PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2009, 36(11): 1415~1422
www.pibb.ac.cn

p-SAGE: Parametric Statistical Analysis of Gene Sets*

HUANG Bo**, LI Wen-Ting**, LI Wen, XIA Xue-Feng, SUN Zhi-Rong***
(*Institute of Bioinformatics and Systems Biology, MOE Key Laboratory of Bioinformatics, State Key Laboratory of BioMembrane and Membrane Biotechnology, Department of Biological Sciences and Biotechnology, Tsinghua University, Beijing* 100084, China)

Abstract Tumor genesis and development often result from deregulation of important biological pathways at the gene expression level. Although there has been much work focused on searching gene sets using gene expression data or other prior information, proper statistical testing of the gene sets is still an open question. Most studies have expanded the testing method of a single gene into the gene sets. Parametric statistical analysis of gene sets (p-SAGE) was presented for determining the significant gene sets or pathways associated with a phenotype of interest. The method was applied to brain tumor experiments to identify many gene sets. Some of the newly discovered gene sets were related to signal transduction and immunity. This simple and effective method gives useful biologically meaningful results.

Key words parametric statistical method, gene sets, deregulated
DOI: 10.3724/SP.J.1206.2009.00325

Since microarrays were introduced, much work has focused on identifying individual genes that exhibit differential expression between two phenotypes from the whole genome in cancer research. Various statistical tests are used to estimate the significance of the differential expressions[1]. However, with a long list of identified candidate genes, the main difficulty in the analysis is how to interpret these differentially expressed genes.

Since most biological processes involve complex interactions and the regulation of multiple genes, it is better that we look through groups of genes with similar biological meanings instead of discrete individual genes. Efforts, in fact, have been made in search of significant gene sets or pathways in recent decades. A knowledge-based approach for interpreting genome-wide expression profiles, Gene Set Enrichment Analysis (GSEA)[2], was one of the first to carry out this idea. Every predefined set of genes is assigned a score calculated as the average of the test statistics of its member genes. Many studies have applied known pathway information to the searching for gene sets within this framework[3].

Besides, other prior knowledge, such as the protein-protein interactions (PPI), has also been used. Ideker *et al.*[4] first performed the analysis based on a PPI network rather than predefined gene sets. They screen the network to identify active subnetworks, by assigning to them scores that resemble previous methods and picking subnetworks with the highest scores.

Those methods give a new view for microarray data analysis. They all share the common procedure: define or identify gene sets and score them with a test statistic; search highly-scored gene sets and predict which of them are related to the response or category of the sample.

As for the search of highly-scored gene sets, many attempts have been made to develop effective

*This work was supported by Hi-Tech Research and Development Program of China (2006AA020403), National Basic Research Program of China (2009CB918801) and The National Natural Science Foundation of China (30770498).

**These authors contributed equally to this work.
***Corresponding author.
Tel: 86-10-62772237, E-mail: sunzhr@mail.tsinghua.edu.cn

Received: May 15, 2009 Accepted: August 25, 2009

algorithms. The significance testing of gene sets is mainly based on two ideas: (1) The one is to average the significant levels of the genes, denoted by ΣT[5, 6]. (2) The other is to first average the gene expression levels and then test the statistical significance, denoted by $T\Sigma$[3, 7].

Great progress has been made in this field, however, these two methods have failed to solve some problems. Both just identify gene sets in which some genes have been up- or down-regulated in the same manner, but they fail to discover pathways containing both up-regulated and down-regulated genes. There are many cases where a transcription factor can simultaneously up-regulate some genes and down-regulate others. The overall mean value in such cases can be unchanged or only slightly changed. The current statistical methods fail to address this issue.

This work presents a parametric statistical procedure to identify deregulated gene sets, which can more accurately calculate the significance of gene sets and discover the gene sets associated with a phenotype of interest. This method, called p-SAGE, overcomes some shortages. Firstly, information loss of the statistical result in traditional approaches, due to averaging scores, is here corrected by constructing metric scores between phenotypes for each gene in a set. Secondly, the correlation structure of the gene sets no longer affects the results since it is not a concern whether the genes in a set are correlated in this method. Finally, in comparison to previous methods, the method is more suitable for finding sets in which the genes are simultaneously distributed into two tails.

1 Materials and methods

1.1 Data

1.1.1 Data preparation. The microarray data were collected from the NIH Gene Expression Omnibus (GDS1813).Here *SNR* (signal to noise ratio) was used as the difference metric scores in an initial analysis. For simplicity, the *SNR* was calculated for each gene as:

$$SNR=\frac{\mu_A-\mu_B}{\sigma_A+\sigma_A} \quad (1)$$

where μ_A and σ_A are the mean and standard deviation for each probe in phenotype A, with μ_B and σ_B for phenotype B.

1.1.2 Gene set selection. This study used three types of sets for the analysis. C2 Sets from MSigDB were used, which were created from several sources including online pathway databases and the biomedical literature (20 cancer related pathways from the NetPath database were added to the C2 sets). In addition, gene sets from the GO were also used, which included cellular components (635 gene sets), molecular functions (2 500 gene sets) and biological processes (3 048 gene sets). The transcription factor target gene sets were created from the Transcriptional Regulatory Element Database.

1.2 Hypothesis testing framework

The overall objective of this analysis was to identify gene sets in which the gene expression levels show prominent differences between the two phenotypes. A parametric statistical approach was used to achieve the goal.

The null hypothesis is: The expression differences for the genes in a gene set between the two phenotypes show the same distribution as that of all the genes in the experiments. To construct a suitable statistic for the hypothesis test, the basic assumption was made that the *SNR*s satisfied the standard Gaussian distribution after normalization. In statistics, $\bar{x}$ or χ^2 could be chosen as the natural test statistic, but the testing efficiency of the two statistic was different. Statistic $\bar{x}$ did identify gene sets in which the genes were expressed up or down in the same manner, but failed to find gene sets in which some genes went up whereas others went down with the overall mean value unchanged or slightly changed. The expressions of genes in a deregulated pathway may remain unchanged in the same way. Tests on some examples showed the limit of the $\bar{x}$ statistic. The statistic χ^2 overcomes this deficiency because of its form, thus suits our hypothesis. Therefore, χ^2 is used in the procedure as the statistic to identify the desired gene sets (Figure 1). In the analysis, for a gene set S with k genes,

$$SDS(S)=\sum_{i=1}^{k} SNR_i^2 \quad (2)$$

where *SDS* refers to set deviation score.

Then, *SDS* (S) was compared with the critical value of $\chi_\alpha^2(k)$ from the $\chi^2(n)$ distribution table for the required significant level (the *P*-value threshold is 0.005).

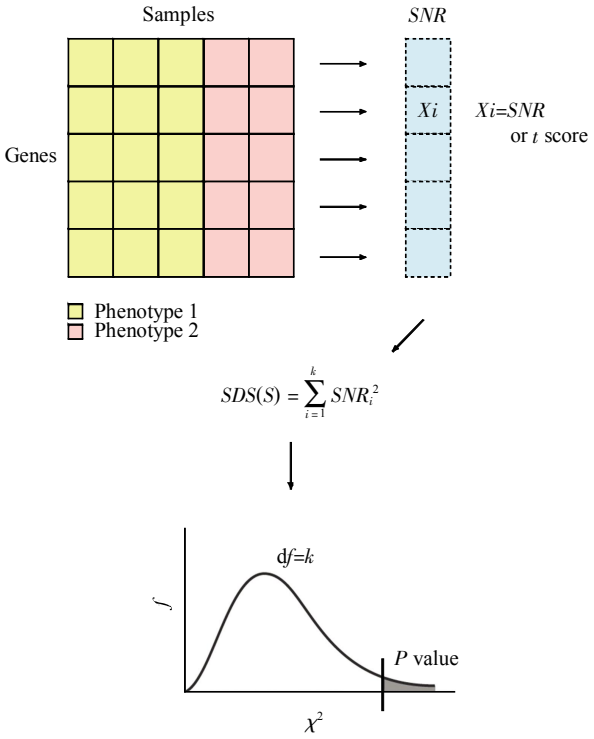

Fig. 1 A schematic diagram of calculating the significant level of gene sets

2 Results and discussion

The analysis was used to identify deregulated gene sets with statistical significance related to brain tumors. The *SNR* can be used as the difference metric for each individual gene. Figure 2 shows the distribution of the normalized *SNR*s and standard Gaussian distribution. Although there are some differences in the figure between the two curves, the *SNR*s' distribution is close enough to be considered as a Gaussian distribution since they have very similar statistical characteristics.

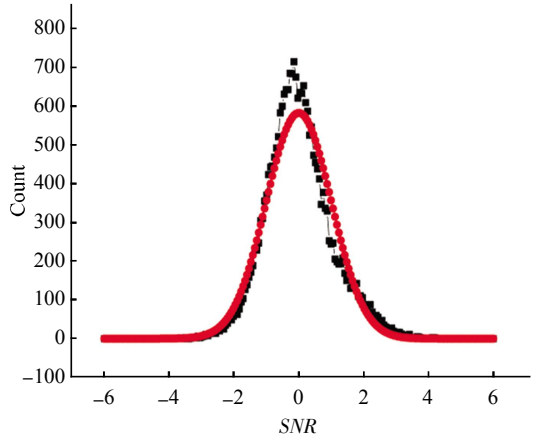

Fig. 2 *SNR* population distribution compared with a standard Gaussian

■—■: Population distribution; ●—●: Standard Gaussian.

2.1 Significant deregulated gene sets

The program did not filter the gene sets by size so as not to miss sets with meaningful biological information. The *P* value was used as the final ranking score. With significance $P \leqslant 0.005$, 400 deregulated gene sets were identified from the C2 gene sets, 202 from the GO gene sets and 12 from the TF target gene sets.

Most of the most significant gene sets are related to brain tumor. It is well-known that some receptor signaling pathways are involved in tumor suppression and cancer progression, such as TGFβ receptor, B cell receptor. TNFα is a proinflammatory cytokine, which plays a role in the pathogenesis of neuronal degeneration[8] and in the neuroinflammatory response. Meanwhile, it is reported that TGFβ and TNFα are involved in the pathogenesis of AD [9]. Mitochondria dysfunction is one of the hallmarks of cancer cell and plays important roles in neuronal apoptosis[10]. Some of the genes are known to be functionally changed in the cancer cell, such as TP53, JUN, MYC[11]. Also, processes related to immune response were discovered in our analysis, such as RUTELLA_HEMATOGFSNDCS_DIFF (hematopoietic growth factors promote the differentiation of tolerogenic dendritic cells).

2.2 Comparison with ΣT

The representative method, GSEA (software downloaded from http://www.broad.mit.edu/gsea) with the default parameters was chosen for comparison with the present results. The parameter 'gene list sorting mode: real or absolute' strongly affected the GSEA results, so comparisons were made with both options.

2.2.1 Comparison with GSEA results sorted in real mode.

After combining the significant gene sets enriched in both normal and tumor phenotypes, the sets were ranked by their absolute NES scores. GSEA identified 532 gene sets from the C2 gene sets, 200 from the GO gene sets and 42 from the TF target gene sets. The two methods agreed on 213 of the C2 sets, 55 of the GO sets and 10 of the TF target gene sets. About 70% of the top 100, 80% of the top 30, and 50% of the top 10 gene sets which are discovered in the GSEA results also were identified from the three databases by the present method. Selected representative sets that were missed in the GSEA results but identified by the present analyzes are listed in Tables 1～3.

Table 1 16 gene sets not identified by the GSEA analysis in the C2 sets

Gene set	Rank	Description
TNF-α	11	TNF-α pathway
VIPPATHWAY	44	Apoptosis of activated T cells is inhibited by vasoactive intestinal peptide (VIP)
FMLPPATHWAY	45	fMLP receptor recognizes formylated bacterial peptides and activates NADPH oxidase
CDMACPATHWAY	53	Ca^{2+} promotes cell proliferation in cultured macrophages by entering the cell *via* calcium channels and activating the MAP kinase pathway
KERATINOCYTEPATHWAY	56	Keratinocyte differentiation, requires the four main MAP kinase pathways
FCER1PATHWAY	76	In mast cells, Fc epsilon receptor 1 activates BTK, PKC, and the MAP kinase pathway to promote degranulation and arachnidonic acid release
TCRPATHWAY	84	T cell receptors pathway induce T cell activation
RACCYCDPATHWAY	100	Ras, Rac, and Rho coordinate to induce cyclin D1 expression and activate cdk2 to promote the G1/S transition
CALCINEURIN_NF_AT_SIGNALING	109	Genes associated with signal transduction through calcium, calcineurin, and NF-AT
PPARAPATHWAY	121	Peroxisome proliferators regulate gene expression *via* PPAR/RXR heterodimers
BCRPATHWAY	122	B-cell receptors activate tyrosine kinases and transiently increase tyrosine phosphorylation
PHOSPHATIDYLINOSITOL_SIGNALING_SYSTEM	126	Phosphatidylinositol signaling system
NFATPATHWAY	139	Cardiac hypertrophy is induced by NF-ATc4 and GATA4
MCALPAINPATHWAY	147	In integrin-mediated cell migration, calpains digest links between the actin cytoskeleton and focal adhesion proteins
CXCR4PATHWAY	190	Responds to the ligand SDF-1 by activating Ras and PI3 kinase to promote lymphocyte chemotaxis
DREAMPATHWAY	192	TF DREAM blocks expression of the prodynorphin gene to blocks pain signaling

Table 2 Gene sets not identified by the GSEA analysis in the TF target gene sets

Genes sets	Rank	Description
TF_NFIC	6	Transcription factor NFIC target genes
TF_NFIA	10	Transcription factor NFIA target genes

Table 3 Representative gene sets not identified by the GSEA analysis in the GO gene sets

Gene set	Rank	Description
GO:0006533	25	Aspartate catabolism
GO:0006468	37	Protein amino acid phosphorylation
GO:0045964[1)]	44	Positive regulation of dopamine metabolism
GO:0005754[1)]	46	Proton-transporting ATP synthase, catalytic core (sensu Eukaryota)
GO:0030308	56	Negative regulation of cell growth
GO:0007264	60	Small GTPase mediated signal transduction
GO:0015991	62	ATP hydrolysis coupled proton transport
GO:0006464	63	Protein modification
GO:0005080	64	Protein kinase C binding
GO:0004861	65	Cyclin-dependent protein kinase inhibitor activity
GO:0007010	70	Cytoskeleton organization and biogenesis
GO:0004422	71	Hypoxanthine phosphoribosyltransferase activity
GO:0006915	89	Apoptosis
GO:0042127	122	Regulation of cell proliferation
GO:0007187	123	G-protein signaling, coupled to cyclic nucleotide second messenger
GO:0008635	143	Caspase activation *via* cytochrome c
GO:0042102	145	Positive regulation of T cell proliferation
GO:0006020[1)]	146	myo-Inositol metabolism
GO:0042771	150	DNA damage response, signal transduction by p53 class mediator resulting in induction of apoptosis
GO:0016337	156	Cell-cell adhesion

[1)] Gene sets contain only one gene.

The reasons that these gene sets were identified by the current approach but not by GSEA will be explained by examining the characteristics of examples from each gene set. The three examples are TNFα, GO:0006915 (apoptosis), and TF_NFIC.

The statistical data for these gene sets' *SNR*s (normalized) are compared with that of a GSEA specifically identified gene set (ALZHEIMERS_INCIPIENT_UP[12] from C2 is used as a reference) in Table 4.

Table 4 Statistical data for three examples from Tables 1～3

Gene set	Average *SNR*	Average absolute *SNR*	Set size	*P* value of Chi square test
ALZHEIMERS_INCIPIENT_UP	–0.48	0.82	296	0.38
TNF-α	–0.066	0.95	753	3.6×10^{-11}
TF_NFIC	–0.041	0.93	98	0.002 4
GO_apoptosis	–0.039	0.91	261	0.000 60

Gene sets with genes' *SNR* distribution farthest from zero are of most interest, which means that either (1) the *SNR*s are mainly distributed at one of the extremes (top or bottom) or (2) the *SNR*s are mainly at both tails. The top three sets in Table 4 (examples of our result) belong to the second type. The average *SNR*s of these three gene sets are all very close to zero, but their average absolute *SNR*s are approximately 1 (one standard deviation from the center of the population distribution). This means that there are a lot more genes distributed at both tails in these sets than those in non-significant sets. On the contrary, top genes in GSEA ranking with the real sorting value accord well only with the first type because *ES* reflects the degree to which set *S* is overrepresented at either extremes (top or bottom) of the list, but it is not sensitive to sets overrepresented at both ends. In this way, *ES* is very similar to the statistic $\bar{x}$ and may fail to identify some sets with biological meanings, since gene expressions in deregulated gene sets or pathways may go in different directions, exhibiting distributions at both ends.

Before we explore more from this comparison, it is necessary that we look into more details of our approach. The sum of the normalized *SNR*'s square (*SDS*) was used to estimate each gene set's distance from the normal state assuming that the *SNR* population distribution is approximately a standard Gaussian distribution. For a gene set *S* with *k* genes, the *SDS* (*S*) gives the significance of the set by comparison with critical values corresponding to specific significant levels in the chi-square test table. However, the *SNR* distribution may be not a standard Gaussian distribution so the *SDS* distribution may deviate from the standard chi-square distribution. To prove the validity of this method, *k* *SNR*s were randomly selected from population to generate a reference set *S'* with a *SDS* (*S'*). The process was repeated 1 000 times to get a distribution of *SDS*(*S'*). The distribution was almost a chi-square distribution with the standard deviation slightly higher than expected and with the same mean value.

Theoretically, *SDS* can be applied to cases where the genes are distributed at only one tail and where they are distributed at both tails, but the method turns out to be more sensitive to the second case. The results show that for the first case, the method identified desired gene sets, but the sensitivity was not good as the GSEA approach. For example, the genes in set ALZHEIMERS_INCIPIENT_UP in Table 4 are mainly distributed at the bottom (left tail), so GSEA gives it a high *NES*. In our method, however, the average of the absolute *SNR* is so small (about 0.82), indicating few genes with *SNR* higher than 1, that the statistic *SDS* cannot be very significant.

When the genes were distributed at both tails, the current method identifies a number of biologically meaningful gene sets. For example, for the TF_NFIC distribution shown in Figure 3, the two tails of the curve turned up while the center remains low. This shape indicates that the genes in this set were distributed away from the center but not at one end, so the GSEA *ES* score with the real *SNR* ranking did not identify this set.

2.2.2 Comparison with GSEA results sorted in absolute mode. GSEA identified 432 gene sets from the C2 sets, 47 from the GO sets and 22 from the TF target genes sets. The TF_NFIC set was identified, but

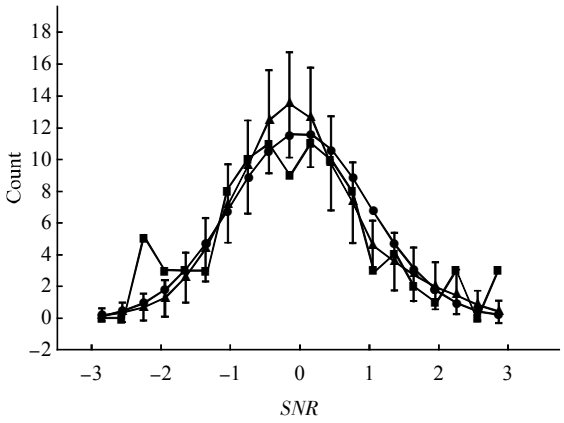

Fig. 3 *SNR* distribution of set TF_NFIC compared with standard a Gaussian and random sets from the *SNR* population with the same set size

■—■:TF_NFIC; ●—●: Standard Gaussian; ▲—▲: Randomly sampled from population.

GO_apoptosis and some other sets were still not found in its significant gene sets list. Moreover, this search missed some gene sets with important biological functions that were identified in the first result, such as TF_TP53, TP_MYC (from the TF target gene sets) and PGC (from the C2 sets). Our initial expectation was that the gene sets discovered using the absolute *SNR* to rank the genes would include all the real *SNR* ranking results, but there were fewer gene sets identified in the absolute mode. Some sets may have been lost because the gene sets were more dispersed in the list than in the real mode so their *ES* values were lower.

2.2.3 Gene sets identified only by the present method. Four representative gene sets from the C2 sets that were identified by the present method with relatively high rankings but not by either GSEA method are listed in Table 5.

Table 5 Four gene sets identified only by our method

Gene set	Average *SNR*s	Average absolute *SNR*s	Set size	*P* value of Chi square test
NFATPATHWAY	0.397	1.14	47	8.61×10^{-5}
MCALPAINPATHWAY	0.320	1.28	22	0.000 1
DREAMPATHWAY	0.540	1.36	13	0.000 3
CCR5PATHWAY	-0.003 40	1.36	16	0.000 5

P values of these gene sets barely pass the significance threshold we set. Most sets play important roles in tumor genesis and have been used as targets in drug design to treat human brain tumors. For example, in the NFATPATHWAY, the NFAT proteins are a family of Ca^{2+}/calcineurin-responsive transcription factors primarily recognized for their central roles in T lymphocyte activation. Yet, they have also been shown to regulate other genes related to cell cycle progression, cell differentiation and apoptosis, revealing a broader role for these proteins in normal cell physiology. Several reports have addressed the participation of NFATs in various aspects of malignant cell transformation and tumorigenic processes [13]. There are also reports that three other gene sets MCALPAINPATHWAY [14], DREAMPATHWAY [15], CCR5PATHWAY[16] are associated with brain tumor.

2.3 Comparison with TΣ

In this comparison both up-regulated and down-regulated genes in the same gene set or pathway are considered. Thus, an absolute expression formula was used to calculate the expression level of the *j*th gene set or pathway:

$$S_j = \sum_{i=1}^{k} I(SNR_i>0)X_{ij} \quad (3)$$

where function *I* is the indicator function returning 1 if the argument is true and −1 otherwise, X_{ij} is the gene expression level of *i*th gene in the *j*th gene set.

Since a simple function cannot be established to describe the distribution of the *SNR* of *S*, the *SNR* was simply ranked according to this value. The results overlap very little with the present method for the top ranked sets, agreeing on only 12 of the top 100 C2 sets, 4 of the top 50 GO sets, and 3 of the top 10 TF target genes sets. At the same time, the results from TΣ overlapped with the results from ΣT on only 2 of the top 100 C2 sets, 1 of the top 50 GO sets, and 0 of the top 10 TF target genes sets, which was similar to that found by Nacu *et al*.[7].

2.4 Discussion

Unlike the previous two methods, the present approach focused on the distributions of all the *SNR*s and the *SNR*s in each gene set. Parametric statistics

were used with a suitable statistic, *SDS*, which was constructed for each gene set to estimate the difference between the set's *SNR* distribution and the *SNR* distribution with the population. The method does not require a ranked list for all genes. This method is very sensitive to gene sets containing both up- and down-regulated genes, which is quite meaningful in biology. *SDS* was chosen as the test statistic for two reasons: previous studies try to take into account gene sets that contain both up- and down-regulated genes and use an absolute mode as the test statistic, but there is no simple classical statistic to describe the distribution of the absolute mode. However, the chi-square distribution can characterize the square of the variance following a normal Gaussian distribution in a convenient, statistically significant manner. The second reason is that many classical statistical testing methods for a single gene, such as the t test and F test, are very difficult to apply to test the gene sets. The major bottleneck is that the sizes of various gene sets are quite different. Therefore, the average significance of the genes is always used. However, this normalization assumes that the score is approximately linear with the size of the gene sets, which probably introduces a bias. In addition, these processing methods always tend to select smaller gene sets. On the contrary, *SDS*, which follows a Chi-square distribution, can analyze all sizes of gene sets; the chi-square test does not assume a linear size relationship.

In conclusion, the *SDS* method can be regarded as a new statistical framework for scoring multiple genes, and also one that gives more biologically meaningful results.

References

1 Kim R D, Park P J. Improving identification of differentially expressed genes in microarray studies using information from public databases. Genome Biology, 2004, **5**(9): R70

2 Subramanian A, Tamayo P, Mootha V K, *et al*. Gene set enrichment analysis:A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA, 2005, **102** (43): 15545～15550

3 Lee E, Chuang H Y, Kim J W, *et al*. Inferring pathway activity toward precise disease classification. Plos Computational Biology, 2008, **4**(11): e1000217

4 Ideker T, Ozier O, Schwikowski B, *et al*. Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics, 2002, **18**(Suppl 1): S233～S240

5 Horvath S, Dong J. Geometric interpretation of gene coexpression network analysis. Plos Computational Biology, 2008, **4**(8): 27

6 Tian L, Greenberg S A, Kong S W, *et al*. Discovering statistically significant pathways in expression profiling studies. Proc Natl Acad Sci USA, 2005, **102**(38): 13544～13549

7 Nacu S, Critchley-Thorne R, Lee P, *et al*. Gene expression network analysis and applications to immunology. Bioinformatics, 2007, **23**(7): 850～858

8 Meda L, Cassatella M A, Szendrei G I, *et al*. Activations of microglial cells by beta-amyloid protein and interferon-gamma. Nature, 1995, **374**(6253): 647～650

9 Mattson M P, Barger S W, Furukawa K, *et al*. Cellular signaling roles of TGF beta, TNF alpha and beta APP in brain injury responses and Alzheimer′s disease. Brain Research Reviews, 1997, **23**(1～2): 47～61

10 Keller J N, Kindy M S, Holtsberg F W, *et al*. Mitochondrial manganese superoxide dismutase prevents neural apoptosis and reduces ischemic brain injury: Suppression of peroxynitrite production, lipid peroxidation, and mitochondrial dysfunction. J Neuroscience, 1998, **18**(2): 687～697

11 Hanahan D, Weinberg R A. The hallmarks of cancer. Cell, 2000, **100**(1): 57～70

12 Blalock E M, Geddes J W, Chen K C, *et al*. Incipient Alzheimer′s disease: microarray correlation analyses reveal major transcriptional and tumor suppressor responses. Proc Natl Acad Sci USA, 2004, **101**(7): 2173～2178

13 Mosieniak G, Pyrzynska B, Kaminska B. Nuclear factor of activated T cells (NFAT) as a new component of the signal transduction pathway in glioma cells. J Neurochemistry, 1998, **71**(1): 134～141

14 Ray S K, Patel S J, Welsh C T, *et al*. Molecular evidence of apoptotic death in malignant brain tumors including glioblastoma multiforme: Upregulation of calpain and caspase-3. J Neuroscience Research, 2002, **69**(2): 197～206

15 Jo D G, Kim M J, Choi Y H, *et al*. Pro-apoptotic function of calsenilin/DREAM/KChIP3. Faseb J, 2001, **15**(3): 589～591

16 Kouno J, Nagai H, Nagahata T, *et al*. Up-regulation of CC chemokine, CCL3L1, and receptors, CCR3, CCR5 in human glioblastoma that promotes cell growth. J Neuro-Oncology, 2004, **70**(3): 301～307

p-SAGE: 基因集合的参数统计分析方法 *

黄 波 ** 李文婷 ** 李 雯 夏雪峰 孙之荣 ***

(清华大学生物信息与系统生物学研究所，生物信息学教育部重点实验室，生物膜与膜技术国家重点实验室，
清华大学生物科学与技术系，北京 100084)

摘要 肿瘤的发生与发展通常是由重要的细胞通路表达水平的反常导致的．尽管目前已经有很多工作利用基因表达谱数据以及一些其他的先验知识来寻找那些与肿瘤相关的基因集合，但还是没有一个恰当的基因集合的统计学方法．大多数研究都是直接将单基因的检验方法直接应用到基因集合上来．提出了基因集合的参数统计分析方法(p-SAGE)，这个方法应用到大脑肿瘤的实验中，识别了许多显著的基因集合．一些新发现的基因集合是与信号转导和免疫相关的．这个简便有效的方法可以得出有生物学意义的结果．

关键词 参数统计分析方法，基因集合，反常基因

学科分类号 Q **DOI:** 10.3724/SP.J.1206.2009.00325

* 国家高技术研究发展计划(863)(2006AA020403), 国家重点基础研究发展计划(973)(2009CB918801)和国家自然科学基金(30770498)资助项目.

** 共同第一作者.

*** 通讯联系人.

Tel: 010-62772237, E-mail: sunzhr@mail.tsinghua.edu.cn

收稿日期：2009-05-15，接受日期：2009-08-25