Smoothing Gene Expression Using Biological Networks

Yue Fan, Mark Kon
Dept. of Mathematics and Statistics
Boston University
Boston, USA
Email: yue@bu.edu, mkon@bu.edu

Shinuk Kim, Charles DeLisi
Bioinformatics and Systems Biology
Boston University
Boston, MA
Email: kshinuk@bu.edu, delisi@bu.edu

***Abstract*—Gene expression (microarray) data have been used widely in bioinformatics. The expression data of a large number of genes from small numbers of subjects are used to identify informative biomarkers that may predict or help in diagnosing some disorders. More recently, increasing amounts of information from underlying relationships of the expressed genes have become available, and workers have started to investigate algorithms which can use such a priori information to improve classification or regression based on gene expression. In this paper, we describe three novel machine learning algorithms for regularizing (smoothing) microarray expression values defined on gene sets with known prior network or metric structures, and which exploit this gene interaction information. These regularized expression values can be used with any machine classifier with the goal of better classification. In this paper, standard smoothing (denoising) techniques previously developed for functions on Euclidean spaces are extended to allow smoothing of microarray expression feature vectors using distance measures defined by biological networks. Such *a priori* smoothing (denoising) of the feature vectors using metrics on the index space (here the space of genes) yields better signal to noise ratios in the data. When tested on two breast cancer datasets, support vector machine classifiers trained on the smoothed expression values obtain better areas under ROC curves in two cancer datasets.**

***Keywords*-Gene Expression, Smoothing, Clustering, Support Vector Regression**

I. Introduction

In computational biology, genome-wide microarray expression profiles have proved to be informative for phenotype classifications, for example classifications involving disease diagnosis, prognosis, and etiology [1], [2], [3]. Some technical challenges in the use of such information include problems related to feature selection and noise reduction. Feature selection, in the context of phenotype classification, aims to select a small group of features (e.g. genes) as biomarkers which can help eliminate irrelevant information by restricting to these most informative features (e.g. highly differentially expressed genes), in addition to triggering further biological/clinical investigation of the disorder mechanism [4], [5], [6], [7], [8]. Nevertheless, the tremendous amount of noise residing in the expression profiles limits the performance of classification. This is due to the randomized character of the microarray tissue selection and hybridization process, as well as measurement noise and other factors. In addition, because a biological disorder is usually caused by a smaller number of genes, irrelevant genetic correlations due to the dynamics of biological processes mask the disease-related biomarkers under a huge amount of information (the problem of too many genes), with limited improvement through the use of regular feature selection methods.

Recently, methods for denoising gene expression data have drawn attention from researchers. In addition to methods using posterior information (which is internal to the data), some techniques take into account prior information (external to data) involving previously known genetic relationships. Of particular interest are metric-type or 'distance' relationships among genes. Such prior relationships are usually represented in the form of networks or graphs to indicate similarity among genes based on gene ontology annotations [9], biological pathways [10], membership in protein modules [11], etc. Gene expression patterns can then be denoised by imposing expression similarity on groups of highly related genes. Along this approach, different methods have been used to obtain denoised expression features [12], [13], [14], [15], [16]. Techniques involving adaptations of more functional and numerical analytic approaches have involved spectral decompositions of gene expression functions [17] and graph wavelet-based denoising of gene expression functions [18].

In this paper we advocate a multifaceted approach which adapts various denoising methods used in statistics (e.g., [19]) and numerical analysis [20]. The denoised feature vectors can help the later implementation of machine learning algorithms and improve the classification/regression performance. In particular, we focus on representing the expression feature vector $\mathbf{z} = (z_1, \ldots, z_n)$ (z_j = expression value of gene v_j) as a function $f(v)$ over a space V consisting of genes or their indices. We then adapt analytic smoothing techniques to functions on V, for example, by defining on V a global metric based on network information from the graph $G = (V, E)$ [21], [22]. We apply analogs of the above-mentioned analytic tools, e.g., locally constant regularization [23], kernel smoothing [24], and support

This work was partially supported by NCI grant 2548-5, and NIH grant GM080625-01

vector regression smoothing [25] to denoising microarray expression features, to illustrate a more general regularization method for feature vectors in machine learning (ML). The regularized (pre-processed) expression features can then be used in training and prediction in standard ML algorithms. As an application we use this method to pre-process expression feature vectors used in support vector machine (SVM) classification of two breast cancer datasets [6], [26] involving prediction of cancer metastasis. Comparing this to the same classifiers using raw (unregularized) expression features, the SVM using denoised expression can improve the area under the prediction ROC curve by 19.6 and 8.6 percentage points, respectively, on the two metastasis datasets.

In contrast to other microarray denoising methods (e.g., PAM [27]), such algorithms use prior knowledge on gene-gene relationships (in a gene space V) to smooth the values of their expression feature vectors, while most previous methods have used information from the expression feature space Z itself. More importantly, by considering Z as a function space on V, this approach to 'smoothing' of expression feature vectors (in Z) using network relationships enables adaption to the gene space V and more generally to any space which is the index set Ω of ML feature vectors of various smoothing/denoising techniques which have been successful in Euclidean spaces. Among these methods, some kernel-based methods (kernel smoothing and support vector regression) can also be shown to be scalable, in that multiple prior networks can be integrated transparently. Though here we use a single network (based on protein-protein interactions) to demonstrate the idea and its implementation, the method can be expanded to multiple networks using standard machine learning kernel combination methods.

II. Method

A. Overview of the Methodologies

We assume our biological machine learning (ML) problem is to estimate the conditional probability $P(Y = 1|\mathbf{z})$, given the observed expression feature vector $\mathbf{z} = (z_1, \ldots, z_n)$. Here $Y = 1$ or -1 indicates, e.g., disease or normal (or case and control) phenotypes, respectively. We remark that our focus is on the unsupervised denoising of expression feature vectors $\mathbf{z}$ (i.e., without reference or knowledge of the label Y), rather than on the classification problem, which is downstream from our unsupervised procedure. Depending upon the selection of subsequent ML classification algorithms, this approach can be used to pre-process feature vectors for various ML tasks, including multi-class classifiers and regression predictors.

Our model assumes the observed expression value z_j of gene v_j consists of signal plus noise, i.e.,

$$z_j = f(v_j) + \epsilon_j$$

where $f(\cdot)$ is the underlying target (e.g. gene expression) function to be approximated, while ϵ_j represents noise, i.e., any additional signal. We assume the ϵ_j are independent and identically distributed (iid).

Formulating the recovery of f as a regression-like problem requires an analog of a distance metric on the underlying (gene) space V, characterizing relationships of genes v_j in V. In this paper, we will use the above-mentioned *a priori* network relationships to derive a prior distance metric on the underlying space V (the index set of genes on which the features are defined) independently of the feature vectors $\mathbf{z}$ defined on it. In particular, we use the network structure $G = (V, E)$, where the vertices $v_j \in V$ are genes (or their protein products) and edges $e_j \in E$ (edges) are a set of interactions between genes (or proteins). Examples of such networks are protein-protein interaction (PPI) networks [11], metabolic pathways [10], transcriptomes [28], and any gene co-expression networks [30]. Under a distance metric derived from such a graph G, strongly connected gene pairs (usually connected through multiple network routes) will be closer to each other than weakly connected genes. The biological content of such a distance metric can include measurement of functional similarity, membership in a metabolic process or regulatory module, co-expression in other experiments, and other network-derived information. The denoising of observed expression feature vector is based on the intuition that $f(v_i)$ is similar to $f(v_j)$ if v_i and v_j are close to each other. More generally, such a regularization can also be done on any index set V of the feature vectors in an ML algorithm, assuming V admits some a priori (network or metric distance) relations on it.

B. Methodologies

This section presents mathematical details from the previous section, (and is separate from the biological/machine learning content of this paper). We want first to derive a metric (distance measure) among genes in the space V, from the network relationship. One intuitive metric is the *geodesic* distance on the graph G (i.e., the length of the shortest path connecting two vertices). With such a distance d defined on the gene space V, we can group genes into disjoint clusters V_k $(1 \leq k \leq K)$ with similarity in clusters corresponding, e.g., to functional or regulatory modules or other characteristics contained in the network G. Each cluster defines a neighborhood, while the total number of clusters K controls the sizes of neighborhoods. An expression feature vector $f(v)$ can then be regularized using local averaging. Thus $f(v)$ (viewed as a function on V) is projected onto constants (i.e., averaged) within each cluster to obtain a local cluster activity, defined by $f_k = |V_k|^{-1} \sum_{j:v_j \in V_k} z_j$. The expectation is that this locally smoothed (flattened) version of $f(v)$ can eliminate a good deal of random measurement and other error in order to improve ML inferences based on the feature vector $\mathbf{f} = (f_1, \ldots, f_k)$.

With the definition of such a distance metric d and corresponding neighborhoods on V, we can alternatively also adopt kernel smoothing [24] techniques from Euclidean space for a function f on V. Specifically, unlike the above clustering-based smoothing using equal weights and common neighborhoods, each gene now has its own set of neighbors (varying with the gene), with a weight given to each neighbor determined by its distance to the center gene and a kernel function $K(v_i, v_j) = k(d_{ij}/\sigma)$. The resulting smoothed approximation is

$$f(v_j) = \frac{\sum_i K(v_i, v_j) z_i}{\sum_i K(v_i, v_j)}$$

To avoid computation of the distance metric d, we can also test graph diffusion kernels [32] and support vector regression [25] to approximate the underlying expression function $f(v)$, globally or locally within a gene cluster. In particular, we have approximated the expression function for each *coexpression cluster* [29] V_k by a function from the family

$$f_k(v_j) = \sum_{i:v_i \in V_k} \alpha_{ki} K(v_i, v_j) z_i + b_k$$

and optimized the objective function

$$\sum_j C \cdot L(f_k(v_j), z_j) + \|f\|_K^2,$$

where $\|f\|_K$ represents the norm in the reproducing kernel Hilbert space with graph diffusion kernel K. $L(f_k(v_j), z_j) = (|f_k(v_j) - z_j| - \epsilon)^+$ is the loss function and C is a regularization parameter balancing the weights between prediction accuracy and smoothness f. Though it is similar to the second approach mentioned in the form of the function f (weighted average of neighboring expression values z_i), this approach does not explicitly use the distance metric d in its computation. Similarly, the distance metric induced from the graph diffusion kernel K measures the connectivity between two vertices (genes).

These three approaches have their own advantages.

1) **Control of over-denoising**: The clustering-based smoother controls over-denoising with the number of clusters k. This parameter is more understandable than the bandwidth σ or functional complexity $\|f\|_K$ used in the other two methods.
2) **Ease of computation**: Though in practice the computation of distance metrics can be very expensive, all three of the above methods can be computed without explicit calculation of the metric. In particular, for clustering-based smoothers, a graph clustering algorithm [31] is used to approximate the best clustering quickly. And diffusion kernels [32] can also be calculated for the kernel-based methods without solving for the distance metric. Beyond that, however, support vector regression is more computationally expensive.
3) **Flexibility of approximation**: Note that support vector regression is more flexible than the averaging-based methods, since it uses a family of functions to approximate locally the expression function f for each patient. When noise is low along with the risk of overfitting, the support vector regression estimate is able to capture more of the signal structure.

III. Application

The best way to test the effectiveness of the above denoising is to train a fixed classification algorithm on the original and denoised expression feature vectors, in order to compare the algorithm performances. We expect that using regularized feature vectors improves classification performance comparing to using raw feature vectors. We used two breast cancer datasets from high-throughput gene expression studies by Wang, et al. [6] and van de Vijver, et al. [26], together with outcome information on metastasis. Because the computational cost to derive the *geodesic* distance is significantly large, we only assessed the clustering-based smoothing (Approach 1) and support vector regression (Approach 3) to demonstrate the concepts of our methods. A protein-protein interaction network G compiled from multiple databases [11], [33] was used to provide *a priori* information of functional relationship for the construction of the denoising operators discussed above. The total number of genes in the raw datasets contains more than 5000+ genes.

We tried different numbers of clusters for the clustering-based smoother above. As expected, the performance on both datasets first improves as cluster size decreases, and then deteriorates as k increases (this non-monotonicity property of denoising accuracy is studied mathematically in [21]). Three measurements of performance used. AUROC and AUPRC are defined as areas under the ROC curve and the precision-recall curve, respectively. They assess the performance by measuring both power and prediction errors. ACC90 measures the prediction accuracy when the cut-off obtains 90% sensitivity. It is a common statistic used in disease prognosis. The best SVM classifier used on clustering-smoothed expression data is more accurate than SVM using raw gene expression features. For example, the area under the ROC curve is improved by clustering smoothing from 53.4% to 73.0% and from 66.0% to 71.2% for Wang's dataset and van de Vijver's dataset, respectively. Details are listed in Table I.

We tested support vector regression on the same datasets with the assistance of the LIBSVM [34] package (ϵ-SVR). To limit over-denoising, we controlled for low local complexity of the estimated expression function f. We also defined *coexpression sub-clusters* [29] to control the requirement that the complexity needed within each cluster be low. For the Wang dataset, the performance of SVR was comparable to cluster smoothing. In the van de Vijver dataset,

SVR further improved the area under the ROC curve from 71.2% to 74.6% (See Table II). Even though SV regression fits a better local expression function than averaging, it needs more data points for training. That explains why the optimal number of clusters for SVR method is smaller than for cluster-based smoothing.

No. Clusters	Wang			van de Vijver		
	AUROC	AUPRC	ACC90	AUROC	AUPRC	ACC90
64	0.658 (0.014)	0.450 (0.019)	0.470 (0.027)	0.687 (0.014)	0.371 (0.015)	0.472 (0.024)
128	0.680 (0.015)	0.462 (0.021)	0.477 (0.023)	0.705 (0.013)	0.399 (0.019)	0.520 (0.023)
256	0.692 (0.019)	0.475 (0.026)	0.526 (0.029)	0.689 (0.016)	0.398 (0.022)	0.490 (0.030)
512	0.684 (0.019)	0.487 (0.031)	0.502 (0.029)	0.686 (0.021)	0.375 (0.023)	0.489 (0.038)
1024	0.708 (0.019)	0.500 (0.032)	0.527 (0.029)	0.712 (0.019)	0.403 (0.026)	0.520 (0.026)
2048	0.730 (0.017)	0.522 (0.029)	0.567 (0.024)	0.500 (0.038)	0.270 (0.026)	0.311 (0.026)
RAW	0.534 (0.044)	0.362 (0.032)	0.430 (0.035)	0.660 (0.027)	0.346 (0.028)	0.535 (0.020)

Table I

PERFORMANCE OF CLUSTERING-BASED SMOOTHER ON WANG AND VAN DE VIJVER BREAST CANCER DATASETS. NUMBERS WITHOUT PARENTHESES ARE MEAN VALUES OF THE PERFORMANCE MEASURES OVER 200 5-FOLD CROSS VALIDATIONS. NUMBERS IN PARENTHESES ARE STANDARD DEVIATIONS OF THE PERFORMANCE MEASURES. AUROC AND AUPRC ARE AREAS UNDER ROC CURVE AND PRECISION-RECALL CURVE, RESPECTIVELY. THEY ASSESS THE PERFORMANCE BY BOTH POWER AND PREDICTION ERRORS. ACC90 MEASURES THE PREDICTION ACCURACY WHEN THE CUT-OFF OBTAINS 90% SENSITIVITY. IT IS A COMMON STATISTIC USED IN DISEASE PROGNOSIS. THE ROW INDICATED BY RAW REPRESENTS THE SVM TRAINED ON UNADJUSTED EXPRESSION FEATURE VECTORS. FOR ALL VALUES OF k TESTED, THE DATA SMOOTHED BY ALGORITHM 1 IMPROVES THE PERFORMANCE OF THE CLASSIFIER.

No. Clusters	Wang			van de Vijver		
	AUROC	AUPRC	ACC90	AUROC	AUPRC	ACC90
1	0.618 (0.013)	0.405 (0.014)	0.456 (0.024)			
64	0.672 (0.018)	0.503 (0.025)	0.476 (0.033)	0.706 (0.017)	0.441 (0.025)	0.468 (0.032)
128	0.698 (0.018)	0.519 (0.026)	0.52 (0.032)	0.738 (0.017)	0.456 (0.024)	0.527 (0.035)
256	0.716 (0.017)	0.526 (0.024)	0.565 (0.030)	0.741 (0.017)	0.465 (0.025)	0.536 (0.033)
512	0.71 (0.016)	0.515 (0.023)	0.567 (0.027)	0.746 (0.015)	0.478 (0.026)	0.552 (0.030)
1024	0.701 (0.015)	0.494 (0.022)	0.558 (0.027)	0.74 (0.013)	0.48 (0.025)	0.536 (0.022)
2048	0.676 (0.019)	0.47 (0.026)	0.521 (0.031)	0.718 (0.015)	0.441 (0.026)	0.532 (0.018)
RAW	0.54 (0.040)	0.364 (0.030)	0.434 (0.035)	0.661 (0.023)	0.351 (0.026)	0.535 (0.020)

Table II

PERFORMANCE OF SUPPORT VECTOR REGRESSION ON WANG AND VAN DE VIJVER BREAST CANCER DATASETS. NUMBERS WITHOUT PARENTHESES ARE MEAN VALUES OF THE PERFORMANCE MEASURES OVER 200 5-FOLD CROSS VALIDATIONS. NUMBERS IN PARENTHESES ARE STANDARD DEVIATIONS OF THE PERFORMANCE MEASURES.

IV. CONCLUSION

In this paper we have presented outlines of three new algorithms for smoothing (denoising) high throughput microarray expression data, all based on the principle of unsupervised regularization (preprocessing of feature vectors) using *a priori* information (in this case from biological networks). The conversion of network information into distance metrics in gene space has also allowed further study of more advanced smoothing/denoising techniques on expression data. When applied on two breast cancer datasets in predicting metastasis, the machine classifiers trained with such denoised expression features achieved better performances than those trained with raw expression features. In particular, areas under ROC curves were improved by 19.6 and 8.6 percentage points in two data sets using denoised expression features.

Conventional classification analyses effectively assume that gene expressions are independent, and treat them as individual features for building classification or regression models. Network information allows introduction of inter-gene metrics into consideration, which more generally encourages a search for modularized sets of biomarkers to supplement or replace individual gene biomarkers. Because genes in the same functional modules have similar denoised expression values, feature selection will be more likely to identify groups of genes in the same modules as a result. In addition, derived distance metrics on gene space can also be useful in quantifying gene interactions. Other statistical principles, such as canonical correlations and principal component analysis, are also good candidates for extension into gene space; such additional work could lead to the development of more advanced modular biomarker identification algorithms.

ACKNOWLEDGMENT

This work was partially supported by NCI grant 2548-5, and NIH grant GM080625-01

REFERENCES

[1] J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, and e. a. Chan, W. C., "Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling." *Nature*, vol. 403, no. 6769, pp. 503–511, 2000.

[2] C. D. Bloomfield and E. S. Lander, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring." *Science*, vol. 286, no. 5439, pp. 531–537, 1999.

[3] A. Perez-Diez, A. Morgun, and N. Shulzhenko, "Microarrays for cancer diagnosis and classification." *Advances in experimental medicine and biology*, vol. 593, pp. 74–85, 2007.

[4] M. West, C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J. A. Olson, J. R. Marks, and J. R. Nevins, "Predicting the clinical status of human breast cancer by using gene expression profiles." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 20, pp. 11 462–11 467, 2001.

[5] S. Ramaswamy, K. N. Ross, E. S. Lander, and T. R. Golub, "A molecular signature of metastasis in primary solid tumors." *Nature genetics*, vol. 33, no. 1, pp. 49–54, January 2003.

[6] Y. Wang, J. G. Klijn, Y. Zhang, A. M. Sieuwerts, M. P. Look, F. Yang, D. Talantov, M. Timmermans, M. E. Meijer-van Gelder, J. Yu, T. Jatkoe, E. M. Berns, D. Atkins, and J. A. Foekens, "Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer." *Lancet*, vol. 365, no. 9460, pp. 671–679, 2005.

[7] L. J. van 't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend, "Gene expression profiling predicts clinical outcome of breast cancer." *Nature*, vol. 415, no. 6871, pp. 530–536, January 2002.

[8] T. Ideker, V. Thorsson, A. F. Siegel, and L. E. Hood, "Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data." *Journal of computational biology*, vol. 7, no. 6, pp. 805–817, 2000.

[9] M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene ontology: tool for the unification of biology," *Nature Genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[10] M. Kanehisa and S. Goto, "Kegg: kyoto encyclopedia of genes and genomes." *Nucleic acids research*, vol. 28, no. 1, pp. 27–30, January 2000.

[11] A. Kanapin, S. Lewis, S. Mahajan, B. May, E. Schmidt, I. Vastrik, G. Wu, E. Birney, L. Stein, and P. D'Eustachio, "Reactome knowledgebase of human biological pathways and processes." *Nucleic Acids Research*, vol. 37, no. Database issue, pp. D619–D622, 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/18981052

[12] J. Su, B.-J. Yoon, and E. R. Dougherty, "Accurate and reliable cancer classification based on probabilistic inference of pathway activity," *PLoS ONE*, vol. 4, no. 12, p. e8161, 12 2009.

[13] J. A. Olson, J. R. Marks, H. K. Dressman, M. West, and J. R. Nevins, "Oncogenic pathway signatures in human cancers as a guide to targeted therapies." *Nature*, vol. 439, no. 7074, pp. 353–357, 2006.

[14] E. Lee, H.-Y. Chuang, J.-W. Kim, T. Ideker, and D. Lee, "Inferring pathway activity toward precise disease classification." *PLoS Computational Biology*, vol. 4, no. 11, p. e1000217, 2008.

[15] Q. Wang and S. Rao, "Towards precise classification of cancers based on robust gene functional expression profiles." *BMC Bioinformatics*, vol. 6, no. 1, p. 58, 2005.

[16] H.-Y. Chuang, E. Lee, Y.-T. Liu, D. Lee, and T. Ideker, "Network-based classification of breast cancer metastasis." *Molecular Systems Biology*, vol. 3, no. 140, p. 140, 2007.

[17] F. Rapaport, A. Zinovyev, M. Dutreix, E. Barillot, and J.-P. Vert, "Classification of microarray data using gene networks." *BMC Bioinformatics*, vol. 8, no. 1, p. 35, 2007.

[18] S. Yang and E. D. Kolaczyk, "Target detection via network filtering." *eprint arXiv:0902.3714*, 2009.

[19] D. Donoho, "De-noising by soft-thresholding." *IEEE Transactions on Information Theory*, vol. 41, no. 3, pp. 613–627, 1995.

[20] T. F. Chan and J. Shen, *Image Processing and Analysis: Variational, PDE, Wavelet and Stochastic Methods*. Society for Industrial Mathematics, 2005.

[21] Y. Fan, S. Kim, M. A. Kon, L. Raphael, and C. DeLisi, "Regularization Techniques for Machine Learning on Graphs and Networks with Biological Applications." *Commun. Math. Anal.*, vol. 8, No. 3, pp. 136–145, 2010.

[22] Y. Fan, S. Kim, M. A. Kon, L. Raphael, and C. Delisi, "Functional analytic tools for machine learning." *preprint*, 2010.

[23] W. Härdle, *Applied Nonparametric Regression (Econometric Society Monographs)*. Cambridge University Press, January 1992.

[24] M. P. Wand and M. C. Jones, *Kernel Smoothing (Chapman & Hall/CRC Monographs on Statistics & Applied Probability)*. Chapman and Hall/CRC, December 1994.

[25] V. Vapnik, *The Nature of Statistical Learning Theory (Information Science and Statistics)*. Springer, November 1999.

[26] M. Parrish, D. Atsma, A. Witteveen, A. Glas, L. Delahaye, T. Van Der Velde, H. Bartelink, S. Rodenhuis, E. T. Rutgers, and e. a. Friend, S H, "A gene-expression signature as a predictor of survival in breast cancer," *The New England Journal of Medicine*, vol. 347, no. 25, pp. 1999–2009, 2002.

[27] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu, "iagnosis of multiple cancer types by shrunken centroids of gene expression," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 10, pp. 6567–6572, May 2002.

[28] V. Matys, E. Fricke, R. Geffers, E. Gossling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. E. Kel, O. V. Kel-Margoulis, D.-U. Kloos, S. Land, B. Lewicki-Potapov, H. Michael, R. Munch, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele, and E. Wingender, "TRANSFAC(R): transcriptional regulation, from patterns to profiles," *Nucl. Acids Res.*, vol. 31, no. 1, pp. 374–378, 2003.

[29] K. Yeung, M. Medvedovic, and R. Bumgarner, "Clustering gene-expression data with repeated measurements," *Genome Biology*, vol. 4, no. 5, pp. R34+, 2003.

[30] B. Zhang and S. Horvath, "A general framework for weighted gene co-expression network analysis." *Statistical applications in genetics and molecular biology*, vol. 4, no. 1, 2005.

[31] I. S. Dhillon, Y. Guan, and B. Kulis, "Weighted graph cuts without eigenvectors a multilevel approach," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 11, pp. 1944–1957, 2007.

[32] R. I. Kondor and J. D. Lafferty, "Diffusion kernels on graphs and other discrete input spaces," in *ICML '02: Proceedings of the Nineteenth International Conference on Machine Learning*. Morgan Kaufmann Publishers Inc., 2002, pp. 315–322.

[33] S. Razick, G. Magklaras, and I. M. Donaldson, "irefindex: A consolidated protein interaction database with provenance," *BMC Bioinformatics*, vol. 9, no. 1, p. 405, 2008.

[34] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001.