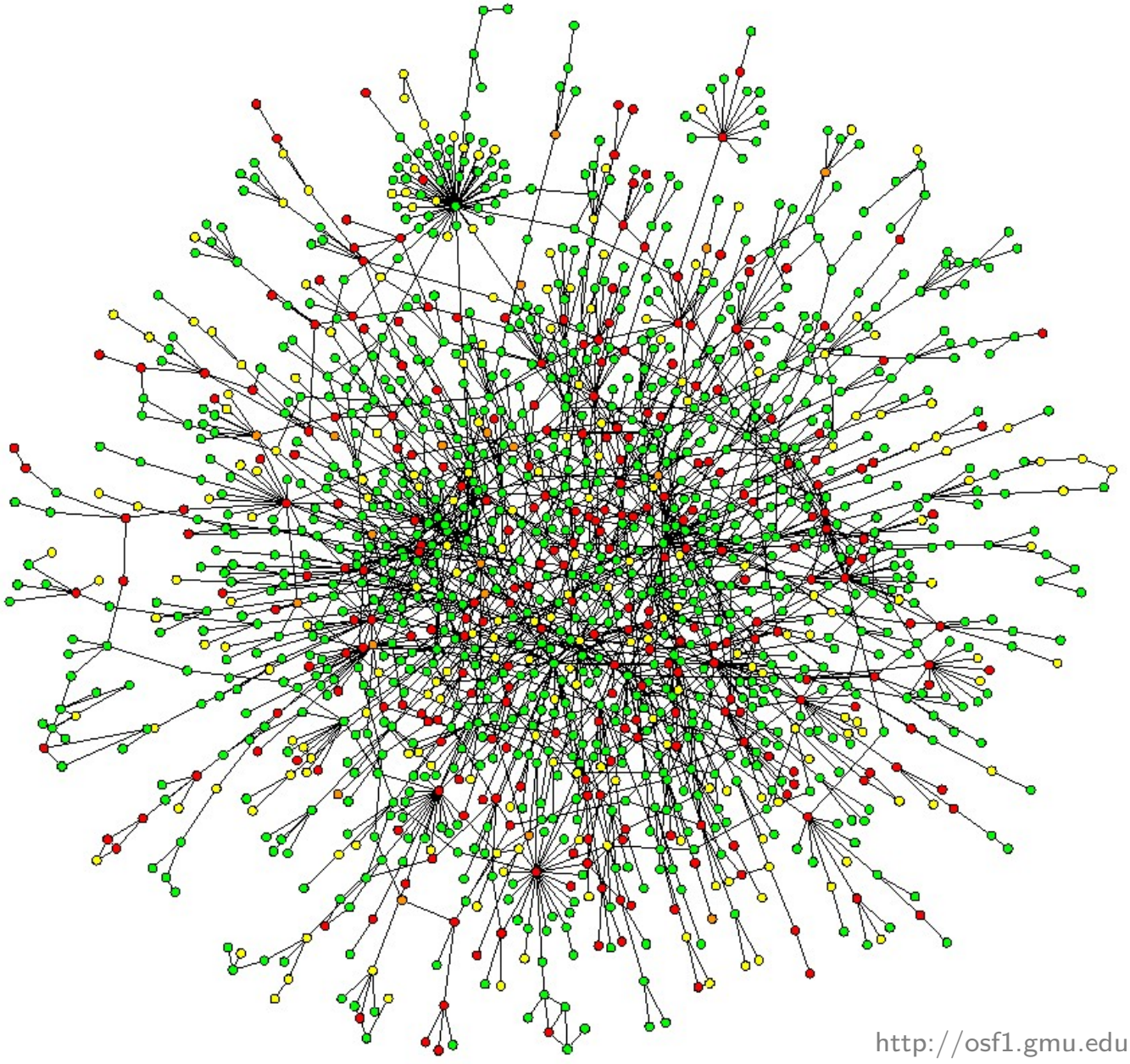
credit: en.wikipedia.org

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

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.:

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

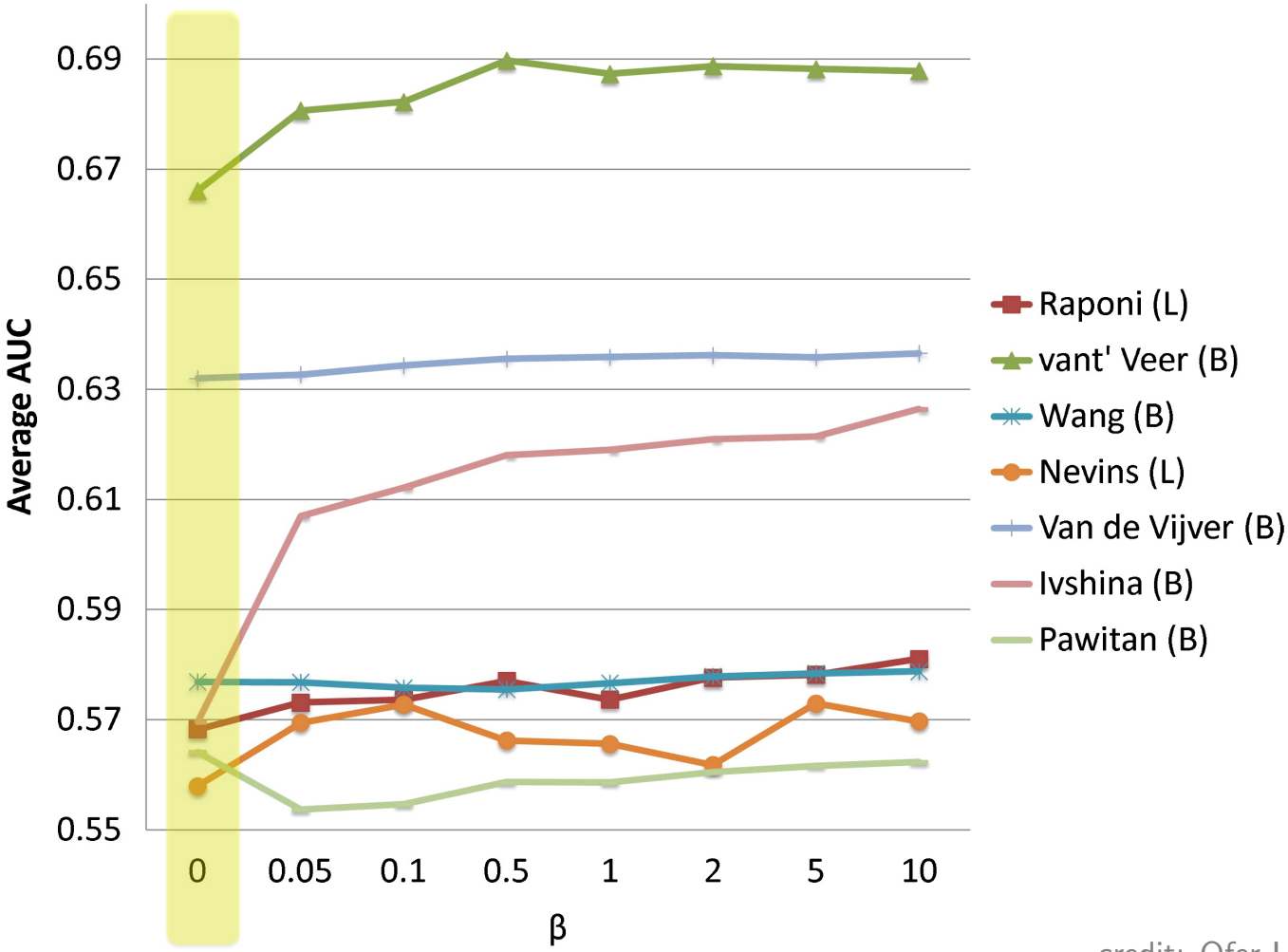
$$\forall i \in \{1, \cdots, n\} : \alpha_i \geq 0$$

3. Dual to Primal

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

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:

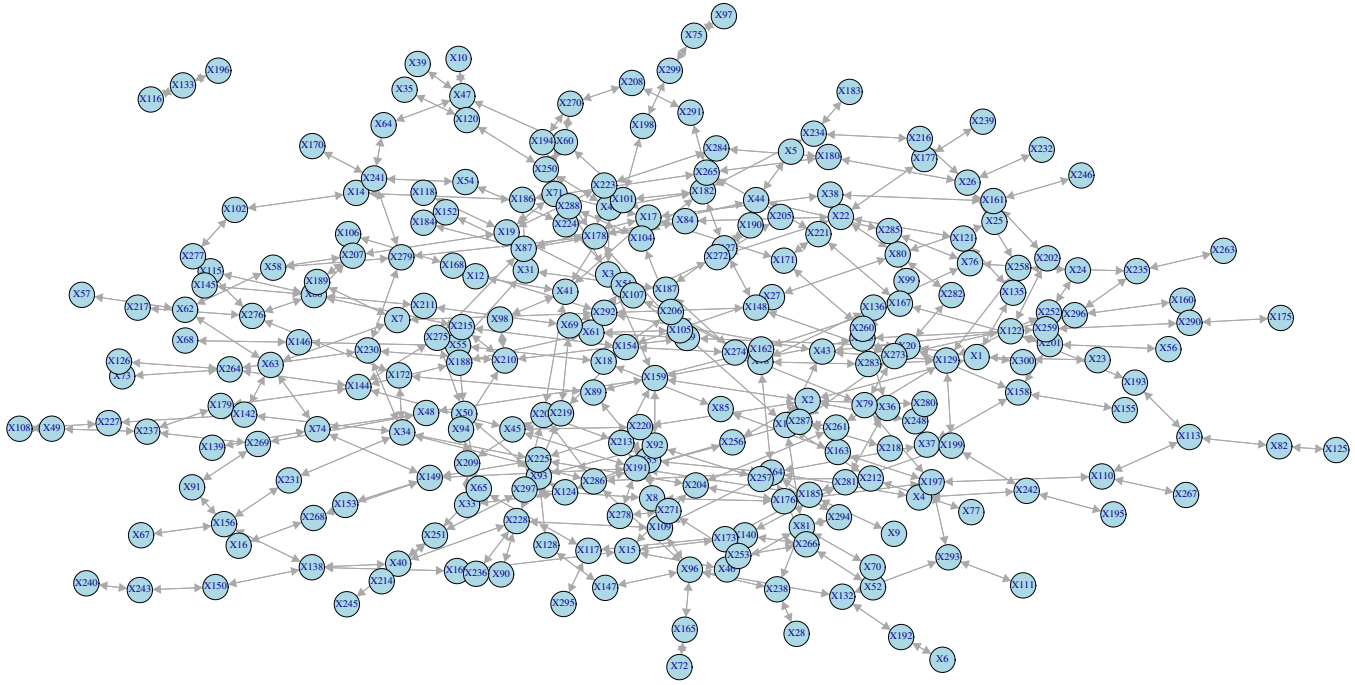
$$f(n) = \begin{cases} N(-\mu, 1) & \text{if } n \text{ is in class 1} \\ N(\mu, 1) & \text{if } n \text{ is in class 2} \end{cases}$$

3. Random nodes:

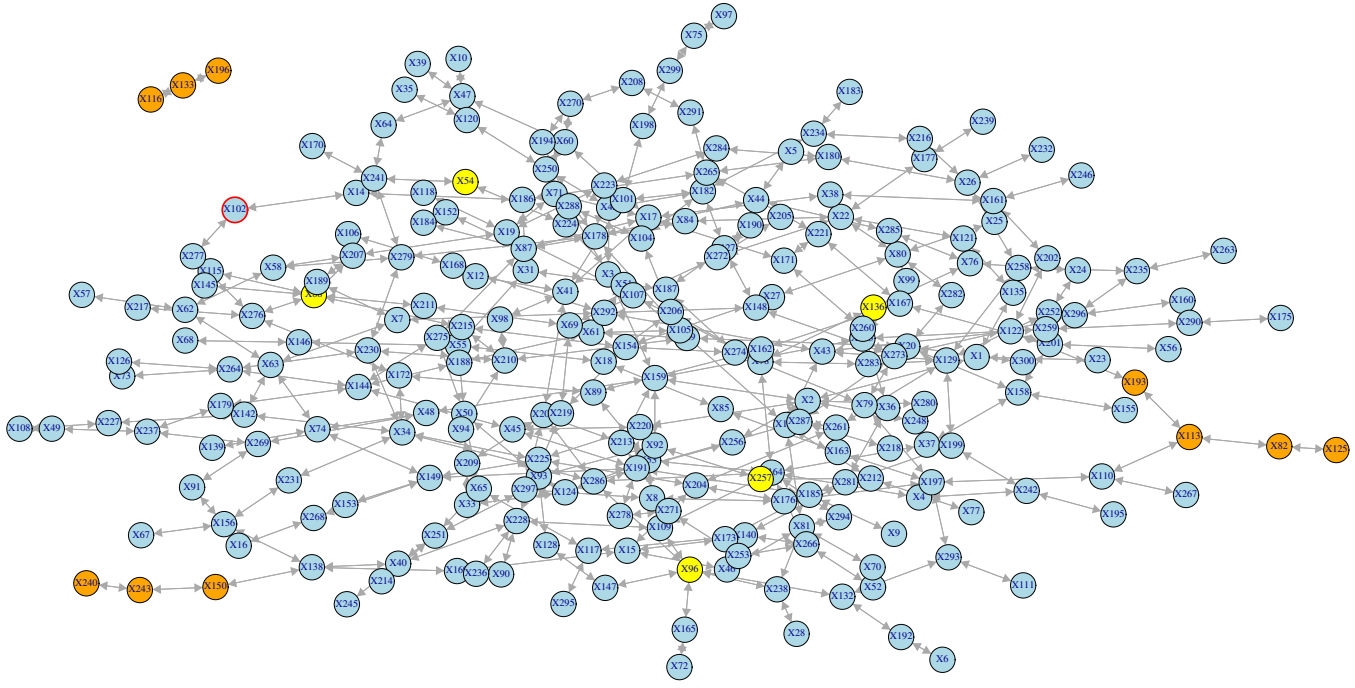
$$f(n) = N(0, 1)$$

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

Synthesized data

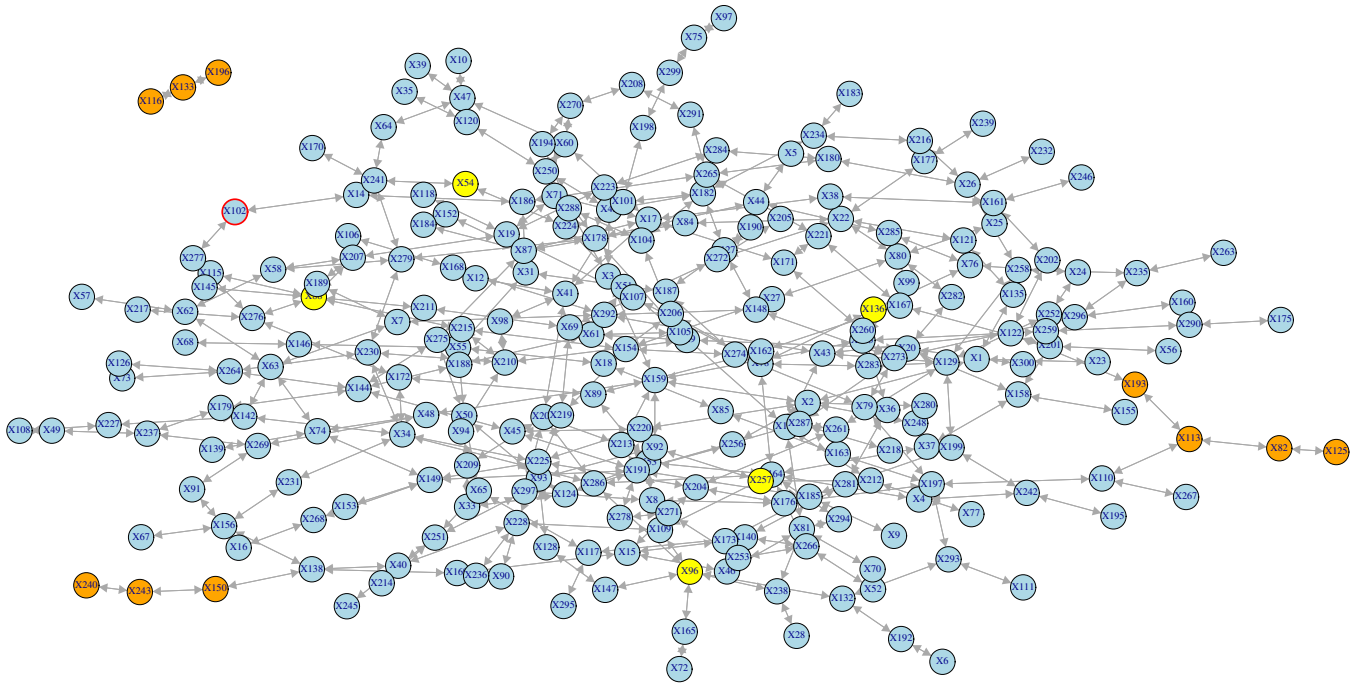


Synthesized data easy scenario



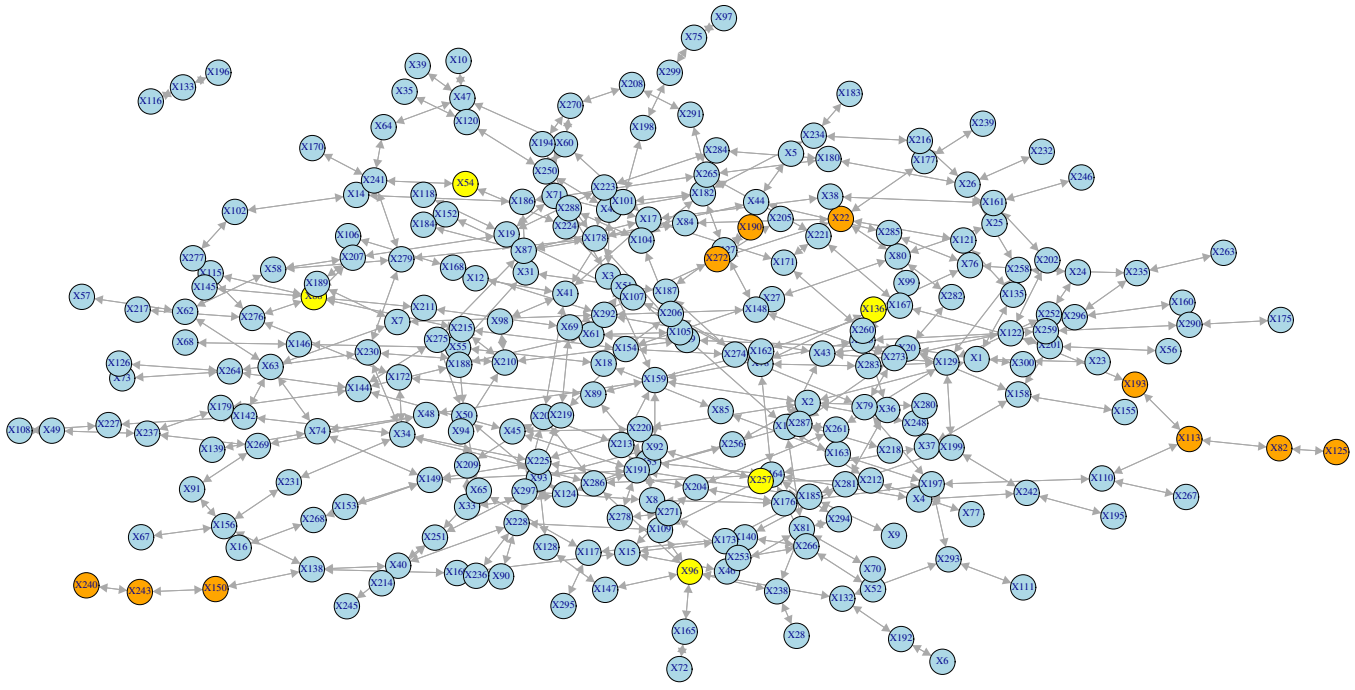
Synthesized data easy scenario

X88	X88
X96	X96
X116	X116
X150	X150
X116	X133
X82	X82
X196	X196
X136	X136
X82	X125
X196	X133
X82	X113
X125	X125
X133	X133
X150	X243
X54	X54
X113	X113
X240	X240
X257	X257
X240	X243
X113	X193
X243	X243
X193	X193
X88	X115
X150	X138
X88	X215
X96	X147
X88	X207
X136	X76
X96	X165
X88	X145



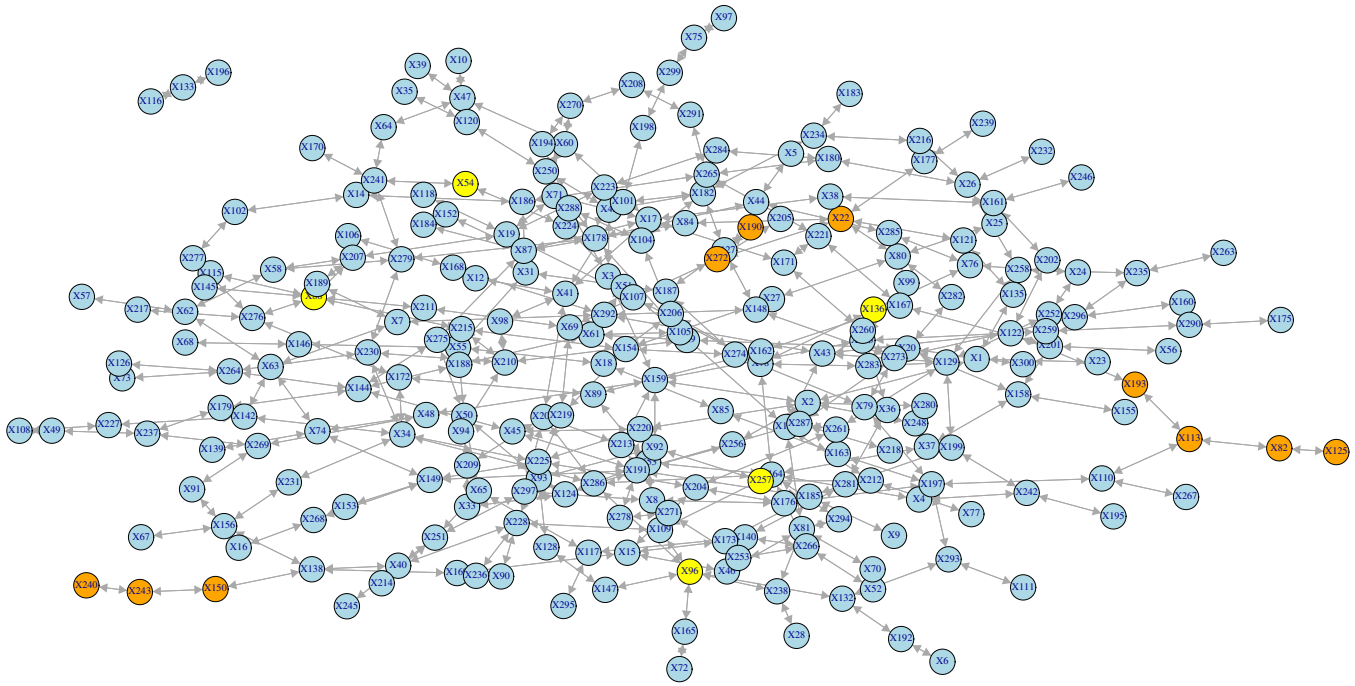
X196	X196
X196	X133
X133	X133
X133	X116
X116	X116
X240	X240
X196	X116
X125	X125
X240	X243
X125	X82
X243	X243
X82	X82
X243	X150
X82	X113
X150	X150
X113	X113
X113	X193
X193	X193
X113	X110
X150	X138
X193	X122
X240	X150
X125	X113
X233	X233
X110	X110
X110	X267
X267	X267
X82	X193
X122	X122
X95	X95

Synthesized data medium scenario



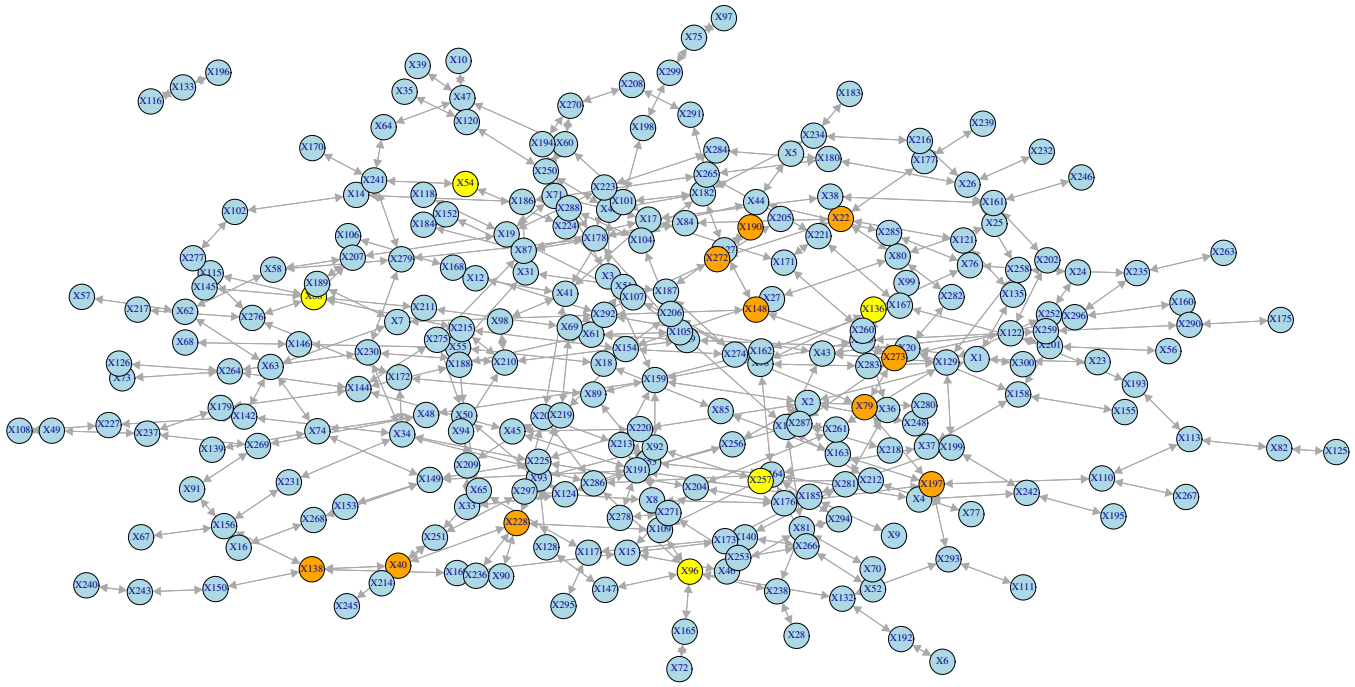
Synthesized data medium scenario

X88	X88
X96	X96
X22	X22
X190	X190
X150	X150
X136	X136
X82	X82
X22	X272
X190	X272
X82	X113
X82	X125
X113	X113
X54	X54
X125	X125
X150	X243
X272	X272
X113	X193
X240	X240
X240	X243
X257	X257
X243	X243
X193	X193
X96	X109
X96	X165
X136	X76
X88	X276
X88	X211
X150	X138
X88	X207
X88	X115



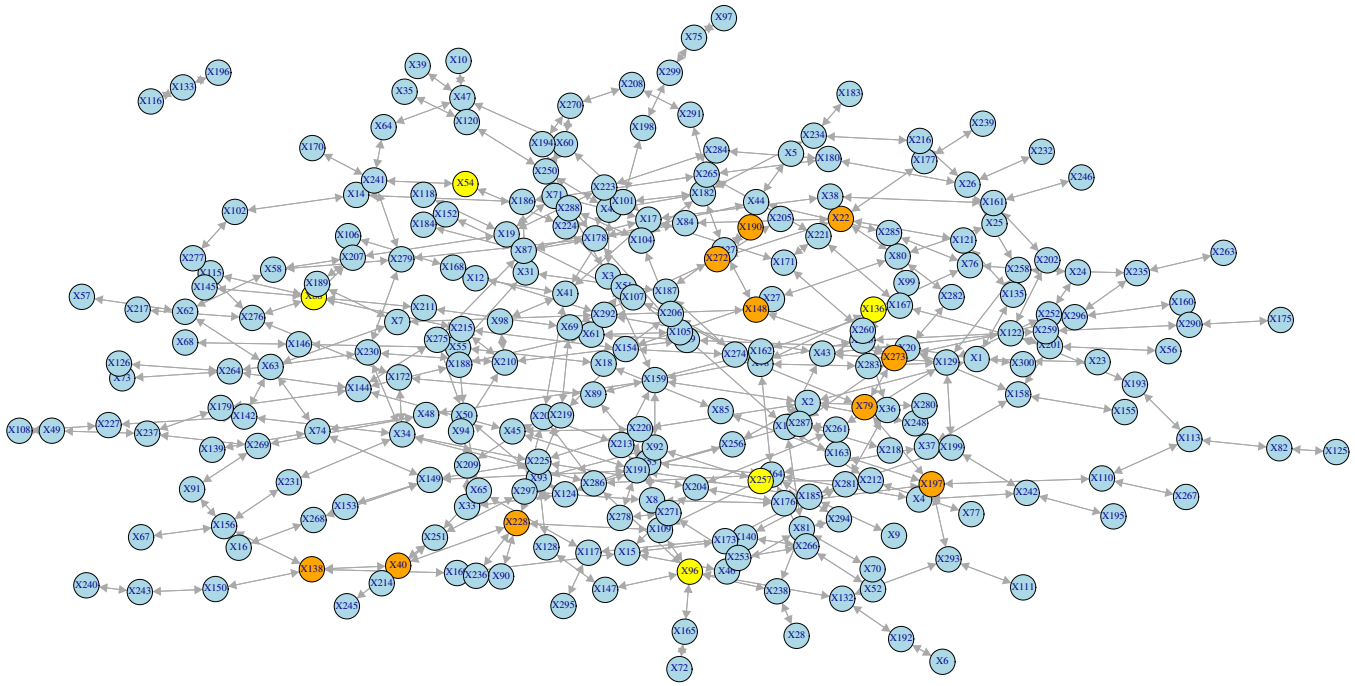
X240	X240
X125	X125
X240	X243
X125	X82
X243	X243
X82	X82
X243	X150
X82	X113
X150	X150
X113	X113
X113	X193
X193	X193
X113	X110
X150	X138
X193	X122
X240	X150
X110	X110
X125	X113
X110	X267
X190	X190
X267	X267
X233	X233
X190	X272
X272	X272
X82	X193
X272	X205
X138	X138
X122	X122
X205	X205
X82	X110

Synthesized data hard scenario



Synthesized data hard scenario

X88	X88
X138	X138
X190	X190
X96	X96
X138	X40
X79	X79
X40	X40
X22	X22
X136	X136
X40	X228
X190	X272
X79	X273
X79	X197
X54	X54
X22	X272
X148	X148
X228	X228
X148	X273
X273	X273
X197	X197
X272	X272
X257	X257
X88	X115
X96	X109
X88	X211
X88	X207
X138	X150
X88	X145
X88	X276
X136	X76



X190	X190
X190	X272
X272	X272
X272	X205
X205	X205
X233	X233
X190	X127
X272	X69
X272	X22
X40	X40
X40	X228
X228	X228
X228	X90
X90	X90
X40	X245
X228	X236
X40	X138
X69	X69
X228	X219
X112	X112
X148	X148
X245	X245
X236	X236
X69	X219
X138	X138
X257	X257
X148	X127
X69	X211
X219	X219
X54	X54

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.

Why it doesn't work

Weighted idea

Finished!

Thank You!