A New Method for Identifying Cancer-Related Gene Association Patterns

Hong-Qiang Wang[1,⋆], Xin-Ping Xie[2], and Ding Li[1]

[1] Intelligent Computing Lab, Hefei Institute of Intelligent Machine, CAS, P.O. 1130, 230031, Hefei, China
{hqwang126,ld0510104}@126.com
[2] Department of Mathematics and physics, Anhui University of Architecture, 230022, Hefei, China
xpxie@yahoo.com.cn

Abstract. Gene association plays important roles in complex genetic pathology of cancer. However, development of methods for finding cancer-related gene associations is still in its infancy. Based on a biological concept of gene association module (GAM) comprising a center gene and its expression-related genes, this paper proposes a gene association detection model called kernel GAM (kGAM). In the model, we assume that the expression of the center gene can be predicted by the expression-related genes. Based on defining a cost function, a kernel ridge regression algorithm is developed to solve the kGAM model. Finally, to identify a compact GAM for a given center gene, a heuristic search procedure is designed. Experimental results on three publicly available gene expression data sets show the effectiveness and efficiency of the proposed kGAM model in identifying cancer-related gene association patterns.

Keywords: Microarray data, kernels, ridge regression, gene association.

1 Introduction

Genes in a cell working together and functioning in a coordinated manner plays an important role in the generation of cellular phenotypes and fine coordination between gene activities is essential for the formation of a signaling pathway [12,17]. These coordinated activities are manifested in the form of correlated expression levels of genes [2,3]. Therefore, it is critical and necessary to detect and utilize gene associations to understand complex genetic diseases. Another motivation of this work is that, although the large volume of gene expression data have been accumulated and are available online, it is still difficult and challenging to mine biological knowledge from these data in terms of methodology [10]. Generally, there are two main challenges in analyzing gene expression data: the complexity of invisible biological systems and the non-typical features of gene expression data including high noise, high-dimensionality but small sample size.

⋆ Corresponding author.

D.-S. Huang et al. (Eds.): ICIC 2011, LNBI 6840, pp. 115–122, 2012.

Many studies on various model systems have suggested that a gene can be combinatorially regulated by a relatively small number of transcription factors simultaneously or under different conditions, leading to strikingly complex patterns of gene expression. From these findings, we abstract a gene association structure, named gene association module (GAM), which consists of a center gene and its associated (unnecessarily regulating or regulated) elements (genes). In the GAM, the links, only appearing between the center gene and its associated genes, represent the influence of the associated elements on the center gene. As a hub topology, the GAM has been found to be universal in biological systems due to its robustness and sparseness for signal transduction [13,4,6].

In this paper, based on the GAM structure, we develop a kernel GAM model (kGAM) for detecting cancer-related complex gene associations. In the model, the main idea is to use the associated genes to regress the expression of the center gene. To characterize the model, a cost function is defined as the regression error. The cost function allows determining the structural parameters of the kGAM and potentially provides a way to use kGAMs to classify cancer. To find a compact GAM for a given center gene, a heuristic compact-kGAM searching procedure is developed based on the cost function.

In experimental section, we collect three publicly available real-world gene expression data sets, binary or multi-class, to evaluate the performance of the proposed method in detecting gene association patterns. To evaluate the cancer classification performance of the proposed model, we also implement and apply several previous classification methods including Fisher discriminant analysis (FDA), K nearest neighbor(KNN), support vector machines with linear kernel (linear-SVM) and radial basis function kernel (rbf-SVM) to these data sets, and their classification accuracies are compared with those of our model.

2 Methods

2.1 kernel Gene Association Model

Considering a gene association structure composed of a center gene g and p associated elements (genes), we assume that the expression of the center gene can be predicted by the associated genes. Let y denote the expression level of gene g and $\mathbf{x} = [x_1, x_2, \cdots, x_p]$ the expression levels of the p associated genes, such kind of gene association structure can be linearly modeled as:

$$\begin{cases} \hat{y} = f(\mathbf{x}) = A\mathbf{x}^T + b \\ y = \hat{y} + \epsilon \end{cases} \tag{1}$$

where $A = [a_1, a_2, \cdots, a_p]$, b is a constant and $\epsilon \sim N(0, 1)$. The element a_i measures the association of gene i on the center gene g, and its positive value denotes an expression promotion on gene g while its negative value denotes an expression repression.

For such a structure, we define a cost function E as

$$E = \frac{1}{2}(\hat{y} - y)^2 = \frac{1}{2}(\bar{A}\bar{P})^2 \tag{2}$$

where $\bar{A} = [A, -1, 1]$ is referred to as an extended association coefficient vector and $\bar{P} = [x_1, x_2, \cdots, x_p, y, b]$ as an extended expression profile. From Eq.2, the cost function is dependent on the internal relationship of the structure. Given an expression profile sample, only when the association pattern implicity embedded in it, instead of the explicit gene expression values, agrees with the internal relationship will the value of the cost function approach zero. This agrees with the fact that the coordination between the genes, rather than the expression values themselves, plays a crucial role in determining gene activity, and the cost function reflects the level of this activity.

The complexity of biological systems suggests that gene associations may not proceed in a linear manner. We introduce a nonlinear kernel function to approximate the expression value y of the center center in Eq.1, and the resulting gene association structure is referred to as the kernel gene association model (kGAM). The kernel function is constructed as follows. We first consider such a kind of nonlinear transformation

$$\Phi(\mathbf{x}) = [\phi_1(\mathbf{x}^T\mathbf{e}_1), \phi_2(\mathbf{x}^T\mathbf{e}_2), \cdots, \phi_p(\mathbf{x}^T\mathbf{e}_p)]^T, \tag{3}$$

where $\phi_i, i = 1, 2, \cdots, p$ represents a nonlinear function, $\mathbf{e}_i = [e_{ij}; i, j = 1, 2, \cdots, p]^T$ and $e_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$. We use the sigmoid function and form ϕ_i as

$$\phi_i(\mathbf{x}) = \left(1 + exp(-\beta(\tfrac{x_i - \mu_i}{\sigma_i})^2)\right)^{-1}, i = 1, 2, \cdots, p \tag{4}$$

where μ_i and σ_i are the location and width parameters, respectively, which can be estimated as the mean and standard deviation of gene expression levels, and $\beta \in (0, 1]$ is a constant. As a result, by combining Eqs.3 and 4, we construct the kernel function as:

$$\kappa(\mathbf{x}, \mathbf{z}) = \sum_{i=1}^{p} (Logsig(\beta, x_i) Logsig(\beta, z_i)) \tag{5}$$

The parameter β is referred to as the kernel parameter, which controls the approximation to gene associations.

To efficiently solve the above gene association model, we introduced a ridge parameter $0 < \lambda < 1$ to impose a sparsity constraint on the values of the association coefficients. The ridge parameter controls the relative trade-off between the sparsity constraint and data approximation fidelity, and a proper value of it will compensate for the information insufficiency so that an effective solution can be found. For the kGAM model, we use the kernel ridge regression technique [14,7] to solve its parameter, A, and the kernel parameter β is optimally chosen by varying its value among (0,1).

2.2 kGAM-Based Cancer Classification and the Searching of a Compact kGAM

As described above, the cost function in Eq.2 reflects the association information encapsulated in a kGAM, and will approach zero when the expression profile of a

sample agrees with the internal relationship of the kGAM. This property can be used to design a association-based cancer classification rule as follows. Consider C sample classes, and for each class, with the center gene g and its p associated genes, a kGAM model, $H_i, i = 1, 2, \cdots, C$, has been built, respectively. For a given test sample t, we predict its class to be

$$c = \arg\min_i \{E_i(t)\} \tag{6}$$

where E_i are the cost functions associated to kGAMs H_i.

There is little or no knowledge about how many genes known truly correlated to a given center gene. To find a compact kGAM for a given center gene from a gene pool, we present a heuristic searching procedure. Simply speaking, the procedure begins with, and iteratively searches and attaches the element most associated with the center gene to the list of associated genes. Because gene networks tend to be sparse and only a small group of genes are involved in a particular biological process, the search procedure would converge within a small number of steps, and has a low computational cost.

3 Experimental Results

To evaluate the proposed approach, we collected three publicly available gene expression data sets, two binary datasets, Golub data [9], Singh data [16], and one multi-class dataset, Armstrong data [1]. The three data sets each have a standard training/test split [9,16,1]: For the Golub data, the training and test sets contain 38 and 34 samples, respectively; 102 and 34 samples for the Singh data; and 77 and 15 samples for the Armstrong data.

We analyzed the three data sets based on the standard splits: the training sets are used to detect significant kGAMs and construct classifiers, and the test sets are used for validation. In order to avoid the influence of noisy genes to cancer classification, only 200 genes, with the highest signal-to-noise ratio (SNR) [9] for the binary datasets or the highest variance between samples for the multi-class data set, were used for performance evaluation in our experiments. For each dataset, we tried the 50 genes with the highest SNR/variance values as center genes to search for significant kGAMs for cancer classification.

3.1 Detection of kGAMs for the Three Data Sets

The association coefficients capsulated in a kGAM reflect the gene association patterns in cancer classes. Fig. 1 shows the association coefficients in three kGAMs with "Human common acute lymphoblastic leukemia antigen (CALLA)" being the center gene, for the three classes of the Armstrong data set. Note that, to highlight significant association differences, the three kGAMs are simplified by trimming the association coefficients less than 5% of the maximal values to 0. A number of studies have shown that the CALLA gene plays a potential role

as a functional neutral endopeptidase in both normal and malignant lymphoid function. In particular, the gene associates with a number of small secreted peptides whose abnormal misfolding and aggregation may be a cause of a number of diseases [15]. As shown in Fig. 1, the three kGAMs identified suggest that the gene is differently regulated in the three leukemia cancer classes. For the three kGAMs, the TOP2B gene with Accession no. 36571_at is most closely associated with the CALLA gene. The TOP2B gene encodes a DNA topoisomerase, which can control and alter the topological states of DNA during transcription [5]. The three kGAMs disclose that the TOP2B gene represses the expression of the CALLA gene in all the three leukemia classes, as shown in Fig. 1. Some associated genes exhibit remarkably different effects on the CALLA gene in the three leukemia classes. For example, the gene with Accession no. 40797_at, known as ADAM10, promotes the expression of the CALLA gene in Class 2 while represses in Class 3; the gene with Accession no. 1602_at, known as PRKCI (Protein kinase C, iota), promotes the expression of the CALLA gene in Classes 1 and 2 while no significant impact occurs in Class 3. The PRKCI gene has been found to control the dynamics of microtubules within the early secretory pathway [8], and the ADAM10 gene to encode a sheddase, which performs cleaving of the membrane proteins and plays a role in a number of peptide hydrolysis reactions [18].

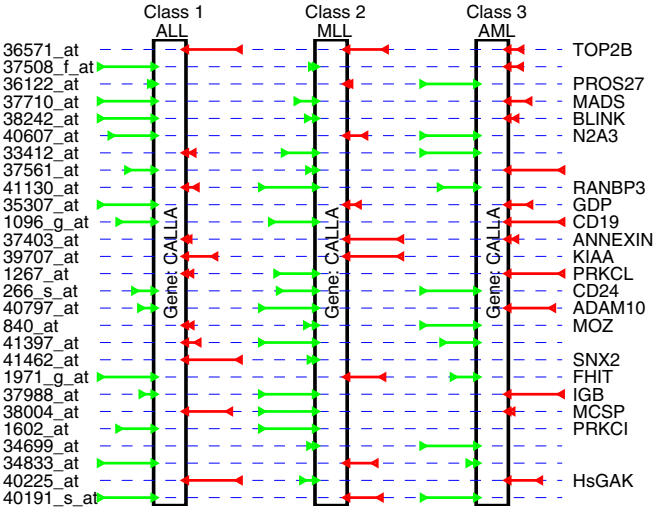

Fig. 1. Association patterns (maximum of the association coefficients is 9) captured in a kGAM model for the three cancer classes of the Armstrong data. The red lines represent negative expression association, the green lines represent positive expression association, and the length of lines represent the association strength. These associated genes have significantly different effects on the hub gene CALLA in the three cancer classes, and these differences in turn determine the characteristics of the three classes.

A kGAM encapsulates a stable association pattern common to a particular cancer class, and its cost function measures how a sample disagree with the class in association patterns. For samples belonging to the class, the values of the cost function will remain low due to the similar association pattern. To illustrate this property, Fig.2 shows the distribution of the cost values of the three classes for the Armstrong data set. From this figure, it can be seen that most of the samples in each class have a low cost value (less than 10^{-4}) according to the corresponding cost functions.

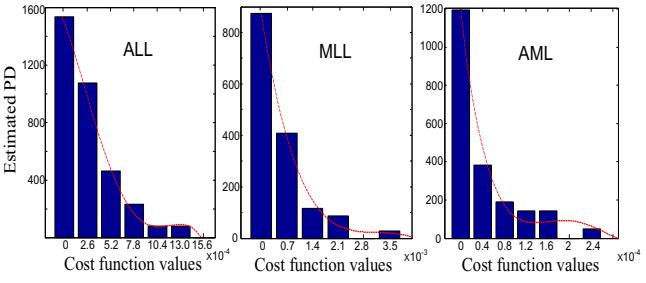

Fig. 2. Distribution of the cost values of the kGAMs found for the three cancer classes of the Armstrong data. The red doted lines are the fitting curves with 4-degree polynomial function. PD is short for probability density.

3.2 Evaluation of the Classification Performance of the kGAM Model

To further show the power of the kGAM model in capturing gene association patterns, we applied the kGAMs identified above to classify cancer according to the kGAM classification rule. Table 1 summarizes the classification accuracies on the test sets by three kGAMs for each of the three data sets. For comparison, based on the same genes as the three kGAMs contained, several conventional methods were implemented to classify the three data sets, which include two support vector machines (SVMs) with linear (linear-SVM) and radial basis function (rbf-SVM) kernels (http://sourceforge.net/projects/svm/), k(k=3))-nearest neighbor (KNN) and Fisher linear discriminant (FLD). The regularization parameter of the linear-SVM was optimally chosen from the range $\{2^{12}, 2^{11}, \cdots, 2^{-1}, 2^{-2}\}$, and the two parameters of the rbf-SVM, regularization factor and kernel width, were optimized based on a two-dimensional grid search technique within the ranges, $\{2^{12}, 2^{11}, \cdots, 2^{-1}, 2^{-2}\}$ and $\{2^{4}, 2^{3}, \cdots, 2^{-9}, 2^{-10}\}$. For the multi-class problem, the voting strategy [11] is used along with these previous methods to make optimal classification decision. The results by the conventional methods are compared with ours in Table 1. From Table 1, it can be seen that our kGAM models achieve much better classification accuracies than the other methods, irrespective of the binary problems or the multi-class problem.

Table 1. Comparison of classification accuracies between our kGAM model and several conventional methods for the Golub (binary), Singh (binary) and Armstrong (3-class) data

Datasets	Methods	kGAM I	kGAM II	kGAM III
Golub data	kGAM model	1	1	0.97
	rbf-SVM	0.94	0.97	0.94
	linear-SVM	0.91	0.97	0.88
	KNN	0.94	0.88	0.85
	FLD	0.88	0.88	0.85
Singh data	kGAM model	1	0.97	1
	rbf-SVM	0.94	0.91	0.91
	linear-SVM	0.91	0.76	0.91
	KNN	0.91	0.87	0.97
	FLD	0.38	0.60	0.48
Armstrong data	kGAM model	1	1	0.93
	rbf-SVM	0.93	0.86	0.80
	linear-SVM	0.73	0.60	0.73
	KNN	0.87	0.80	0.80
	FLD	0.67	0.53	0.47

4 Conclusion

We have proposed a model (kGAM) for detecting cancer-related gene associations. The model can flexibly approximate complex association patterns between genes and overcome the problem of small sample in microarray data analysis. The proposed approach was evaluated on three publicly available microarray data sets. The experimental results show the effectiveness and efficiency of the proposed approach in both capturing gene associations. Future work will focus on the optimal construction of the kGAM model and applications on more real-world microarray data sets.

Acknowledgments. This work was supported by the grants of the National Science Foundation of China, Nos. 31071168, 30900321, 60975005, 61005010, 60873012, 60973153 and 60905023.

References

1. Armstrong, S.A., Staunton, J.E., Silverman, L.B., Pieters, R., den Boer, M.L., Minden, M.D., Sallan, S.E., Lander, E.S., Golub, T.R., Korsmeyer, S.J.: Mll Translocations Specify a Distinct Gene Expression Profile That Distinguishes a Unique Leukemia. Nat. Genet. 30(1), 41–47 (2002); 1061-4036 10.1038/ng76510.1038/ng765
2. Basso, K., Margolin, A.A., Stolovitzky, G., Klein, U., Dalla-Favera, R., Califano, A.: Reverse engineering of regulatory networks in human b cells. Nature Genetics 37(4), 382–390 (2005)

3. Calvano, S.E., Xiao, W., Richards, D.R., Felciano, R.M., Baker, H.V., Cho, R.J., Chen, R.O., Brownstein, B.H., Cobb, J.P., Tschoeke, S.K., Miller-Graziano, C., Moldawer, L.L., Mindrinos, M.N., Davis, R.W., Tompkins, R.G., Lowry, S.F.: ProgramInflamm, L.S.C.R., Host Response to, I.: A Network-based Analysis of Systemic Inflammation in Humans. Nature 437(7061), 1032–1037 (2005); 0028-0836 10.1038/nature03985 10.1038/nature03985
4. Carter, S.L., Brechbuhler, C.M., Griffin, M., Bond, A.T.: Gene Co-expression Network Topology Provides a Framework for Molecular Characterization of Cellular State. Bioinformatics 20(14), 2242–2250 (2004)
5. Champoux, J.J.: DNA Topoisomerases: Structure, Function, and Mechanism. Annu. Rev. Biochem. 70(1), 369–413 (2002)
6. Cooper, T., Morby, A., Gunn, A., Schneider, D.: Effect of Random and Hub Gene Disruptions on Environmental and Mutational Robustness in Escherichia Coli. BMC Genomics 7(1), 237 (2006)
7. Du, K.L., Swamy, M.N.S.: Neural Networks in a Soft-computing Framework. Springer-Verlag London Limited, London (2006)
8. Fields, A.P., Regala, R.P.: Protein kinase c: Human Oncogene, Prognostic Marker and Therapeutic Target. Pharmacol. Res. 55(6), 487–497 (2007)
9. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., Lander, E.S.: Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science 286, 531–537 (1999)
10. Lee, I., Lehner, B., Crombie, C., Wong, W., Fraser, A.G., Marcotte, E.M.: A Single Gene Network Accurately Predicts Phenotypic Effects of Gene Perturbation in Caenorhabditis Elegans. Nat. Genet. 40(2), 181–188 (2008); 1061-4036 10.1038/ng.2007.70 10.1038/ng.2007.70
11. Schmann, J.: Pattern Classification: A Unified View of Statistical and Neural Approaches. Wiley Interscience, Hoboken (1996)
12. Segal, E., Friedman, N., Kaminski, N., Regev, A., Koller, D.: From Signatures to Models: Understanding Cancer Using Microarrays. Nat. Genet. 37, S38–S45 (2005)
13. Seo, C.H., Kim, J.R., Kim, M.S., Cho, K.H.: Hub Genes with Positive Feedbacks Function as Master Switches in Developmental Gene Regulatory Networks. Bioinformatics 25(15), 1898–1904 (2009)
14. Shawe-Taylor, J., Cristianini, N.: Kernel Methods for Pattern Analysis. Cambridge University Press, Cambridge (2004)
15. Shipp, M.A., Vijayaraghavan, J., Schmidt, E.V., Masteller, E.L., D'Adamio, L., Hersh, L.B., Reinherz, E.L.: Common Acute Lymphoblastic Leukemia antigen (CALLA) Is Active neutral endopeptidase 24.11 ("enkephalinase"): direct evidence by cDNA transfection analysis. Proceedings of the National Academy of Sciences of the United States of America 86(1), 297–301 (1989), http://www.pnas.org/content/86/1/297abstract
16. Singh, D., Febbo, P.G., Ross, K., Jackson, D.G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A.A., D' Amico, A.V., Richie, J.P., et al.: Gene Expression Correlates of Clinical Prostate Cancer Behavior. Cancer Cell 1, 203–209 (2002)
17. Tlsty, T.: Cancer: Whispering Sweet Somethings. Nature 453(7195), 604–605 (2008); 0028-0836 10.1038/453604a 10.1038/453604a
18. Yang, J., Price, M.A., Neudauer, C.L., Wilson, C., Ferrone, S., Xia, H., Iida, J., Simpson, M.A., McCarthy, J.B.: Melanoma Chondroitin Sulfate Proteoglycan Enhances Fak and Erk Activation by Distinct Mechanisms. J. Cell Biol. 165(6), 881–891 (2004)