Identification of Oncogenic Genes for Colon Adenocarcinoma from Genomics Data

Changhe Fu, Ling Jing
College Of Science
China Agricultural University
Beijing, China

Changhe Fu, Su Deng
School Of Mathematics
And System Science
Shenyang Normal University
Shenyang, China

Guangxu Jin
Department of Systems Medicine and
Bioengineering
The Methodist Hospital Research Institute,
Houston, US

***Abstract*—Identification of oncogenic genes from comprehensive genomics data with large sample size is of challenge. Here, we apply a well-established computational model, Bayesian factor and regression model (BFRM), to predict unknown colon cancer genes from colon adenocarcinoma genomic data. The BFRM takes advantages of its latent factors to characterize the underlying association between genes and the large number of colon cancer patients. Based on the known cancer genes in Online Mendelian Inheritance in Man (OMIM), we addressed three important latent factors focusing on characterization of heterogeneity of expression patterns related to specific oncogenic genes from the microarray data of 174 colon cancer patients. We found that the three latent factors can be employed to predict unknown colon cancer genes using the known oncogenic genes. These predicted unknown cancer genes were extensively validated by using the new somatic genes identified in the same patients from DNA sequencing data.**

***Index Terms*—Bayesian analysis; genomics data; somatic mutation; GO enrichment analysis**

I. Introduction

Rapid development of experimental techniques for genomic study accelerates the generation of a large number of genomics data. The decreasing costs of genomics assays make it feasible that even only one cancer type has a large data set involving in hundreds and thousands of cancer patients, for example, the rich data sets for cancer genomics in The Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov/). However, the growing sample number of genomics data leads to the challenges of identification of consistent oncogenic genes for a cancer of interests from the huge genomics data sets. Although the gene signatures studies [1-7] based on statistical p values can alleviate this problem, these studies are insufficient to indicate the underlying associations between the genes and the samples considered. Therefore, to better manipulate the emerging genomics data with large number of samples, we need to consider more powerful bioinformatics methods to address the unknown oncogenic genes for cancer studies.

Bayesian factor and regression model (BFRM) [8] has been widely used in identification of pathway activation [9] and repositioning of drugs for cancer studies [10]. Different from the general gene signature studies, the model uses linear latent factor in which the factor loadings matrix is sparse, i.e., each factor is related to only a relatively small number of genes, representing a sparse, parsimonious structure indicating the associations between genes and a large number of samples. A flexible class of sparsity-inducing priors helps to introduce a pre-defined patterns of zeros in the factor loadings matrix, enabling the associations between genes and factors satisfy the sparsity property. BFRM makes use of evolutionary stochastic search and MCMC method [11] to update the priors for the factor loadings matrix. Thus, the BFRM model provides new opportunities to analyze the genomics data and address unknown oncogenic genes for cancer studies.

In this paper, we develop a new computation framework to predict unknown cancer genes for colon adenocarcinoma. We applied the BFRM model on the cDNA microarray data from 174 colon cancer patients of TCGA. The BFRM employed 3 latent factors to address unknown colon cancer genes. In each factor, the genes are ranked by the values in the corresponding column of the factor loadings matrix. We found that the genes with high ranks includes the known and unknown colon cancer genes. The unknown top genes with high ranks are validated by the somatic mutation information derived from DNA sequencing. The results show that most of them have somatic mutations on its genome sequences. The further GO analysis indicates that these genes are significantly enriched in the cell cycle and DNA repair cellular functions, which are important for the progression of colon cancer.

II. Methods and Data

A. BFRM

In BFRM, genomics data can be denoted by a $p \times n$ matrix x whose ith sample is written as a regression term on some known covariates plus with a latent factor structure term. The model structure is as follows:

$$\mathbf{x}_i = \mu + \mathbf{B}h_i + \mathbf{A}\lambda_i + v_i \qquad i = 1:n \tag{1}$$

or

$$\begin{aligned} x_{g,i} &= \mu_g + \beta'_g + \alpha'_g \lambda_i + v_{g,i} \\ &= \mu_g + \sum_{j=1}^{r} \beta_{g,j} h_{j,i} + \sum_{j=1}^{k} \alpha_{g,j} \lambda_{j,i} + v_{g,i} \end{aligned} \tag{2}$$

where,

$\mu = (\mu_1,\ldots,\mu_p)'$ is a p vector of intercept terms.

B is a $p \times r$ regression parameters matrix whose elements are $\beta_{g,j}$, (*g = 1:p, j = 1:r*), whose rows are denoted by β'_g.

A is a $p \times k$ factor loadings matrix whose elements are $\alpha_{g,j}$, (*g = 1:p, j = 1:k*), whose rows are denoted by α'_g.

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

$h_i = (h_{1,i},\ldots,h_{r,i})'$ is a known r-vector of covariates for sample i. In some cases, the known covariates are unavailable, and the model contains only latent factor structure.

$\lambda_i = (\lambda_{1,i},\ldots,\lambda_{k,i})'$ is a latent factor k-vector.

$v_i = (v_{1,i},\ldots,v_{p,i})'$ is an residual error term for sample i which is assumed normal distribution.

For gene expression profiles, the factor loadings matrix is able to specify sparsity as other applications of factor models. Sparsity model induces a great many zeros in the high-dimensional matrix. That means, one gene will associate with a small number of as well as one factor may still include a small number of genes. Therefore, BFRM applies sparsity model on **A**. The covariate may relate to a few genes in an experiment for the statistical parsimony view. Then, **B** also is modeled as sparsity in BFRM. Usually, sparsity model assumes a series of prior distributions which denotes the probability of each element **A** and **B**. Besides assuming sparsity priors specified for **A** and **B**, the model requires the priors of other variances.

According to these priors, BFRM uses Markov chain Monte Carlo (MCMC) analysis for posterior simulation. For the conditional distributions of Markov chain, the model simulates it using standard routine of Bayesian theory [12]. MCMC methods are a class of algorithms for sampling from probability distributions based on constructing a Markov chain that has the desired distribution as its equilibrium distribution. Then. Monte Carlo methods uses the state of the chain after as a sample satisfied the desired distribution. BFRM implements MCMC analysis by a Gibbs sampling format [13] that is a standard method of Monte Carlo methods. The results of the MCMC analysis are the simulation of these matrixes, vectors, and their posterior probability.

B. GO Enrichment Analysis

In this paper, we use WebGestalt [14][15] to implement GO [16] enrichment analysis. WebGestalt provides visualization (Fig. 2 is a example) of significantly enriched GO categories for Biological Process, Molecular Function, and Cellular Component using three separate Directed Acyclic Graphs (DAGs). Each GO category is a node in the DAG. GO categories in red, which has a p value < 0.05, are the enriched GO categories while the black ones are their non-enriched parents.

C. Data

The dataset of colon adenocarcinoma gene expression analysis were derived from TCGA that is a community resource project. Gene expression analysis of TCGA colon samples used Agilent Expression 244K microarrays, and all samples were run on AgilentG4502A_07_3. The set of colon adenocarcinoma genomics has 174 samples.

The disorder information of genes are from OMIM which is a comprehensive, authoritative compendium of human genes and genetic phenotypes that is freely available and updated daily according to published literatures. Based on the cancer information of OMIM, we chose 173 genes which are related with cancer in OMIM [17].

Another dataset downloaded from TCGA is about somatic mutations which samples are the same as the dataset of colon adenocarcinoma gene expressions.

III. RESULTS

A. Gene signatures derived from latent factors

Before running BFRM, we need to set a series of parameters. Most of them that can be set as default are used for prior distributions mentioned as above. In this paper, we set the number of latent factors as 20. In the outputs of the model, the factor loadings matrix and the posterior estimation matrix of loadings matrix are important for following studies. The factor loadings matrix is **A** matrix mentioned in section Ⅱ that represent the weight of every gene related to every latent factor. The posterior estimation matrix of loadings matrix **P** shows the posterior distribution for the factor loadings matrix **A**. In order to select the genes included in the latent factors of interests, we have to ensure that their posterior probability estimations in the posterior estimation matrix are higher than a threshold. Here, we set $\theta = 0.99$ as the threshold.

B. Most of predicted colon cancer genes are mutated in the DNA-Seq analysis

After analysis of the posterior estimation matrix **P** of loadings matrix, there are 134 genes remained in **A** matrix for they pass the threshold. Among these genes, there are 27 genes related with colon cancer announced by OMIM, 67 genes labeled somatic mutations in the DNA-Seq analyses by TCGA, and 15 genes for both. We showed the results for all factors in Fig. 1. According to the percentages, we selected 3 factors with highest percentages, that is, 1, 3, 7.

C. Later factors are helpful to predict colon cancer genes

We ranked the genes in each factor by the absolute value of elements of **A**. The results are shown in Tables 1-3. The values in **A** represent the weights of genes related to factors in BFRM analysis.

We added the known colon cancer information curetted from OMIM and somatic mutation information derived from TCGA in the Tables. Among all genes of the three factors, most of the top genes are verified by the information. The results imply that our method is helpful to identity those unknown colon cancer genes.

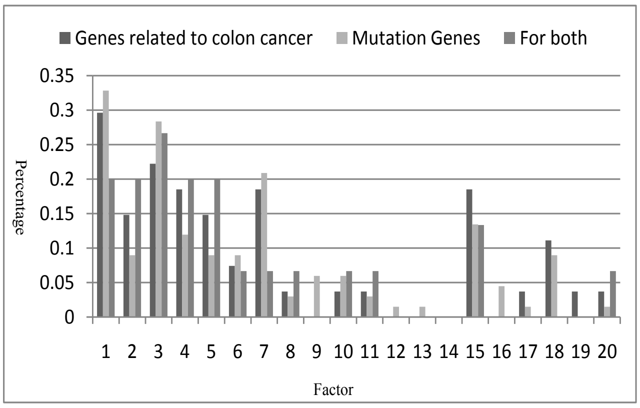

Fig. 1. Percentage of genes across factors

TABLE I. FACTOR 1 AND THE GENES

Rank	*Gene*	*/aij/ of A*	*Colon cancer related*	*Somatic mutation*	*P-value*

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

Rank	*Gene*	*$/a_{ij}/$ of A*	*Colon cancer related*	*Somatic mutation*	*P-value*
1	BRCA2	0.617		Yes	1.0
2	BUB1	0.556	Yes		1.0
3	BUB1B	0.527	Yes		1.0
4	AURKA	0.507	Yes	Yes	1.0
5	RAD54B	0.507	Yes	Yes	1.0
6	CHRNA3	0.479			0.999
7	RAD54L	0.452		Yes	1.0
8	HMMR	0.401		Yes	1.0

TABLE II. FACTOR 3 AND THE GENES

Rank	*Gene*	*$/a_{ij}/$ of A*	*Colon cancer related*	*Somatic mutation*	*P-value*
1	CTHRC1	0.450			0.999
2	IL1RN	0.416		Yes	0.999
3	NRAS	0.389	Yes	Yes	1.0
4	IGF2R	0.359		Yes	1.0
5	IL1B	0.352			0.991
6	CDH1	0.336		Yes	1.0
7	HMMR	0.316		Yes	1.0
8	PDGFRL	0.300	Yes		0.999

TABLE III. FACTOR 7 AND THE GENES

Rank	*Gene*	$/a_{ij}/$ *of A*	*Colon cancer related*	*Somatic mutation*	*P-value*
1	AXIN2	0.634	Yes	Yes	1.0
2	PLA2G2A	0.545	Yes		0.999
3	SAMD9	0.487		Yes	1.0
4	MLH1	0.448	Yes		1.0
5	SLC26A3	0.438	Yes		1.0
6	MAP3K8	0.377		Yes	0.999
7	EPHB2	0.351		Yes	1.0

D. GO enrichment analysis on the predicted genes

We applied GO analysis on 22 genes in the Table 1-3. The analysis is implemented by a web-based GSEA tool, called as WebGestalt. From the results shown on Fig. 2 and Table 4, we found that the genes are enriched in certain important cellular functions, such as, cell cycle. Table 5 displays the enrichment genes in GO biological process: cell cycle (GO:0007049).

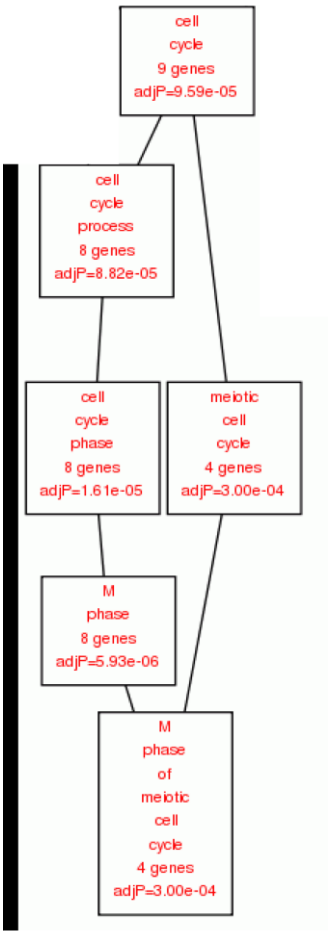
cell cycle 9 genes adjP=9.59e-05
cell cycle process 8 genes adjP=8.82e-05
cell cycle phase 8 genes adjP=1.61e-05
meiotic cell cycle 4 genes adjP=3.00e-04
M phase 8 genes adjP=5.93e-06
M phase of meiotic cell cycle 4 genes adjP=3.00e-04

Fig. 2. Screenshot of GO enrichment analysis from WebGestalt

TABLE IV. GO ENRICHMENT ANALYSIS (#GENE>=7)

GO	*# Gene*	*p-value*
Biological process: M phase	8	5.93e-06
Biological process: cell cycle phase	8	1.61e-05
Biological process: cell cycle process	8	8.82e-05
Biological process: cell cycle	9	9.59e-05
Biological process: cellular component organization	13	0.0002
Molecular function: protein kinase activity	7	0.0003
Molecular function: phosphotransferase activity, alcohol group as acceptor	7	0.0005
Molecular function: kinase activity	7	0.0006
Molecular function: transferase activity, transferring phosphorus-containing groups	7	0.0009
Molecular function: ribonucleotide binding	9	0.0014
Molecular function: purine ribonucleotide binding	9	0.0014
Molecular function: ATP binding	8	0.0015
Molecular function: adenyl ribonucleotide binding	8	0.0015

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

p-value from hypergeometric test

TABLE V. ENRICHMENT GENES OF BIOLOGICAL PROCESS: CELL CYCLE

Gene	*Colon cancer related*	*Somatic mutation*
BRCA2		Yes
MAP3K8		Yes
AURKA	Yes	Yes
BUB1B	Yes	
BUB1	Yes	
RAD54L		Yes
MLH1	Yes	
IL1B		
RAD54B	Yes	Yes

IV. Discussion and conclusion

From the results in Tables 1-3, we predict that the 12 genes, such as BRCA2, CTHRC1, RAD54L, which are not reported for colon cancer related in OMIM, are highly related to colon cancer. The results are extensively validated by the somatic mutations identified from the same patients. In theory of latent factor, the goal is to find the similarity between the items in the same latent factor. Statistically, the percentage 0.833 (10/12) is far greater than 0.468 that is computed from the ratio between 81 mutation genes from DNA sequencing data of TCGA and 173 genes.

From GO analysis, it is obvious that the enriched cellular processes and molecular functions are important at many cancer studies [18][19]. Another finding is that the genes in Table 5 are almost from factor 1, as shown in Table 1. It illustrates the rationale of latent factor analysis. Therefore, we can call factor 1 as "Cell Cycle" factor.

In this paper, we aim to identify oncogenic genes from genomics data. Here, we implemented the method on the gene expression data. It is also feasible to identify oncogenic genes from other types of genomics data, such as copy number or other sequencing data. Additionally, we will add more genes into the model using protein-protein interaction (PPI) data, finding key genes of pathways related with colon by BFRM analysis.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (10971223, 11071252), Chinese Universities Scientific Fund (2012YJ130, 2011JS039).

References

[1] E. Huang, E. P. Black, H. Dressman, M. West, and J. R. Nevins, "Gene expression phenotypes of oncogenic signaling pathways," *Cell Cycle,* vol. 2, pp. 415-417, Sep-Oct 2003.

[2] E. Huang, *et al.*, "Gene expression phenotypic models that predict the activity of oncogenic pathways," *Nat Genet,* vol. 34, pp. 226-230, Jun 2003.

[3] E. P. Black, T. Hallstrom, H. K. Dressman, M. West, and J. R. Nevins, "Distinctions in the specificity of E2F function revealed by gene expression signatures," *Proc Natl Acad Sci U S A,* vol. 102, pp. 15948-15953, Nov 1 2005.

[4] R. Spang, *et al.*, "Prediction and uncertainty in the analysis of gene expression profiles," *In Silico Biol,* vol. 2, pp. 369-381, 2002.

[5] M. West, *et al.*, "Predicting the clinical status of human breast cancer by using gene expression profiles," *Proc Natl Acad Sci U S A,* vol. 98, pp. 11462-11467, Sep 25 2001.

[6] J. R. Nevins, *et al.*, "Towards integrated clinico-genomic models for personalized medicine: combining gene expression signatures and clinical factors in breast cancer outcomes prediction," *Hum Mol Genet,* vol. 12 Spec No 2, pp. R153-157, Oct 15 2003.

[7] J. Pittman, E. Huang, J. Nevins, Q. Wang, and M. West, "Bayesian analysis of binary prediction tree models for retrospectively sampled outcomes," *Biostatistics,* vol. 5, pp. 587-601, Oct 2004.

[8] C. M. Carvalho *et al.*, "High-Dimensional Sparse Factor Modeling: Applications in Gene Expression Genomics," *J Am Stat Assoc,* vol. 103, pp. 1438-1456, Dec 1 2008.

[9] J. T. Chang, et al., "A genomic strategy to elucidate modules of oncogenic pathway signaling networks," Mol Cell, vol. 34, pp. 104-114, Apr 10 2009.

[10] G. Jin*, C. Fu, et al.*, "A novel method of transcriptional response analysis to facilitate drug repositioning for cancer therapy," *Cancer Res,* vol. 72, pp. 33-44, Jan 1 2012.

[11] W.R Gilks., S. Richardson and D.J.Spiegelhalter, "Markov Chain Monte Carlo in Practice". *Chapman & Hall/CRC*, 1996.

[12] M. W. Omar Aguilar, "Bayesian Dynamic Factor Models and Portfolio Allocation," Journal of Business and Economic Statistics, vol. 18, pp. 338-357, 2000.

[13] G. Casella and E. I. George. "Explaining the Gibbs sampler". *The American Statistician*, vol. 46, pp. 167–174, 1992.

[14] B. Zhang, S. Kirov, J. Snoddy., "WebGestalt: an integrated system for exploring gene sets in various biological contexts," *Nucleic Acids Res,* vol. 33, pp. W741-748, Jul 1 2005.

[15] D.T.Duncan, N. Prodduturi, B. Zhang., "WebGestalt2: an updated and expanded version of the Web-based Gene Set Analysis Toolkit," BMC Bioinformatics, col. 11(Suppl 4), pp. 10, Jul 23 2010

[16] M. Ashburner, et al., "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium," Nat Genet, vol. 25, pp. 25-29, May 2000.

[17] A. Hamosh*, et al.*, "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders," *Nucleic Acids Res,* vol. 33, pp. D514-517, Jan 1 2005.

[18] A. C. Mita, et al., "Survivin: key regulator of mitosis and apoptosis and novel target for cancer therapeutics," Clin Cancer Res, vol. 14, pp. 5000-5005, Aug 15 2008.

[19] C. Catzavelos, et al., "Decreased levels of the cell-cycle inhibitor p27Kip1 protein: prognostic implications in primary breast cancer," Nat Med, vol. 3, pp. 227-30, Feb 1997.