BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 11 2012, pages 1544–1545
doi:10.1093/bioinformatics/bts169

Gene expression

Advance Access publication April 6, 2012

PaGeFinder: quantitative identification of spatiotemporal pattern genes

Jian-Bo Pan[1,2], Shi-Chang Hu[3], Hao Wang[1], Quan Zou[3,*] and Zhi-Liang Ji[1,2,*]

[1]Department of Chemical Biology, College of Chemistry and Chemical Engineering, The Key Laboratory for Chemical Biology of Fujian Province, [2]State Key Laboratory of Stress Cell Biology, School of Life Sciences, and [3]School of Information Science and Technology, Xiamen University, Xiamen 361005, Fujian, P R China

Associate Editor: Trey Ideker

ABSTRACT

Summary: Pattern Gene Finder (PaGeFinder) is a web-based server for on-line detection of gene expression patterns from serial transcriptomic data generated by high-throughput technologies like microarray or next-generation sequencing. Three particular parameters, the specificity measure, the dispersion measure and the contribution measure, were introduced and implemented in PaGeFinder to help quantitative and interactive identification of pattern genes like housekeeping genes, specific (selective) genes and repressed genes. Besides the on-line computation service, the PaGeFinder also provides downloadable Java programs for local detection of gene expression patterns.

Availability: http://bioinf.xmu.edu.cn:8080/PaGeFinder/index.jsp

Contact: appo@xmu.edu.cn; zouquan@xmu.edu.cn

Received on January 2, 2012; revised on February 29, 2012; accepted on April 2, 2012

1 INTRODUCTION

Spatiotemporal variation of gene expression can happen extensively among tissues, developmental stages, physiological conditions and individuals (Lage *et al.*, 2008). The variation is believed to link with gene function and pathology. Benefiting from current applications of high-throughput technologies, e.g. microarray and next-generation sequencing (NGS), simultaneously monitoring gene differential expressions in large scale becomes easier. When digging into these large volume of data, patterns can be detected.

Currently, three kinds of pattern genes, housekeeping genes, specific/selective genes and repressed genes, have received general attentions. Housekeeping genes are generally defined as genes that express ubiquitously in all conditions, which have been adopted as molecular markers in qualitatively or semi-quantitatively measuring gene expression level for a long time (Warrington *et al.*, 2000). The specific (selective) genes are a group of genes whose expressions are enriched in one or several conditions, e.g. tissues, or developmental stages (Liang *et al.*, 2006). Opposite to the specific gene expression, some genes are expressed in almost all conditions except in one or several conditions. These genes are acknowledged as repressed genes or 'disallowed genes' (Thorrez *et al.*, 2011). They are exceptions of housekeeping genes. The spatiotemporal preference of these pattern genes carries crucial information of what the genes do and how they work together to execute certain physiological functions.

Traditionally, these pattern genes were detected by molecular technologies like RT-PCR, *in situ* hybridization etc. However, due to the limitation of technologies, many pattern genes identified by these methods were later found to be inappropriate when including more samples. This problem was significantly reduced with availability of large scale datasets generated by microarray, SAGE or NGS. Upon these high-throughput data, various methods were adopted previously on detecting such pattern genes, including cutoff, relative fraction (Chang *et al.*, 2011; Liu *et al.*, 2008) and learning algorithms like Naive Bayes classifier (De Ferrari and Aitken, 2006) and SVM (Dong *et al.*, 2011). Some of them are simple but qualitative (e.g. cutoff); some are quantitative but insensitive (e.g. relative fraction); some are powerful but instable and hard to be implemented (e.g. learning algorithms). Therefore in this study, we introduced three novel parameters as quantitative indicators to describe and automatically identify pattern genes from serial transcriptomic data. An on-line server was constructed as well to provide dynamic analysis service.

2 METHODS

2.1 SPM and identification of specific gene

To quantitatively estimate the relative expression specificity of a gene in a sample, the specificity measure (SPM) was introduced as following. Each gene expression profile was first transformed into a vector X:

$$\mathrm{X}=(x_1,x_2,\ldots,x_i,\ldots,x_{n-1},x_n) \quad (1)$$

where n is the number of samples in a profile. At the same time, a vector X_i was created to represent the gene expression in sample i:

$$\mathrm{X}_i=(0,0,\ldots,x_i,\ldots,0,0) \quad (2)$$

The SPM of a gene in a sample was then determined by calculating the cosine value of intersection angle θ between vector X_i and X in high-dimension feature space:

$$\mathrm{SPM}_i=\cos\theta=\frac{\mathrm{X}_i\bullet\mathrm{X}}{|\mathrm{X}_i|\bullet|\mathrm{X}|} \quad (3)$$

where $|\mathrm{X}_i|$ and $|\mathrm{X}|$ are the length of vector X_i and X, respectively. The value of SPM ranges from 0 to 1.0. A SPM value close to 1.0 indicates the major contribution of gene expression in a designated sample (e.g. vector X_i) against that in all samples (vector X). The higher the SPM value is (e.g. SPM $\geq$ 0.9), the more specific the gene expression is in a sample.

*To whom correspondence should be addressed.

2.2 DPM and identification of housekeeping gene and repressed gene

To evaluate the variability and diversity degree of a gene expression profile, a new parameter of dispersion measure (DPM) was introduced as following. The gene expression profile (X) was first converted to its corresponding SPM profile (X_{SPM}):

$$X_{SPM} = (SPM_1, SPM_2, \ldots, SPM_i, \ldots, SPM_{n-1}, SPM_n) \quad (4)$$

The DPM was then determined by

$$DPM = \sqrt{\frac{\sum_{i=1}^{n}(SPM_i - \overline{SPM})^2}{n-1}} \cdot \sqrt{n} \quad (5)$$

where n is the sample number, and $\overline{SPM}$ is the mean of SPMs in a gene expression profile. Unlike conventional SD analysis, DPM is independent of gene expression level and sample number by scaling into a region of 0–1.0 as above. In this way, DPM makes variability comparable between profiles or datasets. A value of DPM close to 0 suggests equal expressions of gene over samples. Therefore, DPM can serve as a good indicator in quantitative description and identification of 'strict' housekeeping genes that have nearly constant expression in all samples, e.g. DPM ≤ 0.3. As exceptions of housekeeping genes, the repressed genes are detected by verifying gene expressions in all samples except one.

2.3 CTM and identification of selective gene

The contribution measure (CTM) is a statistical parameter developed to measure the enrichment of gene expressions in several samples. The CTM was calculated by

$$CTM = \sqrt{\sum_{i=1}^{k} SPM_i^2} \quad (6)$$

where k is the number of selected samples. In this study, the tissue-selective genes were described and identified as genes whose expressions are enriched in limited samples (e.g. $2 \leq k \leq 4$), in each of samples ($SPM_i \geq 0.4$) and together (e.g. CTM ≥ 0.9 and DPM ≥ 0.9).

3 ACCESS OF PaGeFinder

The PaGeFinder can be freely accessed at http://bioinf.xmu.edu.cn:8080/PaGeFinder/index.jsp. To initiate the interactive data analysis, user is required to upload a local pre-processed gene expression dataset to the remote server. The dataset should be prepared in tab-delimited format as following: the first row contains titles of each column. The first column contains unique identifiers (normally probeset IDs or gene symbols) for genes, which will be used to query or browse the analysis results. The following rows and columns are expression data of samples. Currently, the server only accepts tab-delimited plain text file or its compressed '.zip' file, which cannot exceed 10 Mbits in file size. After successful file uploading, data validation function is called to check for missing data or improper values. If the dataset passes the validation, the server will respond the statistic of valid rows (genes) and columns (samples); otherwise, prompt error messages.

When file is uploaded and validated successfully, an optional expression cutoff value is asked as an indicator of gene absence/presence for further data normalization. Clicking on the button 'continue' will lead to the query page, where analysis results can be downloaded at the right top corner of page or browsed by three different ways: Gene Search, Pattern Gene Search and Pattern Search. When query a designated gene via the 'Gene Search' form, its normalized expression profile, SPM distribution and its DPM evaluation will be shown. The 'Pattern Gene Search' form enables user to retrieve information of specific genes, housekeeping genes, selective genes and repressed genes through four independent sub-forms. The query starts by setting proper cutoffs for detecting these four pattern genes. A set of default cutoffs have been preset for convenience; however, user can freely customize results by modifying cutoff values in respective forms. Query submission will respond a sorted list of genes (identifiers) that satisfy query criteria. Clicking on an identifier will lead to the detailed information page, where the gene patterns can be visualized in charts as well as quantitative measures. The 'Pattern Search' form provides functions for detecting two global gene expression patterns, similarity and correlation analyses, which was previously implemented in the GEPS sever (Wang *et al.*, 2006).

For those large datasets, the PaGeFinder even provides downloadable Java programs for local analysis. Currently, three standalone Java programs for SPM/DPM calculation, similarity calculation and correlation calculation are available.

4 CONCLUSION

In summary, the introduction of three novel parameters in PaGeFinder provides an easier, more sensitive and robust way in quantitative detection of gene expression patterns than current methods like cutoff and relative fraction. PaGeFinder is particularly useful for dynamic and global understanding of gene functions under serial spatiotemporal conditions. Moreover, it also can be widely applied on mining other high-throughput data based on protein, metabolite or other molecule systems.

Funding: Fundamental Research Funds for the Central Universities (#2010121084) and Natural Science Foundation of China (NSFC#30873159).

Conflict of Interest: none declared.

REFERENCES

Chang,C.W. *et al.* (2011) Identification of human housekeeping genes and tissue-selective genes by microarray meta-analysis. *PLoS One*, **6**, e22859.

De Ferrari,L. and Aitken,S. (2006) Mining housekeeping genes with a Naive Bayes classifier. *BMC Genomics*, **7**, 277.

Dong,B. *et al.* (2011) Predicting housekeeping genes based on Fourier analysis. *PLoS One*, **6**, e21012.

Lage,K. *et al.* (2008) A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. *Proc. Natl Acad. Sci. USA*, **105**, 20870–20875.

Liang,S. *et al.* (2006) Detecting and profiling tissue-selective genes. *Physiol. Genomics*, **26**, 158–162.

Liu,X. *et al.* (2008) TiGER: a database for tissue-specific gene expression and regulation. *BMC Bioinformatics*, **9**, 271.

Thorrez,L. *et al.* (2011) Tissue-specific disallowance of housekeeping genes: the other face of cell differentiation. *Genome Res.*, **21**, 95–105.

Wang,Y.P. *et al.* (2006) GEPS: the gene expression pattern scanner. *Nucleic Acids Res.*, **34**, W492–W497.

Warrington,J.A. *et al.* (2000) Comparison of human adult and fetal expression and identification of 535 housekeeping/maintenance genes. *Physiol. Genomics*, **2**, 143–147.