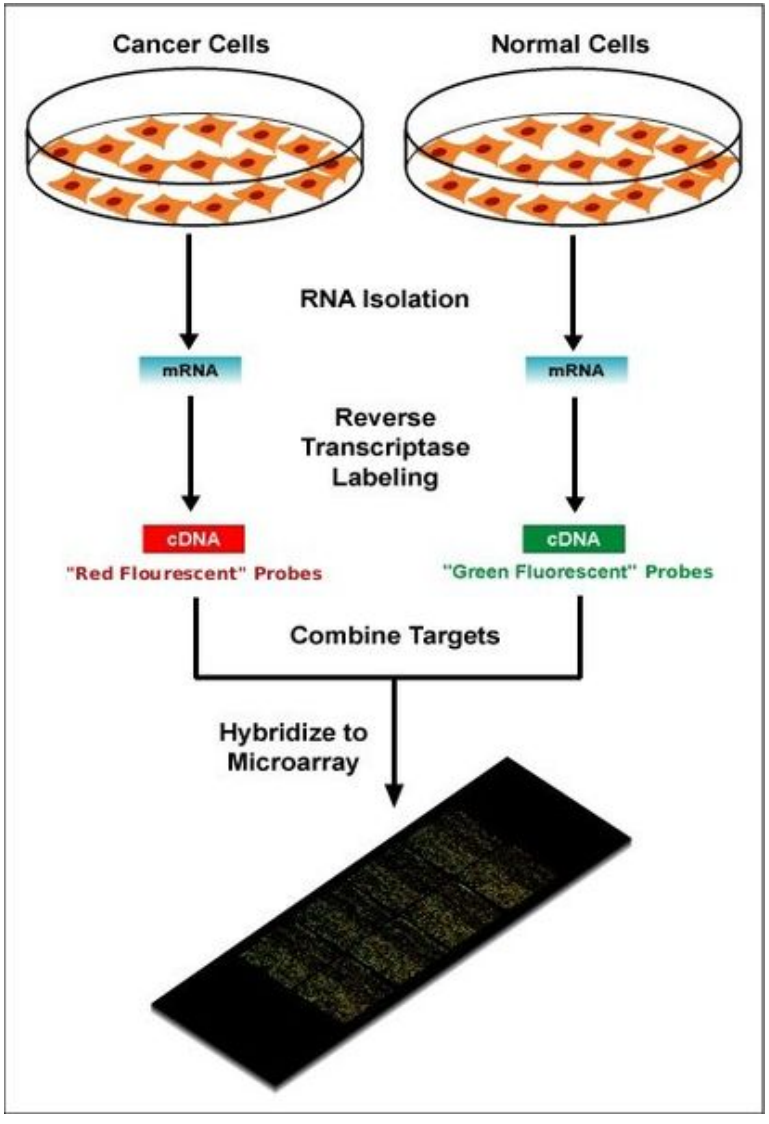
PPI Networks and Gene Expression

Adrin Jalali

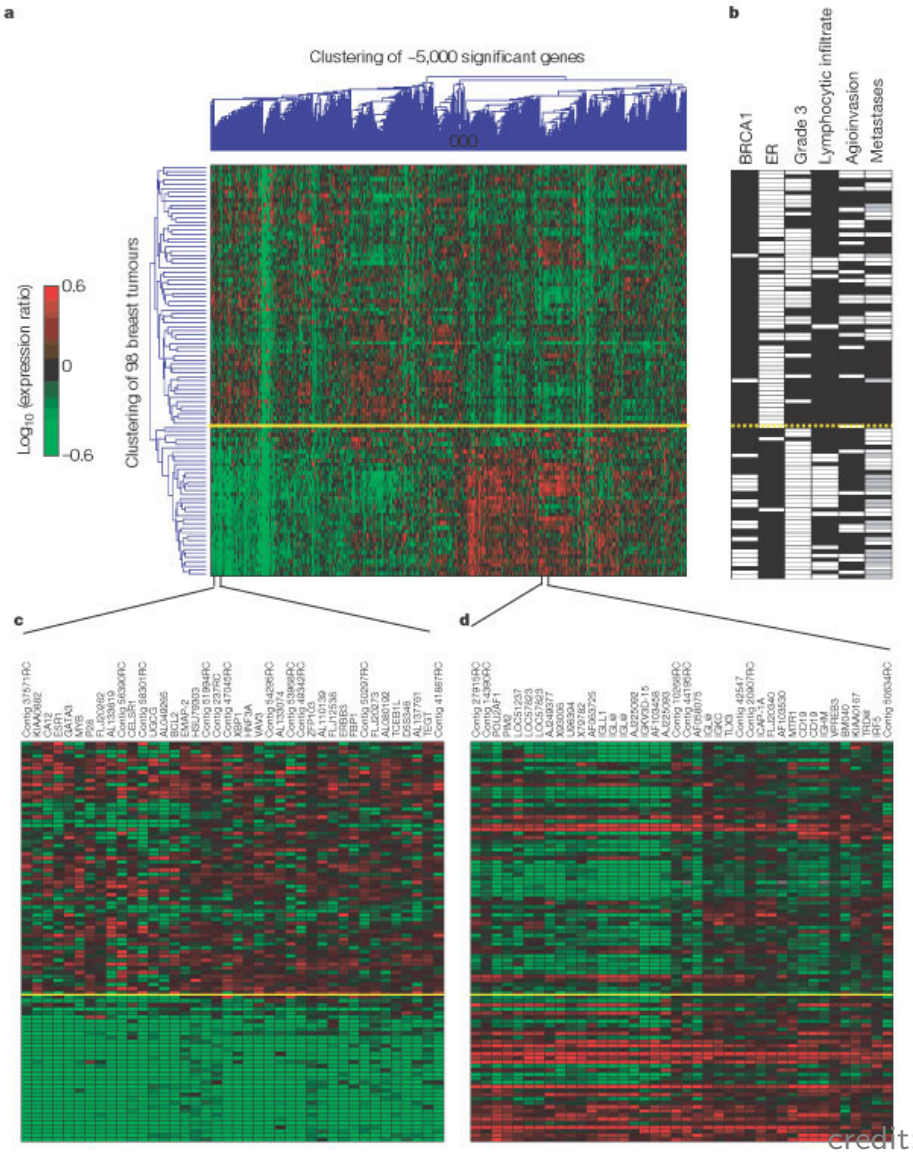
July 7, 2013

Microarray Gene Expression



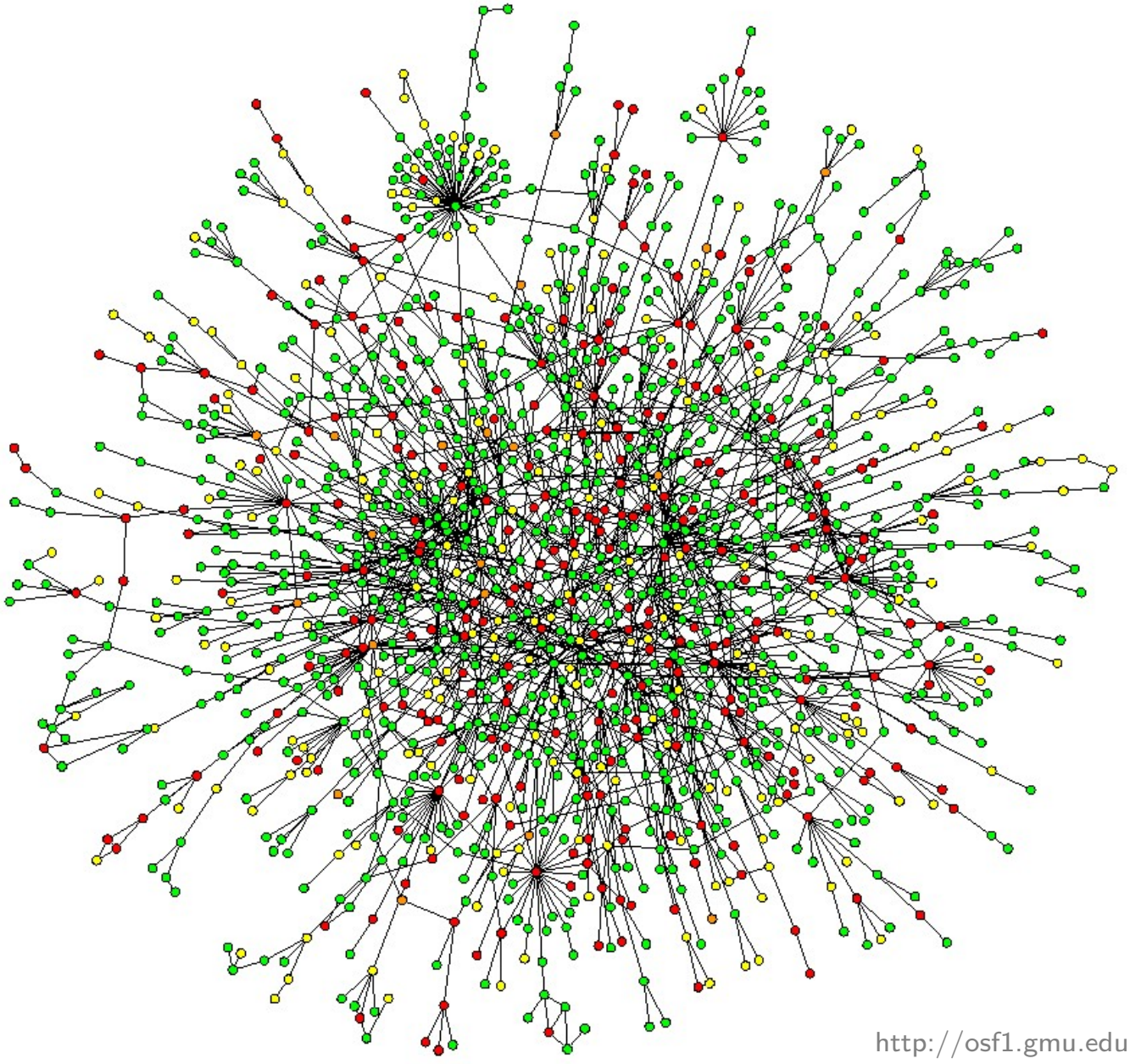
credit: en.wikipedia.org

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.:

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

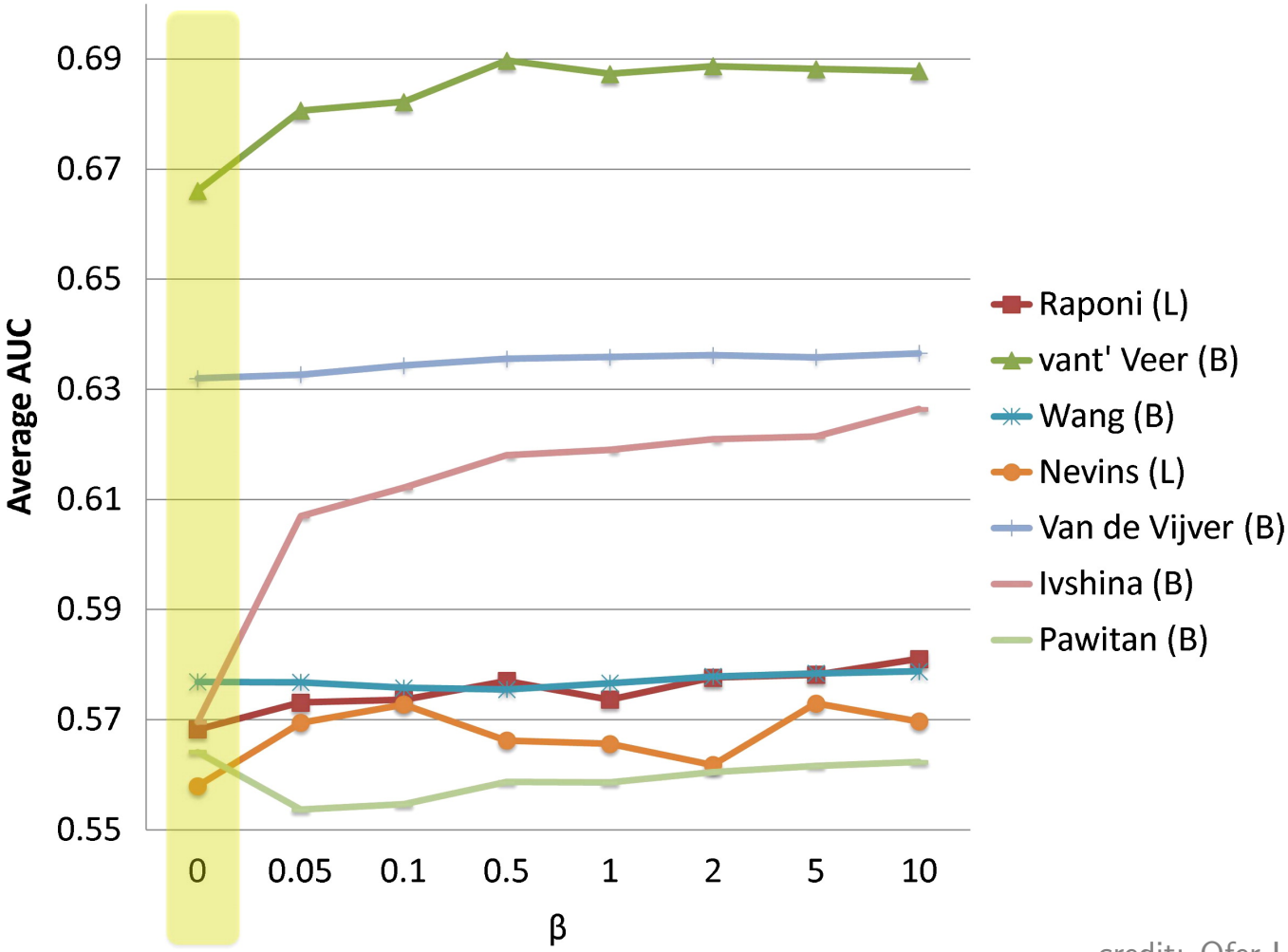
$$\forall i \in \{1, \cdots, n\} : \alpha_i \geq 0$$

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:

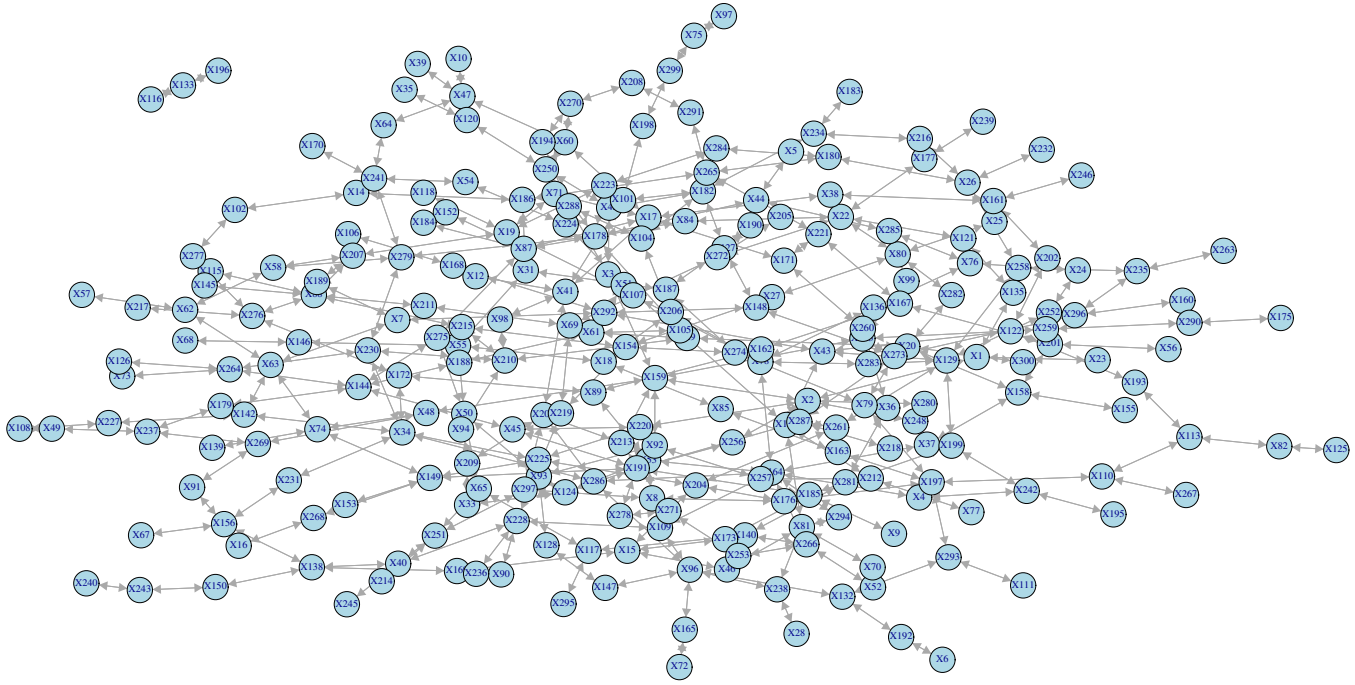
$$f(n) = \begin{cases} N(-\mu, 1) & \text{if } n \text{ is in class 1} \\ N(\mu, 1) & \text{if } n \text{ is in class 2} \end{cases}$$

3. Random nodes:

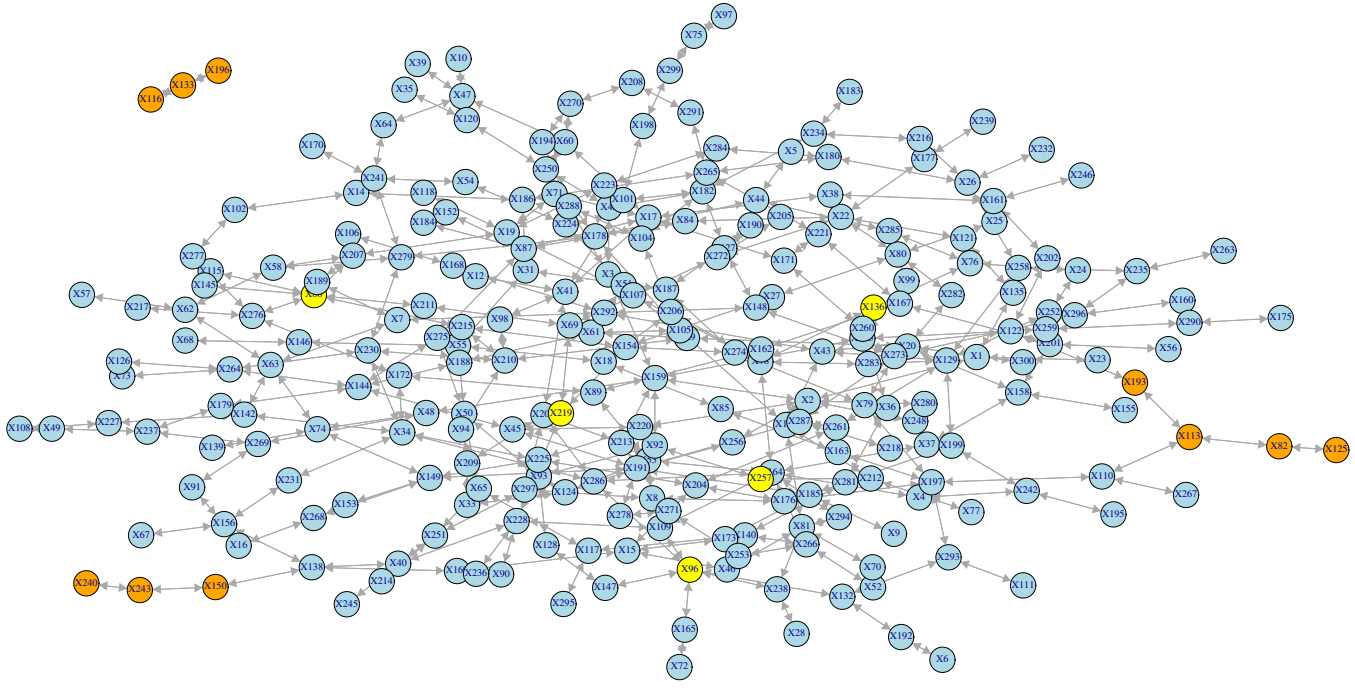
$$f(n) = N(0, 1)$$

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

Synthesized data

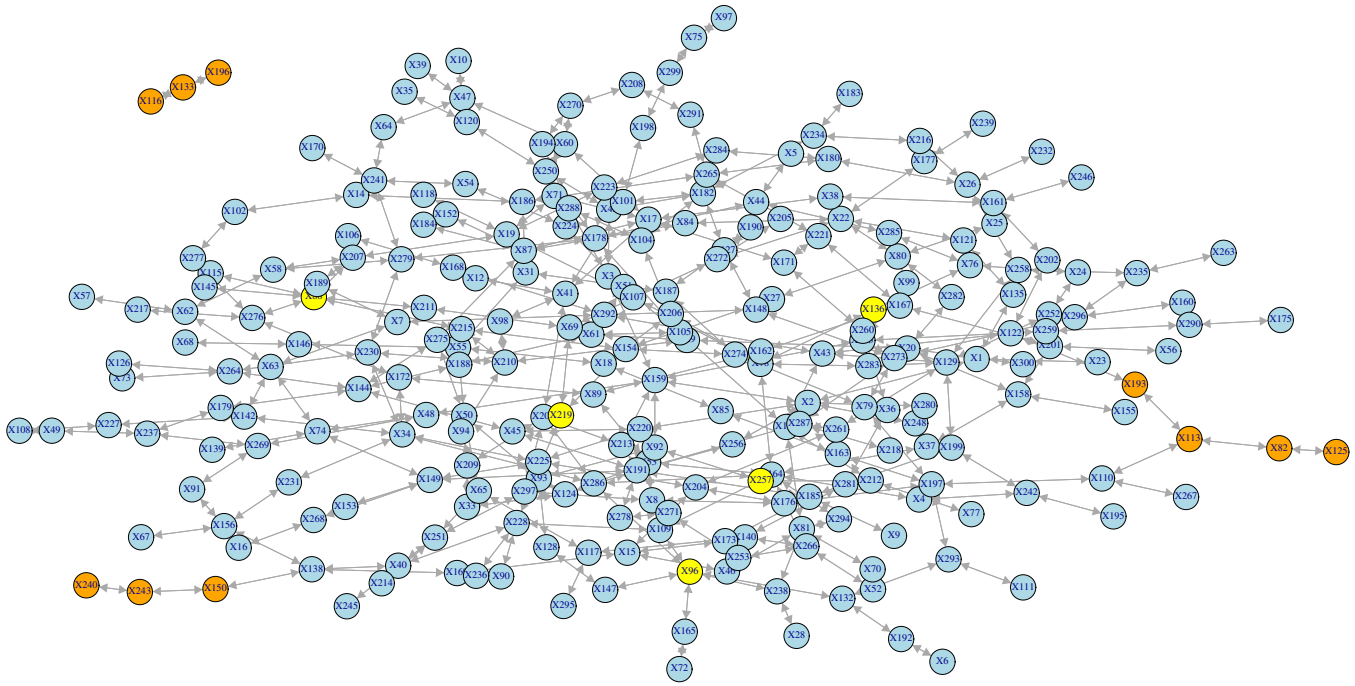


Synthesized data easy scenario



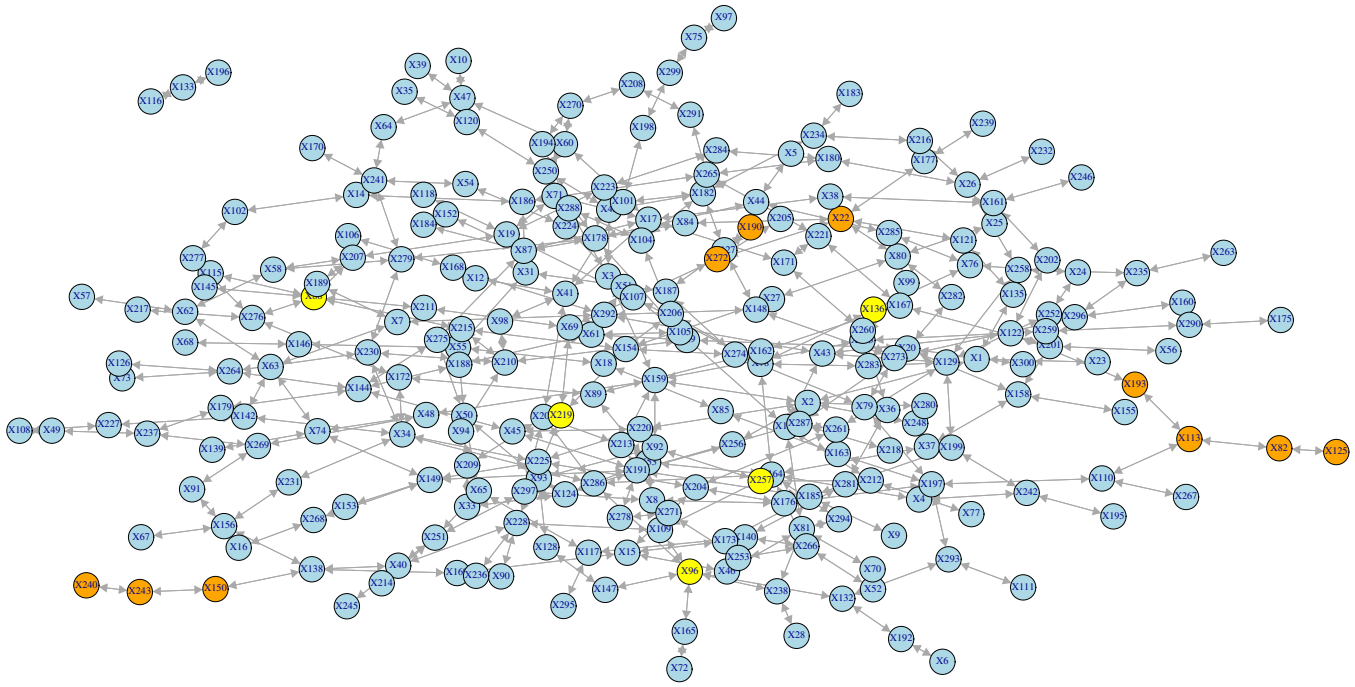
Synthesized data easy scenario

X116	X116
X125	X125
X116	X133
X125	X82
X240	X240
X196	X196
X196	X133
X133	X133
X82	X82
X240	X243
X82	X113
X6	X6
X257	X257
X6	X192
X257	X212
X243	X243
X212	X212
X192	X192
X257	X92
X72	X72
X172	X172
X72	X165
X212	X77
X57	X57
X165	X165
X192	X132
X92	X92
X118	X118
X243	X150
X77	X77



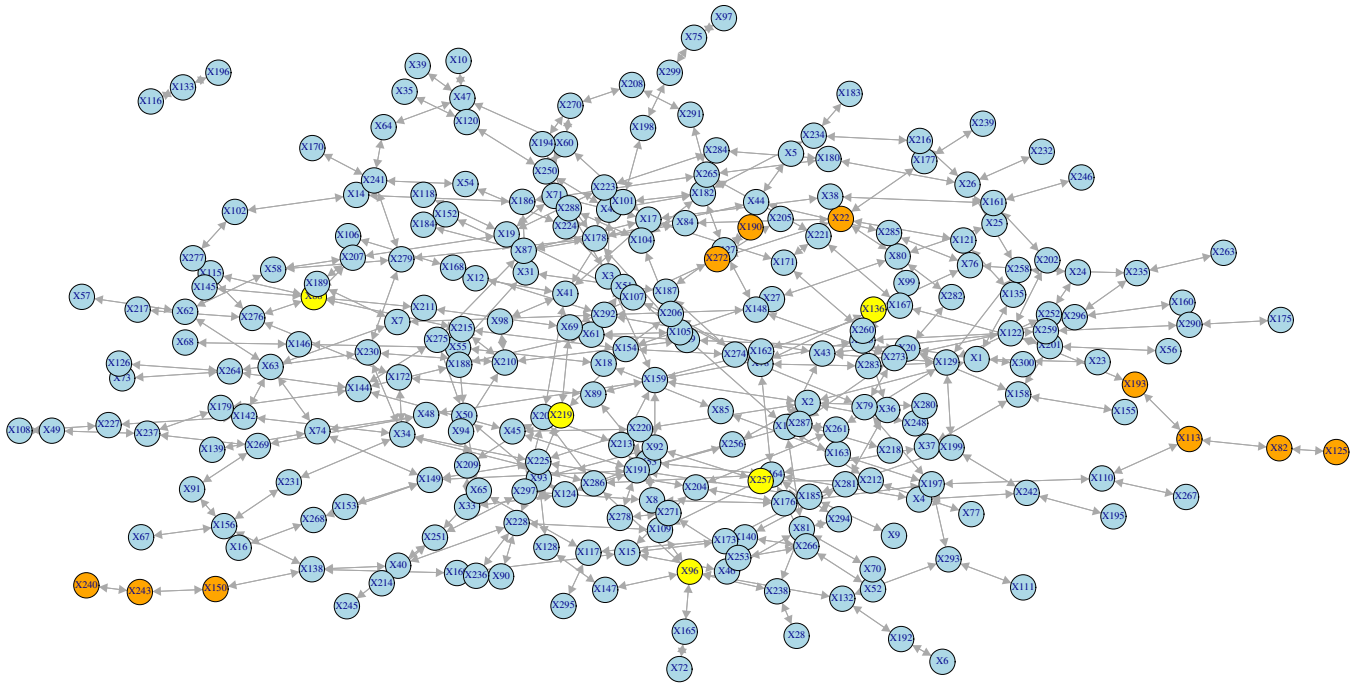
X196	X196
X196	X133
X133	X133
X133	X116
X116	X116
X240	X240
X125	X125
X240	X243
X125	X82
X243	X243
X196	X116
X82	X82
X243	X150
X82	X113
X150	X150
X157	X157
X6	X6
X113	X113
X6	X192
X95	X95
X192	X192
X113	X110
X72	X72
X267	X267
X249	X249
X72	X165
X192	X132
X267	X110
X150	X138
X110	X110

Synthesized data medium scenario



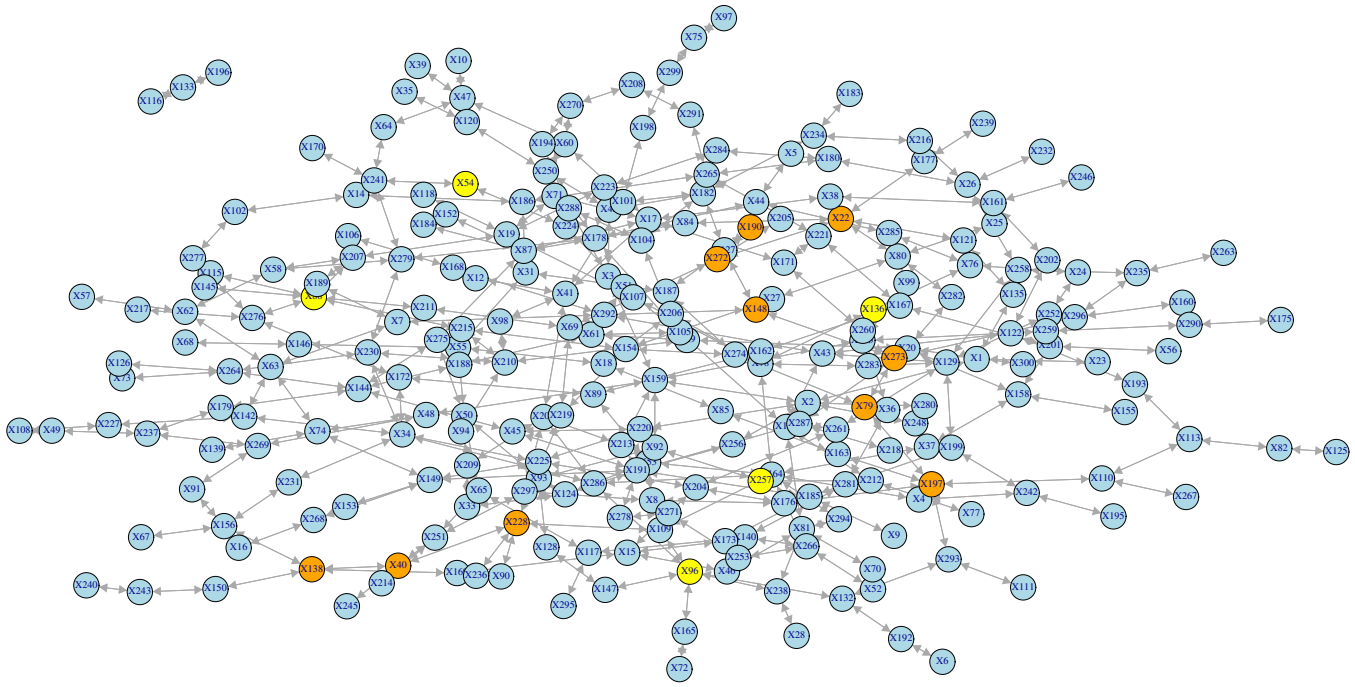
Synthesized data medium scenario

X125	X125
X125	X82
X267	X267
X69	X69
X69	X219
X267	X110
X240	X240
X69	X272
X219	X219
X82	X82
X219	X154
X82	X113
X154	X154
X110	X110
X240	X243
X110	X113
X113	X113
X272	X272
X272	X205
X205	X205
X243	X243
X10	X10
X176	X176
X154	X274
X77	X77
X33	X33
X85	X85
X110	X197
X69	X211
X154	X27



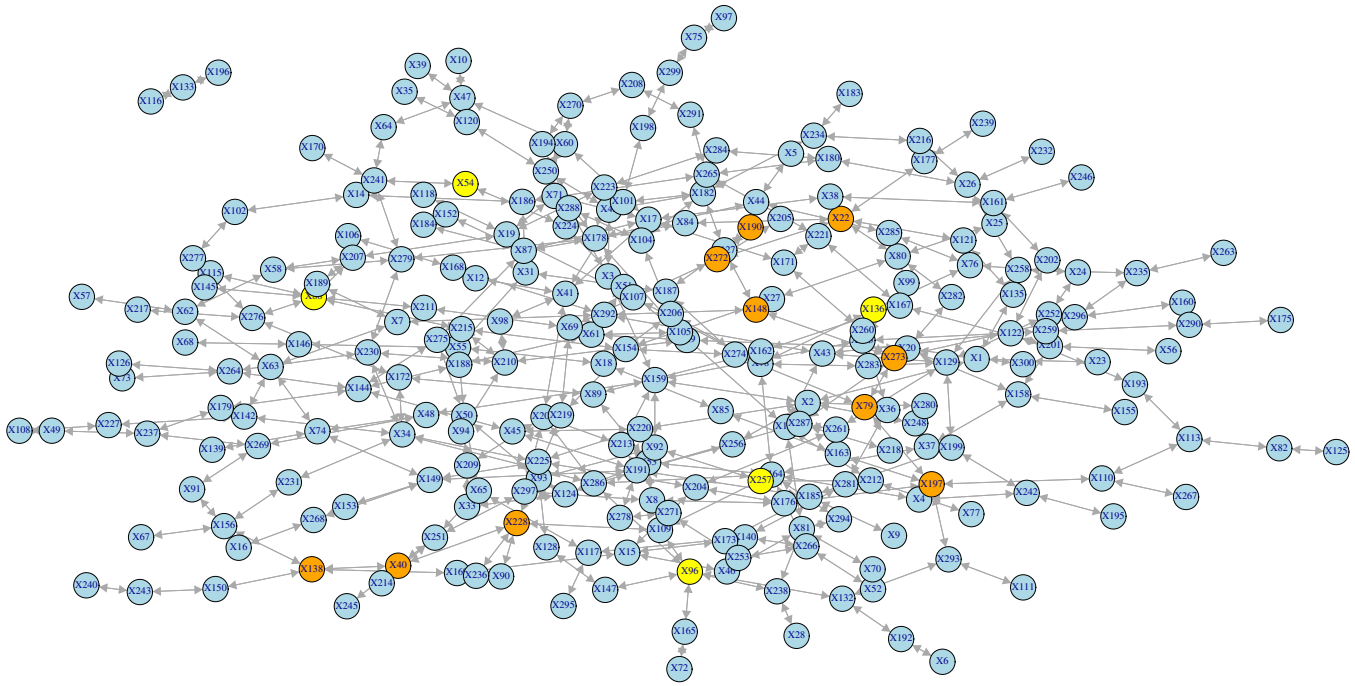
X125	X125
X125	X82
X240	X240
X82	X82
X240	X243
X267	X267
X82	X113
X243	X243
X267	X110
X113	X113
X113	X110
X110	X110
X243	X150
X110	X197
X150	X150
X6	X6
X111	X111
X113	X193
X6	X192
X111	X293
X192	X192
X293	X293
X192	X132
X293	X197
X205	X205
X293	X132
X175	X175
X197	X197
X132	X132
X175	X290

Synthesized data hard scenario



Synthesized data hard scenario

X42	X42
X101	X101
X106	X106
X42	X182
X101	X198
X106	X168
X12	X12
X101	X182
X198	X198
X168	X168
X42	X3
X101	X41
X190	X190
X281	X281
X98	X98
X190	X127
X42	X250
X182	X182
X182	X127
X147	X147
X127	X127
X79	X79
X12	X41
X205	X205
X198	X299
X127	X187
X168	X292
X127	X148
X187	X187
X281	X36



X97	X97
X97	X75
X75	X75
X75	X299
X299	X299
X299	X198
X106	X106
X198	X198
X106	X168
X198	X101
X205	X205
X168	X168
X205	X272
X101	X101
X190	X190
X190	X272
X90	X90
X272	X272
X101	X182
X190	X127
X236	X236
X272	X69
X12	X12
X69	X69
X168	X292
X69	X219
X101	X41
X90	X228
X236	X228
X219	X219

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!