BIOLOGICAL RELEVANCE DETECTION VIA NETWORK DYNAMIC ANALYSIS

Zheng Zhao[†] **Huan Liu**[†] **Jiangxin Wang**[‡] **Yung Chang**[‡]

[†] **Computer Science and Engineering, Arizona State University**
[‡] **School of Life Science, CIDV, The Biodesesign Institute, Arizona State University**
Email: {zhaozheng,huan.liu,jiangxin.wang,yung.chang}@asu.edu

Abstract

Most existing approaches for gene selection are based on evaluating the statistical relevance. However, there are remarkable discrepancies between statistical relevance and biological relevance. It is important to consider biological relevance for crucial genes identification. The task of detecting biological relevance presents two major challenges: first, how to define different types of measures to evaluate the biological relevance from multiple perspectives; and second, how to effectively integrate these measures to achieve better estimations. In this work, we propose to detect biological relevance by applying dynamics analysis using both biological networks and gene expression profiles from different phenotypes, and develop an effective probabilistic model to integrate various types of relevance measures in a unified form. Experimental results show the efficacy and potential of the proposed approach with promising findings.

1 INTRODUCTION

Selecting genes that are critical to a particular biological process has been a major challenge in post-array analysis. Also known as feature selection [18] in machine learning research area, gene selection has attracted intensive research interests and much progress has been made over the last decade in developing effective gene selection algorithms [17, 23]. Given cDNA Microarray data, most existing algorithms try to identify genes that are differentially expressed over the samples via various types of statistical tests. Discriminative genes help classifiers to achieve high accuracy. However, does the better accuracy necessarily indicate higher biological relevance of genes? We applied a supervised gene selection algorithm, *Information Gain* [11] and an unsupervised method, *Gene Variance*, on the expression profiles of 81 glioblastomas patients and 23 normal persons [28] to select genes that may provide insight into the pathogenesis of glioblastomas. Using the top 20 genes selected by Information Gain, the $k$nn classifier achieves an accuracy rate of 0.97, with 5 and 0 selected genes being related to cancer and glioblastoma, respectively, according to literature, respectively. And using the genes selected by gene variance, $k$nn achieves the accuracy of 0.89, while 10 and 7 selected genes are related to cancer and glioblastoma, respectively. The result shows that a gene list of higher accuracy does not necessarily contain more biologically relevant genes. A sensible explanation is that a cDNA Microarray data usually contains more than 10^4 genes but only fewer than 200 samples. A data set of this kind usually leads to the *small sample problem* [22]. With so few samples many genes that may not be biologically relevant can easily gain statistical relevance due to sheer randomness [27]. Hence, selecting genes to achieve high accuracy should not be the sole goal of biological discovery. Genes' biological relevance refers to the relevance confirmed by their involvement in the biological process of interest. For instance, a gene can gain its biological relevance due to its role of triggering cancer. How to systematically detect biological relevance in gene selection is an important problem that need to be addressed.

Robustness is a property that can be observed in most biological systems [13]. Functional robustness allows systems to dynamically adapt to environmental changes or system failures. Various mechanisms, including feedback, redundancy and modularization, have been introduced by evolution to achieve system robustness. However, "systems are evolved to be robust against general perturbations can be extremely fragile against certain types of rare perturbations" [13]. For instance, biological networks usually have scale-free structure [5]. The removal of the hub genes from these networks may cause dramatic topological changes and result in disasters. The defects of a few cancer-related genes can reduce the robustness of the system and trigger cancers [15], which indicates their strong ability to cause fragility. These observations inspire us to detect biological relevance by applying dynamic analysis using both both biological networks

and gene expression profiles of different phenotypes. The major contributions of the paper are: (1) defining three effective measures based on network dynamic analysis to evaluate biological relevance from different angles; (2) developing a novel probabilistic model to integrate different relevance measures to achieve comprehensive evaluation.

2 RELEVANCE DETECTION

We propose to apply dynamic analysis using both biological networks and gene expression profiles to identify important biologically relevant genes. Network dynamic analysis is composed of two distinct lines of research: "the dynamics on networks", and "the dynamics of networks" [10]. In the former, the topology of the network remains static, and each node of the network represents a dynamical system and can change dynamically; in the latter, the topology of a network itself is regarded as a dynamical system. We deal with both types of dynamics in this work. Dynamic analysis provides powerful tools for studying the robustness of networks [1]. We propose three measures to evaluate genes' biological relevance: (1) the evidence of a gene's capability to influence other genes in terms of spreading abnormal expression pattern, which corresponds to the analysis of "the dynamics on networks"; (2) the capability of a gene to cause fragility of the network, which corresponds to the analysis of "the dynamics of networks"; and (3) the abnormity of genes' expression in tumor tissues, which serves as the precondition for a gene to be biologically relevant.

2.1 The Influence Measure

Feedback and redundancy are widely involved in biological networks to isolate the adverse effects of potential individual gene perturbations to ensure system robustness [13]. Earlier study shows that surprisingly few genes (typically 10∼20%) can affect the viability of organisms when knocked off from the genome [15]. However, to cause instability and trigger cancers, genes must take effect by influencing other genes in their downstream biological processes. The fact suggests that to ascertain the relevance of a gene, abnormal expression patterns should be observed on the gene, as well as the genes in its neighborhood. Assume $\mathbf{E}$ is the expression of genes and $\mathbf{G}$ is the given network. Let g_i denote the ith gene with $E(g_i)$ being its expression, let $p(g_i, g_j|\mathbf{G})$ be the probability of reaching gene g_j from gene g_i on the network $\mathbf{G}$, and $s_{abn}(E(g_i), \mathbf{c})$ be the score for measuring the abnormity of the expression pattern of g_j (will be discussed in Section 2.3). Influence of a gene in terms of spreading abnormity can be formulated as:

$$s_{infl}(g_i, \mathbf{E}, \mathbf{G}) = \sum_{j=1, j\neq i}^{N} p(g_i, g_j|\mathbf{G})\, s_{abn}(E(g_i), \mathbf{c}) \tag{1}$$

In the equation, N is the number of genes. With the network $\mathbf{G}$, a simple, yet effective way to estimate $p(g_i, g_j|\mathbf{G})$ can be formulated as:

$$p(g_i, g_j|\mathbf{G}) = e^{-\frac{d^2(g_i, g_j)}{\sigma^2}} \tag{2}$$

Here, $d(g_i, g_j)$ denotes the length of the shortest path between g_i and g_j on $\mathbf{G}$[1] . In the equation, the negative exponential function emphasizes the neighborhood of g_i, and the size of the neighborhood is determined by σ. The modularization of biological network suggests that genes' influence should be mainly extracted from their neighbors.

2.2 The Fragility Measure

The degree distributions of most biological networks follow power law distribution [13, 5]. It is well known that scale-free networks are very resilient to random nodes failure, however, are extreme fragile on failures of the hub nodes. It is theoretically shown that the capability of a gene to cause the fragility of a network is positively correlated to its contribution to the network entropy, which measures the centrality of nodes in the network [20]. Based on the observation, in this work, we propose to use centrality to estimate the capability of a node to cause fragility as follows:

$$s_{frg}(g_i, \mathbf{G}) = centrality(g_i, \mathbf{G}). \tag{3}$$

The concept of centrality has been intensively studied in graph theory and network analysis [14]. A popular centrality measure is the betweeness centrality [9], which is defined as:

$$centrality_B(v) = \sum_{s,t\neq v; s\neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{4}$$

In the equation, σ_{st} is the number of shortest paths from node s to node t, and $\sigma_{st}(v)$ is the number of shortest paths from s to t that pass through the node v[2]. Other typical centrality measures include Degree centrality, Closeness centrality, and Eigenvector centrality. We refer readers to literature for further study on this topic [14].

[1] Using Johnson's algorithm [12], for a sparse graph, the time complexity of obtaining all pairs shortest path is $O(V^2 \log V + VM)$, where V and M denote the number of nodes and edges in the graph, respectively.

[2] Using Johnson's algorithm [12], the time complexity of obtaining the centrality of all nodes is also $O(V^2 \log V + VM)$.

2.3 The Abnormity Measure

Given a set of gene expression profiles of different phenotypes, the abnormally expressed genes can be identified by comparing their expression patterns in tumor tissues with those in normal tissues. This is equivalent to traditional gene selection via detecting gene statistical relevance, where the abnormity of genes are measured by its capability for distinguishing cancerous from normal tissues. Let $E(g_i)$ denote the expression of gene g_i across all samples. Let $\mathbf{c}$ denote the class label, which is used to encode different phenotypes. The abnormity of a gene's expression can be estimated by measuring the consistency between $E(g_i)$ and $\mathbf{c}$:

$$s_{abn}(E(g_i), \mathbf{c}) = consistency(E(g_i), \mathbf{c}) \quad (5)$$

Various measures have been developed for consistency estimation, including Information Gain, Kruskal-Wallis test [11], and ReliefF [26], to name a few. Readers can find comprehensive reviews on this topic [23].

3 CRITERIA INTEGRATION VIA RANK AGGRADATION

After developing multiple biological relevance criteria, our next challenge is to integrate these criteria to comprehensively evaluate genes' relevance from different perspectives. Using different types of biological relevance criteria, we can obtain multiple lists that rank genes in different orders. In this work, we propose a probabilistic model for rank aggregation. Aggregating rankings into a joint one has been studied in both machine learning and information retrieval. Most existing rank aggregation algorithms [4, 25] treat different rank lists equally in the combination, while the proposed method is able to automatically learn a set of combination coefficients according to the importance of different rank lists. This is achieved by maximizing the likelihood of genes' relevance in a specific given gene set, providing a supervised way for rank aggregation. Supervised rank aggregation is also studied in [19], but it requires to provide the supervision information via partial orders among entry pairs, which is not intuitive in our application. Let g_i denote gene i, $1 \le i \le M$, and its rank in rank list l be $r_{l,i}$, we define the probability of g_i to be relevant according to its rank in the rank list l as:

$$P(r_{l,i}) = \frac{1}{B} exp\left(\frac{1}{r_{l,i}}\right),\ B = \sum_{j=1}^{M} exp\left(\frac{1}{j}\right). \quad (6)$$

B is the normalization factor for the distribution. Given L rank lists $R_1, \ldots, R_L$, let the prior probability of picking the lth rank list, R_l, to rank genes be π_l, with $\pi_1 + \ldots + \pi_L = 1$. π_l reflects the reliability of R_l. To construct a mixture model [3], we introduce an L dimensional latent variable $\mathbf{z}_i = \{z_{i,1}, \ldots, z_{i,L}\}$ for each gene g_i, indicating using which rank list we rank g_i, that is if g_i's rank is taken from its rank in R_l, then $z_{i,l} = 1$ and all other elements in $\mathbf{z}_i$ are set to 0. Based on these definitions, we can formulate the joint likelihood of the relevance of $\mathbf{S} = \{g_1, \ldots, g_K\}$ as:

$$p(g_1, \ldots, g_K, Z | R_1, \cdots, R_L, \Theta) = \prod_{i=1}^{K} \prod_{l=1}^{L} \pi_l^{z_{i,l}} P(r_{l,i})^{z_{i,l}}. \quad (7)$$

Here, Z is the set of latent variables $Z = (z_{i,l})_{K \times L}$ $= (\mathbf{z}_1, \ldots, \mathbf{z}_K)$, and the prior $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_L\}$ can be obtained by maximizing the joint likelihood specified in Equation (7) with an EM algorithm.

Proposition 1 *For Computing $\boldsymbol{\pi}$ using EM algorithm, in the E step, $E(z_{i,l})$ is updated by:*

$$\gamma_{i,l} = E(z_{i,l}) = \frac{\pi_l P(r_{l,i})}{\sum_{j=1}^{L} \pi_j P(r_{l,i})}. \quad (8)$$

And in M step, $\boldsymbol{\pi}$ is updated by:

$$\pi_l^{new} = \frac{1}{K} \sum_{i=1}^{K} \gamma_{i,l}. \quad (9)$$

We omit the proof of the proposition due to space constraint. After obtaining $\boldsymbol{\pi}$, the probability of g_i to be relevant can be calculated by marginalizing the joint probability $P(g_i, R_l)$ as:

$$\begin{aligned} P(g_i) &= \sum_{l=1}^{L} P(g_i, R_l) = \sum_{l=1}^{L} P(g_i | R_l) P(R_l) \\ &= \sum_{l=1}^{L} P(r_{l,i}) \pi_l. \end{aligned} \quad (10)$$

The final gene rank list can be obtained by ranking the obtained relevance probabilities in descend order.

4 EXPERIMENTAL RESULTS

We empirically evaluate the proposed approach for gene selection. As it does Gene Selection based on Network Dynamic analysis, we named it **GSND**.

4.1 Data Sets and Experiment Setup

To cDNA Microarray data is used: (1) **Metastatic Prostate** data, the data is extracted from the Gene

Expression Omnibus (GEO) - GSE2545[3]. The data contains the expression profiles of **12,651** genes from **18** normal prostate tissues and **25** metastatic prostate tumor tissues. (2) **Glioblastomas** data, the data is extracted from the GEO - GSE1962. The data contains the expression profiles of **54,163** genes from **23** normal brain tissues and **81** glioblastoma tumor tissues. One biological network data is used: (3) **BioCarta** network data, the data is retrieved from NCI Pathway Interaction Database [24], which contains **12,223** nodes and **15,094** edges. The data is obtained by combining **254** human signaling pathways.

In the experiment, we use the measure defined in Equation 1 and 2 to evaluate genes' influence; use betweens centrality defined in Equation 3 and 4 to evaluate genes' capability to cause fragility; and use Information Gain to detect the abnormal expression patterns of genes. The three relevance criteria are integrated in two ways: $\text{GSND}_{\text{Borda}}$ and $\text{GSND}_{\text{Prob}_{\text{SUP}}}$, which correspond to using the Borda count [7] and the probabilistic model proposed in Section 3 (using only cancer related genes to learn combination coefficients) in GSND, respectively. Borda count is a representative rank aggregation algorithm based on majority voting, which is used as a baseline for comparison in the experiment. We also included Information Gain and Kruskal-Wallis test [11] in the experiment as baseline methods. They correspond to using the traditional supervised gene selection on Microarray data to select genes based on their statistical relevance. To evaluate the performance of different methods, we use two types of criteria. They are: (1) **Accuracy**: accuracy of 1-nearest-neighbor (1NN) classifier achieved on the top ranked genes provided by different methods. (2) **Hit ratio**: HIT_C, HIT_G and HIT_P, correspond to the hit counts of known cancer related, Glioblastoma related and Prostate cancer related genes in the 35 top ranked genes, respectively. The gene-disease association information is obtained from Cancer Gene Index database[4]. The two criteria are used together to evaluate the biological relevance of genes.

4.2 Empirical Findings

Tables 1 and **??** contain the experimental results obtained from ranking genes using different criteria. Based on the results, we report the following observations. Comparing in terms of accuracy, the 2 baseline algorithms: Information Gain and Kruskal-Wallis test, achieve high accuracy on both data sets (accuracy>90%). High accuracy indicates that the top ranked genes in their lists are statistically relevant, since they can separate samples from different categories. The result practically verified their capability on detecting statistical relevance. For the methods derived from GSND, we observed that they also result in high accuracy. The results indicate that the genes selected by the methods derived from the GSND have high statistical relevance. We also noticed that comparing with $\text{GSND}_{\text{Borda}}$, $\text{GSND}_{\text{Prob}_{\text{SUP}}}$ achieve higher accuracy on both data sets, which tells that it is able to detect genes bearing more statistical relevance. Comparing with other criteria, using genes selected by the influence measure and the fragility measure provide lower accuracy. This is reasonable, since the two measures do not directly use expression profiles of genes[5].

Comparing in terms of the hit ratio, the influence measure and the fragility measure achieve good performance (HIT_C >30, HIT_P and HIT_G >15), which indicates their capability on selecting cancer related genes, and suggests that the hypothesis used to design the two measures are effective. According to hit ratio, Information Gain and Kruskal-Wallis test do not perform well, suggests that many top ranked genes in their rank lists may not be cancer related. For the methods derived from GSND, we observed that they all achieve good performance. While comparing with $\text{GSND}_{\text{Borda}}$, $\text{GSND}_{\text{Prob}_{\text{SUP}}}$ achieves higher hit ratios.

On both accuracy and hit ratio, the methods derived from GSND achieve good performance. The two performance measures together suggest that GSND is effective on detecting biological relevance. We also noticed that comparing with $\text{GSND}_{\text{Borda}}$, $\text{GSND}_{\text{Prob}_{\text{SUP}}}$ always performs better, which clearly suggests that the supervision information used in $\text{GSND}_{\text{Prob}_{\text{SUP}}}$ is very helpful in improving performance.

4.3 Study on Biological Relevance

Table 2 contains the top 35 genes selected by $\text{GSND}_{\text{Prob}_{\text{SUP}}}$ on the Metastatic Prostate Cancer data and the Glioblastomas data. There are 8 oncogenes or tumor suppressors detected on the Metastatic Prostate Cancer data, including TP53, HRAS, JUND, AKT1, LYN, RALA, FOS and RAF1. On Glioblastomas data, there are also 8 oncogenes or oncoprotein homologs or tumor suppressor selected, including JUN, MYC, HRAS, STMN1, AKT1, FYN, RELA and BTG1. Besides these oncogenes, there are also genes responsible for cell division and chromosome alternation regulation, such as CDC2, CDK4, PAK1, F2R, and RCC1 from Glioblastomas data; and genes related to regulation of target genes transcription and translation, such as STAT6, EIF1AX, EEF1A1 and PABPC1 from

[3]http://www.ncbi.nlm.nih.gov/geo

[4]https://cabig.nci.nih.gov/inventory/data-resources/cancer-gene-index

[5]To evaluate a gene, the influence measure uses the expression profiles of its neighbors, but not its own.

Table 1: Performance comparison. ACC_{10}, ACC_{20} and ACC_{35} correspond to the accuracy achieved by 1NN classifier using the top 10, 20 and 35 genes, respectively. And ACC_{AVE} corresponds to the averaged accuracy. The upper and lower parts of the table corresponds to on the Metastatic Prostate and the Glioblastomas data respectively.

Methods	ACC_{10}	ACC_{20}	ACC_{35}	ACC_{AVE}	HIT_C	HIT_P
InfoGain	**1.00**	**1.00**	**1.00**	**1.00**	17	4
KrWallis	**0.95**	**0.95**	**0.98**	**0.96**	25	14
Centrality	0.82	0.72	0.86	0.80	**31**	**26**
Influnce	0.84	0.86	0.89	0.86	**31**	**20**
$GSND_{Borda}$	0.89	**0.93**	**0.93**	**0.92**	**32**	**17**
$GSND_{Prob_{SUP}}$	**1.00**	**1.00**	**1.00**	**1.00**	**33**	**25**
InfoGain	**0.99**	**0.97**	**0.98**	**0.98**	11	0
KrWallis	**0.94**	**0.97**	**0.97**	**0.96**	12	1
Centrality	0.86	0.88	**0.90**	0.88	**35**	**22**
Influnce	0.87	0.85	0.88	0.86	**35**	**17**
$GSND_{Borda}$	**0.94**	**0.95**	**0.94**	**0.95**	**33**	14
$GSND_{Prob_{SUP}}$	**0.95**	**1.00**	**0.94**	**0.97**	**32**	**15**

prostate data, and EIF4E, EIF4EBP1 from Glioblastomas data. The genes are all related to tumorigenesis.

For Metastatic Prostate Cancer data,two biomarkers for prostate cancer is detected, EEF1A and GSTP1. PTI-1, encodes a truncated and mutated human EEF1A, gene expression may provide an extremely sensitive indicator for prostate carcinoma progression as reflected by the presence of prostate carcinoma cells in a patients' bloodstream [29]. GSTP1, as a single hypermethylated marker, is informative in 80%∼90% of prostate cancer [8]. For Glioblastomas data, PRKCZ is detected, which is known to be critical for proliferation in human glioblastoma cell lines [6]. Interestingly, we observed some common genes in both lists, indicating their importance and relevance to tumoriogenesis of different cancers. For example, two oncogenes, AKT1 and HRAS: these genes have high centrality and influence values, indicating they might be the fragile cores of the biological network. MAP2K1: it is a member of the dual specificity protein kinase family, which acts as a mitogen-activated protein (MAP) kinase. The four mitogen-activated protein kinases (MAPK1, MAPK3, MAPK8, MAPK14): their protein products have been implicated in diverse cellular processes including cell growth, proliferation, differentiation, and survival.

The pilot study on the two gene lists shows that many of the top ranked genes provided by $GSND_{Prob_{SUP}}$ have strong correlation to tumorigenesis. Several genes in the list are the key components of some important cancer related pathways, such as the MAPK pathway. The results indicate the high potential of the proposed approach on identifying new genes-disease or pathway-disease associations to achieve novel biological discoveries.

5 DISCUSSIONS AND CONCLUSION

The authors in [16] try to select genes that can distinguish samples of different phenotypes, and have consistent expression patterns in its neighborhood. In the approach, genes with different topological importance to the network are treated equality, which is counter intuitive. Recent study also showed that the approach may not work well in practice [2].

In this work, we proposed a general framework, GSND, to address the novel problem of biological relevance detection, via network dynamic analysis. Experimental results showed that the methods derived from GSND can select genes bearing significant biological relevance. To the best of our knowledge, this work is the first explicit attempt to systematically apply network dynamic analysis in gene selection for biological relevance detection. The developed GSND framework forms our preliminary work for knowledge oriented gene selection. Our ongoing work includes: (1) designing other criteria to comprehensively evaluate biological relevance from multiple perspectives; (2) understanding the roles of different types of knowledge [21] in gene selection, and including them in GSND, and (3) developing a user friendly toolbox.

FUNDING

This work is, in part, sponsored by NSF-0812551.

References

[1] P. E. Barbano, etal. A mathematical tool for exploring the dynamics of biological networks. *PNAS*, 104(49):19169–9174, 2007.

Table 2: The top 35 genes selected by $\text{GSND}_{\text{Prob}_{\text{SUP}}}$ on two data sets.

Metastatic Prostate Cancer Data: TP53, HRAS, JUND, ACTA1, MAPK14, MAPK3, EIF1AX, NFATC1, AKT1, PFN1, PRKCA, GSK3B, RAC1, MAPK1, GSTP1, MYLK, RHOA, ILK, MAPK8, FHL2, MAP2K1, PABPC1, CASP8, RAF1, RALA, TPX2, FOS, EEF1A1, NUP210, CDC2, STAT6, UBE2I, MAP3K1
Glioblastomas Data: JUN, HRAS, RCC1, MAPK3, STMN1, AKT1, GPHN, MAPK14, FYN, EGFR, RELA, MAPK1, NONO, CAMK2B, MADD, RANGAP1, EPS15, EIF4E, EIF4EBP1, PRKACB, P4HB, PRKCA, MAP2K1, HMGB2, PRKCZ, MAPK8, CDC2, BTG1, CDK4, EZH2, F2R, RAC1, TIMP4, PAK1, MYC

[2] H. Binder and M. Schumacher. Comment on "network-constrained regularization and variable selection for analysis of genomic data". *Bioinformatics*, 24(21):2566–2568, 2008.

[3] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[4] J. J. Chen, C.-A. Tsai, S. Tzeng, and C.-H. Chen. Gene selection with multiple ordering criteria. *BMC Bioinformatics*, 8:74, 2007.

[5] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. Technical report, arXiv:0706.1062v2, 2009.

[6] A. M. Donson, etal. Protein kinase c zeta isoform is critical for proliferation in human glioblastoma cell lines. *J Neurooncol*, 47(2):109–15, 2000.

[7] C. Dwork, and et al. Aggregation methods for the web. In *WWW*, 2001.

[8] M. Esteller. Aberrant dna methylation as a cancer-inducing mechanism. *Annu Rev Pharmacol Toxicol*, 45:629–56., 2005.

[9] C. L. Freeman. Centrality in social networks: Conceptual clarification. *Social Networks*, 1(3):215–239, 1979.

[10] T. Gross and B. Blasius. Adaptive coevolutionary networks: a review. *J. R. Soc.*, 5:259–71, 2008.

[11] M. Hollander and D. A. Wolfe. *Nonparametric Statistical Methods*. Wiley, 1973.

[12] D. B. Johnson. Efficient algorithms for shortest paths in sparse networks. *Journal of the ACM*, 24(1):1–13, 1977.

[13] H. Kitano. Biological robustness. *NATURE REVIEWS GENETICS*, 5:826–837, 2004.

[14] D. Koschzki, etal. *Network Analysis: Methodological Foundations*, chapter Centrality Indices, pages 16–1. Springer-Verlag, 2005.

[15] J. Lehar, A. Krueger, G. Zimmermann, , and A. Borisy. High-order combination effects and biological robustness. *Mol. Sys. Bio.*, 4:215, 2008.

[16] C. Li and H. Li. Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, 24(9):1175–82, 2008.

[17] T. Li, etal. A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *bioinformatics*, 20:2429–2437, 2004.

[18] H. Liu and H. Motoda, editors. *Computational Methods of Feature Selection*. Chapman and Hall/CRC Press, 2007.

[19] Y.-T. Liu, etal. Supervised rank aggregation. In *WWW*, 2007.

[20] T. Manke, etal. An entropic characterization of protein interaction networks and cellular robustness. *J. R. Soc.*, 3:843–850, 2006.

[21] J. H. Phan, etal. Improving the efficiency of biomarker identification using biological knowledge. In *PSB*, pages 427–38, 2009.

[22] S. J. Raudys and A. K. Jain. Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Trans. Pattern Anal. Mach. Intell.*, 13:252–264, 1991.

[23] Y. Saeys, I. Inza, and P. Larraga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517, Oct 2007.

[24] C. F. Schaefer, K. Anthony, S. Krupa, J. Buchoff, M. Day, T. Hannay, and K. H. Buetow. Pid: The pathway interaction database. *Nucleic Acids Res*, 37:674–9, 2009.

[25] F. Schalekamp and A. van Zuylen. Rank aggregation: Together we're strong. In *Proceedings of the Tenth Workshop on Algorithm Engineering and Experiments (ALENEX)*, 2009.

[26] M. R. Sikonja and I. Kononenko. Theoretical and empirical analysis of Relief and ReliefF. *Machine Learning*, 53:23–69, 2003.

[27] C. Sima and E. R. Dougherty. What should be expected from feature selection in small-sample settings. *Bioinformatics*, 22:2430–2436, 2006.

[28] L. Sun, A. M. Hui, Q. Su, A. Vortmeyer, and et al. Neuronal and glioma-derived stem cell factor induces angiogenesis within the brain. *Cancer Cell*, 9(4):287–300, 2006.

[29] Y. Sun, J. Lin, A. E. Katz, and P. B. Fisher. Human prostatic carcinoma oncogene pti-1 is expressed in human tumor cell lines and prostate carcinoma patient blood samples. *Cancer Res*, 57(1):18–23, 1997.