Gene expression

GAzer: gene set analyzer

Sang-Bae Kim¹, Sungjin Yang¹, Seon-Kyu Kim¹, Sang Cheol Kim², Hyun Goo Woo¹, David J. Volsky³, Seon-Young Kim⁴,* and In-Sun Chu¹,*

¹Korean Bioinformation Center, KRIBB, Daejeon 305-806, ²Department of Applied Statistics, Yonsei University, Seoul, 120-749, Korea, ³Molecular Virology Division, St. Luke’s-Roosevelt Hospital Center, 432 West 58th Street, Antenucci Building, Room 709, New York, NY 10019, USA and ⁴Human Genomics Laboratory, Genome Research Center, KRIBB, Daejeon 305-806, Korea

Received on November 23, 2006; revised on April 1, 2007; accepted on April 6, 2007

Advance Access publication April 27, 2007

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: Gene Set Analyzer (GAzer) is a web-based integrated gene set analysis tool covering previously reported parametric and non-parametric models. Based on a simulation test for the reported algorithms, we classified and implemented three main statistical methods consisting of the z-statistic, gene permutation and sample permutation for ten gene set categories including Gene Ontology (GO) for human, mouse, rat and yeast. This tool identifies significantly altered gene sets scored by z-statistics and P-values from the z-test or permutation test and provides q-values and Bonferroni P-values to correct multiple hypothesis testing. GAzer allows users to observe changes in expression of each gene in a gene set or to see the significance of the gene sets containing a gene(s) of interest, thus allowing interactive data analysis both at the gene and gene set level. Moreover, GAzer offers extensive annotation for each gene.

Availability: The GAzer gene set analyzer is freely available at http://integromics.kobic.re.kr/GAzer/

Contact: kimsy@kribb.re.kr and chu@kribb.re.kr

Supplementary information: This can be found on the web page (http://integromics.kobic.re.kr/GAzer/supplement.jsp)

1 INTRODUCTION

A common approach to microarray data analysis is to identify differentially expressed genes between two sample populations and then to get biologically meaningful interpretations from the selected gene list. In addition to these conventional statistical approaches such as the t-test, SAM (Tusher et al., 2001) and other statistical models (Efron and Tibshirani, 2002; Pan, 2002), new approaches based on biologically categorized gene sets (e.g. Gene Ontology, GO), pathway and chromosomal location, have been recently introduced (Al-Shahrour et al., 2005; Breslin et al., 2005; Mootha et al., 2003; Tu et al., 2005). Under the assumption that weak but coordinated expression changes of specific gene sets can better represent significant flows of biological processes, the gene set analysis approach has shown good potential for interpreting gene expression data since gene set enrichment analysis (GSEA) was introduced (Mootha et al., 2003).

Gene set analysis can be categorized into non-parametric [i.e. GSEA, ErmineJ (Lee et al., 2005) and Catmap (Breslin et al., 2004)] or parametric [i.e. PAGE (Kim and Volsky, 2005) and T-profiler (Boorsma et al., 2005)] methods. Although these approaches have advantages and disadvantages similar to conventional tests such as the rank-sum and two-sample t-test, they are useful for obtaining biological insights from gene expression data.

An extensive collection of predefined gene sets is also important to maximize the analytical power of gene set analysis. In our previous studies, we showed that new and improved biological information can be extracted from composite GO (compGO) (Nam et al., 2006) or cis-regulatory element gene set (Kim and Kim, 2006) categories.

Here, we introduce a web-based and integrated gene set analysis tool designed for gene set analysis using previously reported statistical methods. This tool includes ten functionally annotated gene sets to effectively extract specific biological information; the gene sets comprise pathways, chromosomal locations, InterPro domains, cis-regulatory elements and information about three GO and three compGO subgroups.

2 DESCRIPTION

GAzer consists of a database of predefined gene sets, a utility to evaluate significantly changed gene sets from user input data and a gene annotation tool.

2.1 Gene set database

We constructed a database of ten kinds of predefined gene set categories for human, mouse, rat and yeast. To establish the gene sets, we downloaded annotation information from several public resources such as the NCBI Entrez Gene database (http://www.ncbi.nlm.nih.gov), Affymetrix (http://www. affymetrix.com), TRANSFAC (http://www.gene-regulation. com) and others. We describe the gene set construction processes in detail in our web site (see Supplementary Material online) and all the gene sets can be downloaded
from our web site. In GAzer, the gene symbol is used as the main identifier and Affymetrix identifiers are converted into gene symbols to average replicate genes.

2.2 Methods

We applied previously reported methods, each having different features, to the analysis of statistically significant gene sets (Table 1). Depending on the statistics and statistical test used, we organized the methods into six categories: PAGE ($z$-test), T-profiler ($t$-test), NT (gene permutation with $t$-value), ET (sample permutation with $t$-value) and old and new GSEA (Mootha et al., 2005; Subramanian et al., 2005).

In a series of simulations (see Supplementary Material online), we investigated the performance of each method for the following variables: the total number of genes in a data set, gene set size, sample size, fold change between two groups, the percentage of coordinately changing genes in a gene set and gene–gene correlation. Our simulation study showed following points. First, gene permutation or $z$-test methods generally performed better than sample permutation or $t$-test methods. Second, when sample size was small, sample permutation methods were impractical. Third, when sample size is less than five, using fold change as an input yields better performance than using the $t$-statistic. Fourth, among sample permutation-based methods, the method of Tian et al. (2005) performed better than either the old or new GSEA methods. In our study of the effects of increasing gene–gene correlation within gene sets, we found that permutation-based algorithms (PAGE and NT) tended to produce higher $z$-scores while sample based algorithms (ET, GSEA_N and GSEA_O) produced lower $z$-scores.

We chose to implement PAGE, NT and ET methods to allow testing of two related hypotheses for coordinated association of a group of genes with a phenotype of interest (Tian et al., 2005).

2.3 GAzer system and implementation

GAzer reads two tab-delimited text files—a gene expression data file and a class information file containing the class identifiers (‘0’ or ‘1’)—to identify two groups. Then, users can select several attributes, including one of four species, array types (Affymetrix or dual-channel array), platform type in case of Affymetrix array, ID type (gene symbol, RefSeq, Entrez gene or ORF) in case of a dual-channel array, the minimum number of genes in a gene set, categories of gene sets and analysis methods (PAGE, PAGE_t, NT or ET). Users can choose as many as ten gene sets. For Affymetrix data, GAzer calculates the average value of replicated genes and converts Affymetrix probe IDs into gene symbols. After computing the statistics of a chosen method, GAzer summarizes the significant gene sets filtered and scored by $z$-statistics and $P$-values from the $z$-test or permutation test (Fig. 1A). This system also provides $q$-values and Bonferroni $P$-values to correct multiple hypotheses testing.

GAzer allows users to view individual genes of a gene set with the heatmap of their expression values and to prioritize each gene set in the web interface gene list to see which genes are significant in the overall behavior of the set (Fig. 1B). GAzer also allows users to cross-compare genes in predefined Table 1. Comparison of GAzer with previously reported gene set analysis tools

<table>
<thead>
<tr>
<th>Method</th>
<th>Statistics used</th>
<th>Statistical test</th>
<th>Fold Change</th>
<th>Statistical test</th>
<th>Fold Change</th>
<th>Gene set</th>
<th>Gene set</th>
<th>Gene set</th>
<th>Gene set</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAGE</td>
<td>Fold Change, $z$-test, $t$-test</td>
<td>$z$-test</td>
<td>Fast</td>
<td>Slow</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>YES</td>
</tr>
<tr>
<td>PAGE_t</td>
<td>Fold Change, $z$-test, $t$-test</td>
<td>$z$-test</td>
<td>Fast</td>
<td>Slow</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>YES</td>
</tr>
<tr>
<td>NT</td>
<td>Rank permutation (gene)</td>
<td>$t$-test</td>
<td>Fast</td>
<td>Slow</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>YES</td>
</tr>
<tr>
<td>ET</td>
<td>Permutation (gene)</td>
<td>$t$-test</td>
<td>Fast</td>
<td>Slow</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>YES</td>
</tr>
<tr>
<td>GSEA_N</td>
<td>Permutation (sample)</td>
<td>$t$-test</td>
<td>Fast</td>
<td>Slow</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>YES</td>
</tr>
<tr>
<td>GSEA_O</td>
<td>Permutation (sample)</td>
<td>$t$-test</td>
<td>Fast</td>
<td>Slow</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>YES</td>
</tr>
<tr>
<td>ErmineJ</td>
<td>Rank permutation (gene, sample)</td>
<td>$t$-test</td>
<td>Fast</td>
<td>Slow</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>YES</td>
</tr>
<tr>
<td>MEGO</td>
<td>Rank permutation (gene, sample)</td>
<td>$t$-test</td>
<td>Fast</td>
<td>Slow</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>YES</td>
</tr>
<tr>
<td>GSEA</td>
<td>Rank permutation (gene, sample)</td>
<td>$t$-test</td>
<td>Fast</td>
<td>Slow</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>YES</td>
</tr>
</tbody>
</table>

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/23/13/1697/221402 by guest on 02 May 2019
gene sets. Thus, GAzer has the capacity to analyze gene expression data both at a gene and gene set level interactively. Finally, extensive annotation information for each gene is provided, including all public database identifiers such as NCBI, SWISSPROT, GO and pathway information.

GAzer is implemented in R language (http://www.r-project.org, http://www.bioconductor.org) and uses MySQL (http://www.mysql.com) as a DBMS. This program is wrapped by JAVA (http://www.java.sun.com) to maintain a user-friendly web interface. All R scripts used in the simulation study and GAzer are available on our web site (see Supplementary Material online). Further details, including a User’s Guide, are available on the web site.

ACKNOWLEDGEMENTS

This software development was supported in part by KRRIB Research Initiative Program and 21c Frontier Functional Human Genome Project, Korea. We thank Joshua S. Yang and Kwang-Sik Shin for their excellent technical support.

Conflict of Interest: none declared.

REFERENCES


Fig. 1. Screen shots of GAzer. (A) Process results. Statistical values such as z-scores, P-values, q-values and Bonferroni values are calculated, and significant gene sets ordered by the q-value are shown in the table. (B) Gene list in the selected gene set. Users can display the gene list of a gene set that the user has selected along with the heatmap of the expression values.