Genetics and Molecular Research

IFGFA: Identification of featured genes from genomic data using factor analysis

C.H. Fu[1,2], S. Deng[2], J.H. Wu[3], X.Q. Wu[1], Z.H. Fu[2] and Z.G. Yu[1]

[1]School of Mathematics and Computational Science, Xiangtan University, Xiangtan, China
[2]School of Mathematics and System Science, Shenyang Normal University, Shenyang, China
[3]Foreign Language Department, Shenyang Normal University, Shenyang, China

Corresponding authors: Z.G. Yu / C.H. Fu
E-mail: yuzg@xtu.edu.cn / fuch@synu.edu.cn

Genet. Mol. Res. 15 (3): gmr.15038803
Received May 16, 2016
Accepted June 3, 2016
Published July 25, 2016
DOI http://dx.doi.org/10.4238/gmr.15038803

ABSTRACT. In this study, a software tool (IFGFA) for identification of featured genes from gene expression data based on latent factor analysis was developed. Despite the availability of computational methods and statistical models appropriate for analyzing special genomic data, IFGFA provides a platform for predicting colon cancer-related genes and can be applied to other cancer types. The computational framework behind IFGFA is based on the well-established Bayesian factor and regression model and prior knowledge about the gene from OMIM. We validated the predicted genes by analyzing somatic mutations in patients. An interface was developed to enable users to run the computational framework efficiently through visual programming. IFGFA is executable in a Windows system and does not require other

dependent software packages. This program can be freely downloaded at http://www.fupage.org/downloads/ifgfa.zip.

Key words: Factor analysis; Featured gene; Gene expression profile; Multivariate statistics

INTRODUCTION

With the development of biological research techniques and the lower cost of conducting genomics microarray analysis, a large number of gene expression profiles for genomic studies have been generated. These data can be downloaded from various data sets. For example, genomics data sets for many types of cancer can be obtained from TCGA (http://cancergenome.nih.gov/). However, it remains challenging to identify genes of interest among the large amount of available genomics data.

A number of methods based on gene expression profiles to identify genes of interest have been developed, such as the neural network method (Khan et al., 2001), signal to noise method (Golub et al., 1999), non-parametric test (Tusher et al., 2001; Ruan and Yuan, 2011) and other statistical methods (West et al., 2001; Veer et al., 2002; Spang et al., 2002; Nevins et al., 2003; Pittman et al., 2004). However, these studies were developed for specific types of gene expression profiling data and cannot reveal the underlying relationships between the genes.

Gene expression profiles often contain a portion of redundant genes that have little association with the cancer of interest in classification problems. The existence of redundant information increases the burden of the classification and decreases classification accuracy. Thus, it is necessary to separate featured genes from redundant genes using gene expression profiles to conduct classification. Typically, the standard method for featured gene selection, known as the gene signature, is a double sample *t*-test (Storey et al., 2007; Kim et al., 2008).

Factor analysis is a multivariate statistical method that can interpret variability among observed variables using a smaller number of unobserved variables. Similar to principle component analysis (Alter et al., 2000; Alter et al., 2003), factor analysis can be used to reduce the dimensionality of high-density DNA microarrays data at the probe level (Hochreiter et al., 2006). It has been used to identify signal pathway activation (Chang et al., 2009) and reposition of non-anticancer drugs for anticancer studies (Jin et al., 2012).

We previously developed a computational framework to predict featured genes for colon cancer (Fu et al., 2013). In contrast to generic latent factor analysis studies, the framework's idea is based on the hypothesis that observed variables with higher correlations can be categorized as the same factor in factor analysis. Thus, based on prior information that some genes are related to colon adenocarcinoma in OMIM (Amberger et al., 2015), the framework can predicate other important genes related to colon adenocarcinoma from the same latent factor. A well-established latent factor analysis model, Bayesian Factor and Regression Model (BFRM), which can implement a sparse statistical factor analysis model for high-dimensional multivariate data analysis, was applied in our framework (Carvalho et al., 2008).

In this study, we developed a software tool (IFGFA) based on our previous computational framework (Fu et al., 2013). To facilitate the use of our computational framework or implementation of BFRM by other researchers, we integrated the input parameters of the model, names of the data files, output evaluated factors, and their correlated genes in the visual interface of IFGFA. From the website, users can obtain outputs, including factors, featured

genes, and their correlation degree in a TXT file, without the need to struggle with parameters and procedures to implement the codes. Furthermore, IFGFA can be used for other studies requiring BFRM.

METHODS

We designed the framework to analyze colon adenocarcinoma data (Fu et al., 2013). The framework implemented a factor analysis model, BFRM, using the genomics data of patients to acquire a loading matrix and factor matrix. Next, based on known disease-related gene information or somatic mutation information, the program can select factors with the highest percentages of the prior information. These identified factors can be used for gene signature studies. Furthermore, our framework ranks the genes in each featured factor in terms of the absolute values of entries of the loading matrix and defines the top factors as featured genes. The framework process is shown in Figure 1.

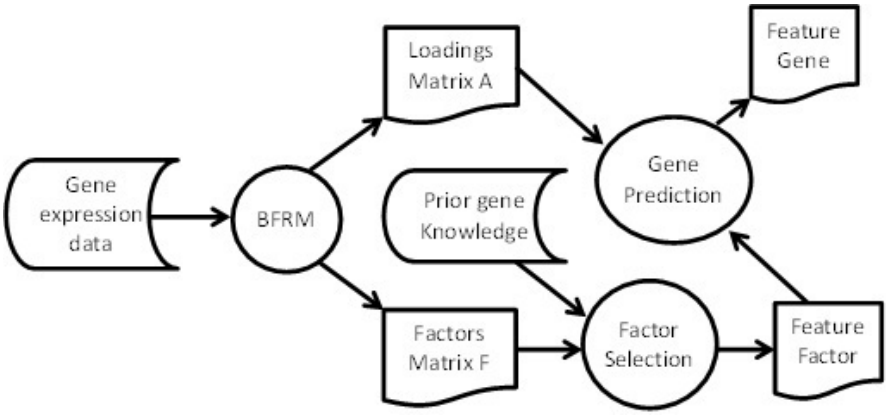

Figure 1. Flowchart of IFGFA framework.

Factor analysis

The simple continuous factor analysis model structure for X is as follows:

$$x_i = Af_i + u_i + e_i \ (i = 1 \ldots n) \qquad \text{(Equation 1)}$$

The vector f_i is "latent" because it is unobserved or called as unknown. A is the factor loading matrix, representing the degrees of latent factors correlated to genes. The vector e_i is a residual error of sample i and its distribution is assumed to follow a normal distribution. The vector u_i is an arbitrary offset vector. The model of equation 1 is referred to as "generative" because it describes how x_i is generated from f_i. If we use all x_i's as columns to form a matrix X and all f_i's as columns of a matrix F with suitably defined U and E, then we can rewrite equation 1 as:

$$X = AF + M + E \quad \text{(Equation 2)}$$

Factor analysis is used to decompose matrix *X*, implemented using BFRM (Carvalho et al., 2008). BFRM utilizes a Markov chain Monte Carlo (MCMC) process to simulate the matrices and their posterior probabilities. Next, Bayesian analysis and other computational methods allow this data to be used in high-dimensional multivariate data analysis.

Featured factor selection

Though factor analysis models can generate useful factors for disease classification, not all of these factors are important for the disease being examined. For example, factors 1, 3, 7, and 15 can be referred to as feature factors because they contain genes enriched in colon cancer as shown in Figure 2 (Fu et al., 2013). The feature factor is an extension of the general factor analysis in our study, making it easier to evaluate gene signatures or classify cancers. We identified these factors by evaluating the percentages of key genes among the total genes for a specific factor. This prior knowledge can be found in OMIM or other data sets.

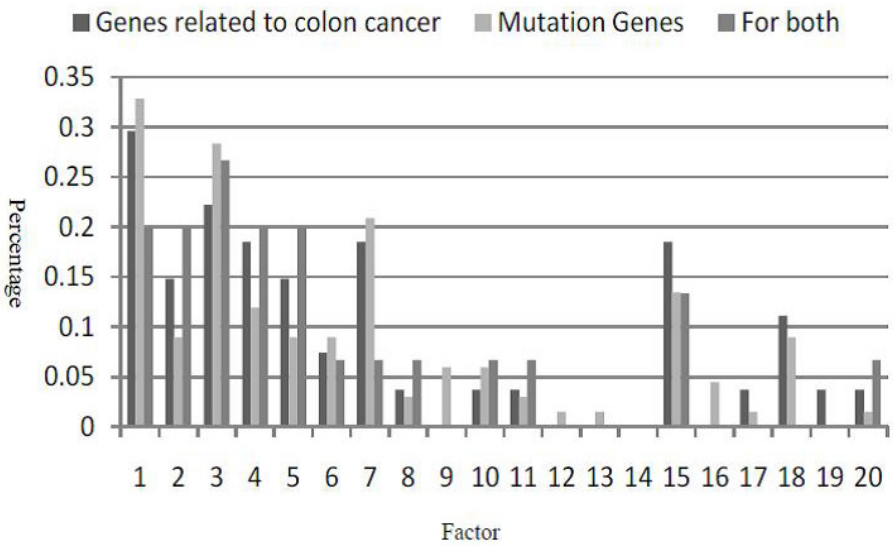

Figure 2. Percentage of genes across factors.

Featured gene prediction

Another function of our tool is that it can identify featured genes that may be correlated with the disease of interest. Based on the latent factor analysis hypothesis, genes from the same latent factor are more likely to possess similar biological characters, such as activation of a signaling pathway. In equation 1, the entries of the loadings matrix *A* represent the relevancy of genes correlated with factors. We ranked genes from the same factor in terms of the absolute values of entries of the loadings matrix. Thus, these top genes can be selected as featured genes of the disease.

RESULTS

Input

The key inputs of IFGFA are two flat text files. One includes gene information containing two columns: gene name and prior information regarding whether the gene is known to be related to the disease of interest. It is recommended to place the known gene on the top of the file, particularly in evolving mode. The second file is a gene expression data file in which the row order must correspond to the first file, with each row representing a variable (gene) and each column representing an observation (sample), with the two columns tab-separated. Every field is numeric and any other kind of data type is not allowed. In IFGFA, missing values in the dataset are indicated by a specific numeric value (such as 0 or 999), and another input file *XMaskFile* must be included which can be set in file *parameters.txt*. *XMaskFile* is a flat text file corresponding to the data file, with each field labeled as 0 for observed and 1 for missing values.

Before running the tool, a few key parameters must be set, as shown in the left panel of Figure 3. *Number of Samples* and *Number of Variables* are the actual numbers of patients and genes in the current run, which can be set to smaller or equal values as those in the data file. *Number of Factors* is the number of factors in the general factor analysis, representing the initial value for the parameter in the evolution model. *Number of Factors* can be set based on experience, and we suggest setting a number between 10 and 20 or using evolution mode for new users. *Evol or Not* must be set to 1 or 0, which indicates whether evolution mode is on. In the evolution model, the tool can estimate the number of factors, but a longer period of time is required to finish the process. *Number of EvolVars* is only necessary if *Evol or Not* is set to 1, which indicates the number of variables used to initialize the evolutionary analysis. *Time of MCMC* is the number of MCMC iterations, which determines the sampling time of Monte Carlo simulation methods (default is 5000). *Time of Burnin* is the number of burn-in iterations in the MCMC, out of which samples from iterations can be included in the final result (default is 2000). Larger burn-in times can eliminate the influence of priors. There are additional parameters included in the file *parameters.txt.* For details, one may require to refer to a previous study (Carvalho et al., 2008).

Output

The outputs of IFGFA are displayed in the right panel of Figure 3. Three selection parameters must be set. *Cutoff of PostPib* is a threshold of posterior probabilities and ensures that the elements of *A* have higher posterior estimation after MCMC iterations. *Number of TopFactor* is the number of factors with the highest percentages of disease-related genes across all factors. *Number of TopGene* is the number of genes with the highest absolute values of entries of *A*, implying the highest correlation with the factor. Figure 3 displays 7 factors in the viewing windows. For example, factor 9 has the highest percentage of 0.50 and includes 4 genes, among which *DLEC1* and *PRKAR1A* are not known to be related to colon cancer. Thus, IFGFA predicts that these genes are related to colon cancer, and thus factor 9 can be referred to as a colon cancer feature factor. Similar results were obtained for an additional 6 factors in this run. In the current directory, the text file *allFactors.txt* saves all factors and their associated genes in this run to allow users to evaluate these parameters in future studies.

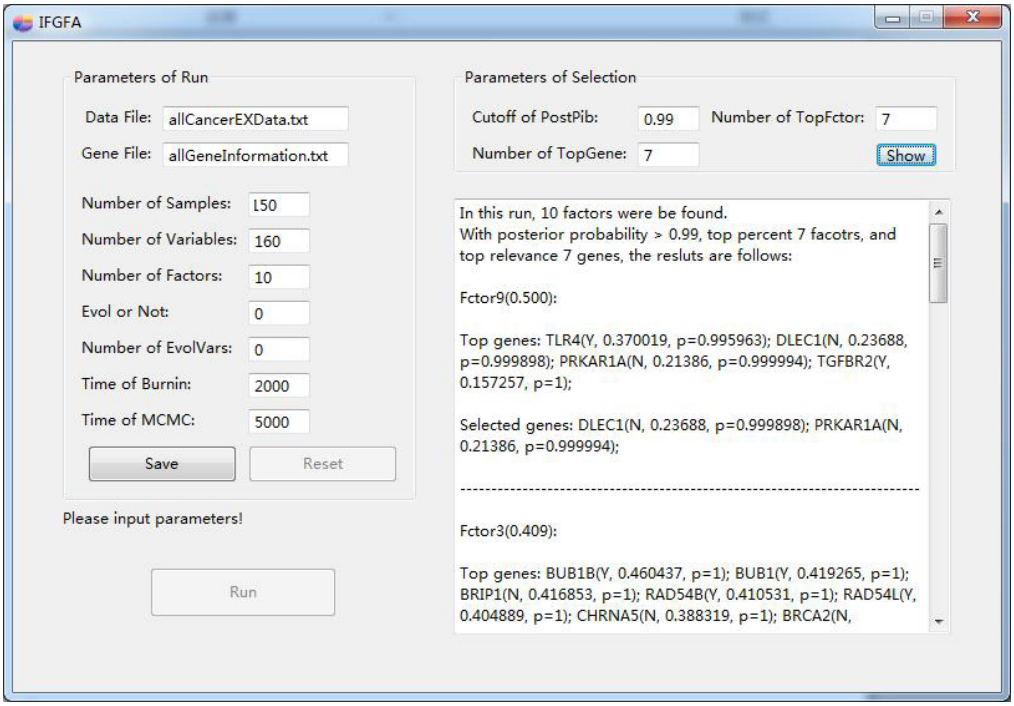

Figure 3. Screenshot of IFGFA.

Performance

The core program of IFGFA is a well-established latent factor model BFRM and is a Bayesian statistics model using MCMC methods to compute large hierarchical models. This program requires the integration of hundreds or even thousands of unknown parameters. The parameters increase with increasing numbers of genes and sample sizes. The running time of IFGFA for every 100 iterates is listed in Table 1 (CPU: Intel Core I7 Q820 1.73GHz, memory: 8 GB). Since it is difficult to identify this number of samples, sampling with replacement from gene expression profiles of colon cancer in the TCGA is conducted. When the sample size is not large, the tool can still provide meaningful results. Based on BFRM, Bild et al. identified important oncogenic pathway signatures in human cancers using only 97 DNA microarray samples (Bild et al., 2006).

Table 1. Running time of IFGFA at every 100 iterates (unit: s).

Gene No.	Sample No.						
	176	500	1000	2000	5000	10,000	17814
174	2.88	5.74	8.9	15.3	33.77	64.4	120.57
500	8.99	15.32	24.11	37.72	80.08	149.1	276.04
1000	32.45	44.49	59.16	87.58	170.1	309.71	567.55
1500	64.47	86.9	106.42	148.87	272.52	484.44	883.86

DISCUSSION

In our previous study (Fu et al., 2013), we showed that a factor analysis method with prior knowledge can be used to identify featured factors and featured genes from genomics data of colon cancer patients. Further GO analysis (Ashburner et al., 2000; Wang et al., 2013) indicated that the genes are meaningfully enriched in the functions of DNA repair and cell cycle, which contribute to the oncogenesis of colon adenocarcinoma. It is encouraging that one of the predicted genes, *RAD54L*, has now been annotated as related to colonic adenocarcinoma in OMIM (Matsuda et al., 1999).

IFGFA can be used not only to identify feature factors and feature genes related to a disease of interest, but also provide a visualization tool for factor analysis. Since the MCMC process is used, the outputs of IFGFA may be not precisely similar in all runs. Results should be evaluated for reliability by running the tool as many times as necessary. In the future, we will develop an online tool and add more functions.

ACKNOWLEDGMENTS

Support for the authors was provided by the Natural Science Foundation of Liaoning Province China (#2015020650) and the National Natural Science Foundation of China (#11371016 and #11201313).

REFERENCES

Alter O, Brown PO and Botstein D (2000). Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. USA* 97: 10101-10106. http://dx.doi.org/10.1073/pnas.97.18.10101

Alter O, Brown PO and Botstein D (2003). Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. *Proc. Natl. Acad. Sci. USA* 100: 3351-3356. http://dx.doi.org/10.1073/pnas.0530258100

Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, et al. (2015). OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.* 43: D789-D798. http://dx.doi.org/10.1093/nar/gku1205

Ashburner M, Ball CA, Blake JA, Botstein D, et al.; The Gene Ontology Consortium (2000). Gene ontology: tool for the unification of biology. *Nat. Genet.* 25: 25-29. http://dx.doi.org/10.1038/75556

Bild AH, Yao G, Chang JT, Wang Q, et al. (2006). Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* 439: 353-357. http://dx.doi.org/10.1038/nature04296

Chang JT, Carvalho C, Mori S, Bild AH, et al. (2009). A genomic strategy to elucidate modules of oncogenic pathway signaling networks. *Mol. Cell* 34: 104-114. http://dx.doi.org/10.1016/j.molcel.2009.02.030

Carvalho CM, Chang J, Lucas JE, Nevins JR, et al. (2008). High-Dimensional Sparse Factor Modeling: Applications in Gene Expression Genomics. *J. Am. Stat. Assoc.* 103: 1438-1456. http://dx.doi.org/10.1198/016214508000000869

Fu C, Deng S, Song Q and Jing L (2013). Latent factor analysis facilitates modelling of oncogenic genes for colon adenocarcinoma. *IET Syst. Biol.* 7: 165-169. http://dx.doi.org/10.1049/iet-syb.2012.0057

Golub TR, Slonim DK, Tamayo P, Huard C, et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286: 531-537. http://dx.doi.org/10.1126/science.286.5439.531

Hochreiter S, Clevert DA and Obermayer K (2006). A new summarization method for Affymetrix probe level data. *Bioinformatics* 22: 943-949. http://dx.doi.org/10.1093/bioinformatics/btl033

Jin G, Fu C, Zhao H, Cui K, et al. (2012). A novel method of transcriptional response analysis to facilitate drug repositioning for cancer therapy. *Cancer Res.* 72: 33-44. http://dx.doi.org/10.1158/0008-5472.CAN-11-2333

Khan J, Wei JS, Ringnér M, Saal LH, et al. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat. Med.* 7: 673-679. http://dx.doi.org/10.1038/89044

Kim SB, Chen VCP, Park Y, Ziegler TR, et al. (2008). Controlling the False Discovery Rate for Feature Selection in High-

resolution NMR Spectra Statistical Analysis and Data Mining: The ASA Data Science Journal Volume 1, Issue 2. *Statistical Analysis & Data Mining* 1: 57-66. http://dx.doi.org/10.1002/sam.10005

Matsuda M, Miyagawa K, Takahashi M, Fukuda T, et al. (1999). Mutations in the RAD54 recombination gene in primary cancers. *Oncogene* 18: 3427-3430. http://dx.doi.org/10.1038/sj.onc.1202692

Nevins JR, Huang ES, Dressman H, Pittman J, et al. (2003). Towards integrated clinico-genomic models for personalized medicine: combining gene expression signatures and clinical factors in breast cancer outcomes prediction. *Hum. Mol. Genet.* 12: R153-R157. http://dx.doi.org/10.1093/hmg/ddg287

Pittman J, Huang E, Nevins J, Wang Q, et al. (2004). Bayesian analysis of binary prediction tree models for retrospectively sampled outcomes. *Biostatistics* 5: 587-601. http://dx.doi.org/10.1093/biostatistics/kxh011

Ruan L and Yuan M (2011). An empirical Bayes' approach to joint analysis of multiple microarray gene expression studies. *Biometrics* 67: 1617-1626. http://dx.doi.org/10.1111/j.1541-0420.2011.01602.x

Spang R, Zuzan H, West M, Nevins J, et al. (2002). Prediction and uncertainty in the analysis of gene expression profiles. *In Silico Biol.)* 2: 369-381.

Storey JD, Dai JY and Leek JT (2007). The optimal discovery procedure for large-scale significance testing, with applications to comparative microarray experiments. *Biostatistics* 8: 414-432. http://dx.doi.org/10.1093/biostatistics/kxl019

Tusher VG, Tibshirani R and Chu G (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA* 98: 5116-5121. http://dx.doi.org/10.1073/pnas.091062498

Veer LJ, Dai H, van de Vijver MJ, He YD, et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415: 530-536. http://dx.doi.org/10.1038/415530a

West M, Blanchette C, Dressman H, Huang E, et al. (2001). Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc. Natl. Acad. Sci. USA* 98: 11462-11467. http://dx.doi.org/10.1073/pnas.201162998

Wang J, Duncan D, Shi Z and Zhang B (2013). WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. *Nucleic Acids Res.* 41: W77-W83.