GEPS: the Gene Expression Pattern Scanner

Yu-Peng Wang¹, Liang Liang¹, Bu-Cong Han¹, Yu Quan¹, Xiao Wang¹, Tao Tao¹ and Zhi-Liang Ji¹,²,*

¹Key Laboratory for Cell Biology and Tumor Cell Engineering, the Ministry of Education of China, School of Life Sciences Xiamen University, Xiamen 361005, Fujian, People's Republic of China and ²The Key Laboratory for Chemical Biology of Fujian Province Xiamen University, Xiamen 361005, Fujian Province, People's Republic of China

Received January 17, 2006; Revised January 31, 2006; Accepted March 1, 2006

ABSTRACT

Gene Expression Pattern Scanner (GEPS) is a web-based server to provide interactive pattern analysis of user-submitted microarray data for facilitating their further interpretation. Putative gene expression patterns such as correlated expression, similar expression and specific expression are determined globally and systematically using geometric comparison and correlation analysis methods. These patterns can be visualized via linear plot with quantitative measures. User-defined threshold value is allowed to customize the format of the pattern search results. For better understanding of gene expression, patterns derived from 329 205 non-redundant gene expression records from the GNF SymAltas and the Gene Expression Omnibus are also provided. These profiles cover 24 277 human genes in 79 tissues, 32 905 mouse genes in 61 tissues and 4201 rat genes in 44 tissues. GEPS is available at http://bioinf.xmu.edu.cn/software/geps/geps.php.

INTRODUCTION

Microarray technologies have been popularly used in the identification of gene expression patterns associated with physiological or pathological states on genome scale (1,2). With their rapidly increasing use in the study of gene function, transcriptional regulation, disease etiology and drug development study of genes/proteins (1–4), a significant challenge has emerged on how to manage the overwhelming amount of transcription data generated by individual gene microarrays. Since inferring function of genes based on direct observation or simple statistical analysis of their expression profiles is both unreliable and arduous, bioinformatics tools have been developed to facilitate data analysis and interpretation (5–9). In many cases, the annotation of genes is assigned automatically using some clustering-based programs, such as GEPAS (7), DNMAD (5), MIDAW (9) and GEMS (6). Such assignments of gene functions are made by discovering the coherent expression patterns. Apart from clustering-based methods, some integrative systems employ various analysis tools such as principal component analysis, supervised classification including feature selection and cross-validation, multi-factorial ANOVA to provide wide range of data analysis (7,8,10,11). The high-level interpretation of data by mapping expression profiles onto currently available regulatory, metabolic and cellular pathways has also been reported (4).

The interpretation of microarray data depends on successful selection of the consensus gene expression patterns such as correlated expression, differential expression and specific expression. These patterns are normally determined by mining gene expression profiles using different algorithms described above. Gene Expression Pattern Scanner (GEPS) is such kind of platform constructed primarily on the basis of systematic and global analysis of the gene expression patterns. One of the advantages of GEPS is the fact that the putative gene expression patterns are identified by comparing the global performance of gene expressions, thus the derived patterns may more properly reflect the true behavior of gene expression. Another advantage is that the relationships of a gene with others can be optionally listed in a descent order according to respective measures, which enables systematic study of gene expression at quantitative as well as qualitative levels. Moreover, besides of the user-submitted data, a number of public gene expression data are also provided to facilitate better understanding of gene expression behaviors.

METHODS

The data of GEPS

GEPS allows users to submit their individual normalized gene expression datasets to the system by calling an underlying dynamic CGI program. The data can be uploaded locally to remote server as a tab-delimited plain text file (‘.txt’ or Gene Expression Omnibus, GEO ‘.soft’ format), or a compressed ‘.gz’ format file in cases of the internet traffic problems. The
format of the dataset is similar to the commonly used format in gene expression datasets: The first column entitled ‘ID_REF’ contains the unique ID for each gene or probeset, which is also used for browsing the analysis results. The second column entitled ‘IDENTIFIER’ is the description (e.g. gene name) of each gene or probeset. The following columns are the expression data. The first row contains the names of each column, while other rows are the expression data with one row per probeset. Null or space is not allowed in any value of the data, which should be replaced by ‘0’ or underscore ‘_’, respectively. In the same row, continued columns with same name will be merged and represented by the average value of their value during data analysis.

In additional to user submitted data, GEPS also provides the pre-scanning patterns of public datasets for better understanding of gene expression. The public datasets come from two important gene expression repositories: the GNF Atlas (http://symatlas.gnf.org/SymAtlas/) (12) and the GEO (http://www.ncbi.nih.gov/geo/) (13). Currently, 19 distinct datasets of 329,205 non-redundant gene expression profiles from GNF SymAltas and GEO are deposited in GEPS, which covers about 24,277 human genes in over 79 tissues, about 32,905 mouse genes in over 61 tissues and about 4,201 rat genes in 44 tissues.

The scanning of gene expression patterns
To initiate the pattern scanning, each gene expression profile is transformed into a vector $X$:

$$X = (x_1, x_2, x_3, \ldots, x_n),$$

where $x_i$ is the gene expression level over tissues, time scale or other conditions and $n$ is the number of tissues or time slots. The pattern scanning is demonstrated in three methods: similarity measure (SM), correlated analysis and specificity measure (SPM). The SM evaluates the geometric similarity between two gene expression profiles in high dimension vector space, which is given by the following equation.

$$SM(\cos \theta) = \frac{X \cdot Y}{|X||Y|},$$

where $\theta$ is the angle of two vectors $X$ and $Y$, $|X|$ and $|Y|$ are the lengths of vectors $X$ and $Y$, respectively. $SM$ ranges from 0 to 1. The correlation of two profiles $X$ and $Y$ can be indicated by the coefficient $r$, which is decided by the following equation.

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n \bar{x} \bar{y}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n \bar{x}^2} \sqrt{\sum_{i=1}^{n} y_i^2 - n \bar{y}^2}},$$

$r$ ranges from $-1$ to 1. $\bar{x}$ is the mean of gene expression levels. SPM is calculated to assess the specificity or abundance of gene expression in tissues. The SM is decided by the following equation.

$$SPM(\cos \alpha) = \frac{x_i}{|X|},$$

where $\alpha$ is the angle between vector and sample axis (either tissue or time) in high dimension sample space, $x_i$ is the expression level in sample $i$, and $|X|$ is the length of vector $X$.

---

Figure 1. The homepage of GEPS.
The interpretation of GEPS
For better interpretation of biological knowledge hidden in the vast volume of data, a gene expression profile can be treated as a distribution curve (a vector during calculation) with respect to tissues, time or other conditions. Comparison between two distributions will be helpful for the identification of the gene expression patterns globally. Geometric comparison (SM) is used to indicate how similar two distribution curves are. A value of SM close to 1 means the high similarity of two distributions. This hints that these two genes may have similar expression patterns regardless of their expression levels. It may be further interpreted that these genes likely play a similar role in biological processes. However, similar expression patterns do not mean that these two gene pairs are related. Correlated analysis is thus demonstrated to tell whether the expression of two genes is correlated. A value of correlated coefficient \( r \) close to 1 or \(-1\) concludes the high correlation of two distributions statistically, while co-expression (close to 1) or inverse-expression (close to \(-1\)) in biological extent. Such correlation further infers that these two genes may have interaction with each other or they are functionally associated proteins (14,15). Tissue-specific expression is very helpful for the understanding of the physiological behavior of a gene. In many cases, the uncertainty of tissue specific genes is due to the short of quantitative measure. In this study, SPM is determined to illustrate how specificity (a value close to 1) of a gene is expressed in a tissue comparing with others. This measure can also be used to differentiate the expression of genes in varied conditions.

Figure 2. The interactive search interface.
Access of GEPS

The GEPS can be freely accessed at http://bioinf.xmu.edu.cn/software/geps/geps.php. To initiate the interactive data analysis, user is required to either provide a previously assigned 6-digit file ID or upload a new dataset to the GEPS server (Figure 1). For new submitted data, user is also requested to select a data type, either count value or log ratio, to continue the analysis. An interactive search interface is generated once the data is successfully uploaded, as well a unique 6-digit ID is assigned to user for future access (Figure 2). GEPS mainly provides three ways for data query: Search patterns for genes, Compare genes and Search specific-expression genes in samples. Through the ‘Search patterns’ form, user is enabled to search expression patterns of a designated gene (represented by the probeset_ID in column ‘ID_REF’) or several genes at one time. Flexible threshold values for different measures are allowed to personalize the query. Probesets satisfying the query criteria are listed separately in three sections: co-expression, inverse-expression and similar expression (Figure 3). Through the ‘Compare genes’ form, user is allowed to compare the expression patterns between multiple genes simultaneously. The comparison results are indicated in a matrix and differentiated in colors (Figure 4). Through the ‘Search specific-expression genes’ form, user is able to browse genes that specifically expressed in designated samples (e.g. tissues or conditions). Probesets satisfying the query criteria are listed in a descending order based on the value of SPM.

In all cases, clicking on a probeset_ID will lead user into the detailed information page. In the detailed information page, analysis results are summarized and visualized in charts (Figure 5). Comments on the results are also made following the rules: a value of SM >0.80 and 0.95 is interpreted as medium similar expression and highly similar expression respectively in this study. A value of correlated coefficient r more than (less than for inverse-expression) 0.75 (−0.60) and 0.90 (−0.80) is considered as medium co-expression (inverse-expression) or highly co-expression (inverse-expression), respectively. A value of SPM >0.90 and 0.99 is taken as highly abundant expression and specific expression, respectively.

CONCLUSION REMARKS

The GEPS is a user-friendly platform for statistical analysis of gene expression patterns. The service of GEPS is real-time and interactive, which allows users to submit data to remote server and manage the analysis results locally. The introduction of a serial of measures enable a user to quantitatively assess the analysis results, based on which preliminary interpretation of the data is also given. The results are also visualized in compact curve charts for better understanding and interpretation of the results. However, efforts have been continuously made to improve the service in such aspects as the identification of local patterns, relationship analysis of...
ACKNOWLEDGEMENTS

This work is supported by following grants: a grant (to Z.L. Ji) from the Program for New Century Excellent Talents in Xiamen University, grants (#30400573 to Z.L.J. and #3047085 to T.T.) from the National Natural Science Foundation of China, a grant (#2004BA711A19-07 to T.T.) from the Ministry of Science and Technology, China, a grant (#C0510003 to T.T.) from the Natural Science Foundation of Fujian Province, a grant (#2005-383 to T.T.) from the Ministry of Education of China and a starting fund (#XK0014 to T.T.) from Xiamen University. Funding to pay the Open Access publication charges for this article was provided by NSFC #30400573.

Conflict of interest statement. None declared.

REFERENCES

1. Chi, J.T., Chang, H.Y., Haraldsen, G., Jahnsen, F.L., Troyanskaya, O.G., Chang, D.S., Wang, Z., Rockson, S.G., van de Rijn, M., Botstein, D. et al. (2003) Endothelial cell diversity revealed by global expression profiling. Proc. Natl Acad. Sci. USA, 100, 10623–10628.
2. Chung, C.H., Bernard, P.S. and Perou, C.M. (2002) Molecular portraits and the family tree of cancer. Nature Genet., 32, 533–540.
3. van Steensel, B. (2005) Mapping of genetic and epigenetic regulatory networks using microarrays. Nature Genet., 37, S18–24.
4. Mlecnik, B., Scheideler, M., Hackl, H., Hartler, J., Sanchez-Cabo, F. and Trajanoski, Z. (2005) PathwayExplorer: web service for visualizing high-throughput expression data on biological pathways. Nucleic Acids Res., 33, W633–W637.
5. Vaquerizas, J.M., Dopazo, J. and Diaz-Uriarte, R. (2004) DNMAD: web-based diagnosis and normalization for microarray data. Bioinformatics, 20, 3656–3658.
6. Wu, C.J. and Kasif, S. (2005) GEMS: a web server for biclustering analysis of expression data. Nucleic Acids Res., 33, W596–W599.
7. Vaquerizas, J.M., Conde, L., Yankilevich, P., Cabezón, A., Mínguez, P., Díaz-Uriarte, R., Al-Shahrour, F., Herrero, J. and Dopazo, J. (2005) GEPAS, an experiment-oriented pipeline for the analysis of microarray gene expression data. Nucleic Acids Res., 33, W616–W620.
8. Shamir, R., Maron-Katz, A., Tanay, A., Linhart, C., Steinfeld, I., Sharan, R., Shiloh, Y. and Elkon, R. (2005) EXPANDER—an integrative program suite for microarray data analysis. *BMC Bioinformatics*, 6, 232.

9. Romualdi, C., Vitulo, N., Del Favero, M. and Lanfranchi, G. (2005) MIDAW: a web tool for statistical analysis of microarray data. *Nucleic Acids Res.*, 33, W644–W649.

10. Psarros, M., Heber, S., Sick, M., Thoppae, G., Harshman, K. and Sick, B. (2005) RACE: remote analysis computation for gene expression data. *Nucleic Acids Res.*, 33, W638–W643.

11. Theilhaber, J., Ulyanov, A., Malanthara, A., Cole, J., Xu, D., Nahf, R., Heuer, M., Brockel, C. and Bushnell, S. (2004) GECKO: a complete large-scale gene expression analysis platform. *BMC Bioinformatics*, 5, 195.

12. Su, A.I., Cooke, M.P., Ching, K.A., Hakak, Y., Walker, J.R., Wiltshire, T., Orth, A.P., Vega, R.G., Sapinoso, L.M., Moqrich, A. et al. (2002) Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl Acad. Sci. USA*, 99, 4465–4470.

13. Barrett, T., Suzek, T.O., Troup, D.B., Wilhite, S.E., Ngau, W.C., Ledoux, P., Rudnev, D., Lash, A.E., Fujibuchi, W. and Edgar, R. (2005) NCBI GEO: mining millions of expression profiles—database and tools. *Nucleic Acids Res.*, 33, D562–D566.

14. Jansen, R., Greenbaum, D. and Gerstein, M. (2002) Relating whole-genome expression data with protein–protein interactions. *Genome Res.*, 12, 37–46.

15. Grigoriev, A. (2001) A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast *Saccharomyces cerevisiae*. *Nucleic Acids Res.*, 29, 3513–3519.