Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

# Gene Class Expression: analysis tool of Gene Ontology terms with gene expression data

**Gislaine S.P. Pereira[1,2], Rodrigo M. Brandão[2], Silvana Giuliatti[1], Marco A. Zago[2,3] and Wilson A. Silva Jr.[1,2]**

[1]Departamento de Genética, Faculdade de Medicina de Ribeirão Preto, USP, Ribeirão Preto, SP, Brasil
[2]Centro Regional de Hemoterapia,
Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto, Ribeirão Preto, SP, Brasil
[3]Departamento de Clínica Médica, Faculdade de Medicina de Ribeirão Preto, USP, Ribeirão Preto, SP, Brasil
Corresponding author: G.S.P. Pereira
E-mail: gislaine@lgmb.fmrp.usp.br

Genet. Mol. Res. 5 (1): 108-114 (2006)
Received January 10, 2006
Accepted February 17, 2006
Published March 31, 2006

**ABSTRACT.** Serial analysis of gene expression (SAGE) technology produces large sets of interesting genes that are difficult to analyze directly. Bioinformatics tools are needed to interpret the functional information in these gene sets. We present an interactive web-based tool, called Gene Class, which allows functional annotation of SAGE data using the Gene Ontology (GO) database. This tool performs searches in the GO database for each SAGE tag, making associations in the selected GO category for a level selected in the hierarchy. This system provides user-friendly data navigation and visualization for mapping SAGE data onto the gene ontology structure. This tool also provides graphical visualization of the percentage of SAGE tags in each GO category, along with confidence intervals and hypothesis testing.

**Key words:** Gene expression, Functional annotation, Tag classification, Web-based bioinformatics tool

 

## INTRODUCTION

A biological event in cells is controlled by the expression of multiple genes in a determined time and in an appropriate way. To monitor the pattern of expression of genes under some pathological and physiological conditions, one of the steps is to understand the biological process (Chen et al., 2000). The success of the development of the serial analysis of gene expression (SAGE) technique has been an important mark. In the SAGE technique, a short 10-base sequence tag, corresponding to an expressed sequence, is concatenated for automatic sequencing, along with tags of differently expressed sequences. This strategy allows maximum coverage of expressed genes for gene identification at the global level of the genome, while it keeps the sequencing analysis to a manageable scale. The application of the SAGE technique has produced valuable information on several biological systems, and the amount of data has grown exponentially (Zhang et al., 1997). One of the challenges in research today is to interpret such data quickly, integrating biological resources of knowledge, such as Gene Ontology (GO) annotation and molecular information. Due to the large number of expressed genes in the eukaryotic genome, bioinformatics tools are necessary to characterize patterns of gene expression (Velculescu et al., 1995). Analytical tools that use a bioinformatics approach and mathematical methods can promote the functional annotation of the expression data and, consequently, present these data in a format that propitiates a better agreement with the biology of the corresponding processes (Lee et al., 2005). The project Gene Ontology Annotation of the European Bioinformatics Institute has a dynamic and controlled vocabulary of GO, which describes the biological process, cellular component and molecular function of generic cells, making it possible to characterize gene products (Harris et al., 2004). Biological knowledge concerning GO, when combined with experimental results and computational approaches, can be useful for biomedical research and for the discovery of new drugs (Lee et al., 2005). The integration of cancer genome data and GO annotations can be a valuable strategy for the selection of biomarkers, for identifying new therapies and for determining the effects of drug treatment (Arciero et al., 2003; Cunliffe et al., 2003). The knowledge on GO also provides a systematic inquiry and functional classification of etiologies of multifactorial diseases, increasing information that could improve the planning and the treatment of these illnesses (Philip-Couderc et al., 2004; Prabakaran et al., 2004; Lee et al., 2005). There are several examples where the information of the GO database can be applied to data generated on a wide scale, and there are currently many tools that provide graphical interfaces, as well as hierarchical and functional annotation of data of the genome based on GO annotations (http://www.geneontology.org/GO.tools.shtml). Examples of such tools include MAPPFinder (Doniger et al., 2003), GoMiner (Zeeberg et al., 2003), FatiGO (Al-Shahrour et al., 2004), Onto-Express (Khatri et al., 2004), EASE (Hosack et al., 2003), GOTree (Zhang et al., 2004), and NetAffx (Cheng et al., 2004). The tool, Gene Class, which we present here, provides an interactive visualization of functional annotation of two lists of SAGE tags in the corresponding GO category, which is chosen by the user. Using the data of SAGE classified in GO hierarchy, the system presents statistical results for each GO category of classification, such as the plot of the percentage of classification for each list of SAGE tags in the corresponding GO category, the result of the test of hypothesis in each category and the confidence interval. The combination of biological annotation with the calculated values will provide the user with the possibility to include biological characteristics that are occult in data, thus increasing the information derived from its database.

## MATERIAL AND METHODS

The new tool, Gene Class, associates SAGE tags with data on GO, providing functional classifications using the GO database (http://www.godatabase.org/dev/database/). First, relationships were established between the SAGE tags and gene products in *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Bos taurus*, *Danio rerio*, *Gallus gallus*, and *Caenorhabditis elegans*, which are included in the GO database. Our objective was to provide the user with a web interface with the option of choosing SAGE tags for classification, based on their frequency in the organized hierarchy of the GO. The classification can be obtained by selecting organism and GO term (biological process, cellular component or molecular function) at different levels (2-5) from two lists of SAGE tags of interest. The GO category is built in directed acyclic graphs (Ashburner et al., 2000) so that each GO term can have several parents in the hierarchy. Our program searches both lists, and it associates each SAGE tag in the term and classification level in the GO hierarchy. The tool then generates a plot of the percentage of genes in each classified category, showing the result of the test of hypothesis in each category, along with the confidence interval. The test of hypothesis is solved considering the following null ($H_0$) and alternative ($H_1$) hypotheses (Triola, 1998).

$$H_0 : p_1 = p_2$$

$$H_1 : p_1 \neq p_2$$

where $p_1$ and $p_2$ are the two lists of SAGE tags targeted for classification. For example, list one would refer to the most frequent tags derived from normal tissue and list two would refer to the most frequent tags derived from tumor tissue. In this case, we assume that the original information is $p_1 \neq p_2$. Using the values of amount in the submission, $n$, and the amount in the classification, and $x$ in each GO category, the total proportion is determined in each term by applying

$$p_1^{'} = \frac{x_1}{n_1}$$

$$p_2^{'} = \frac{x_2}{n_2}$$

where $n_1$ and $n_2$ refer to the total number of genes sent in lists one and two, and $x_1$ and $x_2$ are the number of genes from lists one and two, classified in each GO term, respectively. The statistical test is calculated as follows:

$$z = \frac{p_1^{'} - p_2^{'}}{\sqrt{\frac{\tilde{p}\tilde{q}}{n_1} + \frac{\tilde{p}\tilde{q}}{n_2}}}$$

where

$$p_1 - p_2 = 0$$

and the combined estimate of $p_1$ and $p_2$ are

$$\tilde{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

$$\tilde{q} = 1 - \tilde{p}$$

The choice of the level of significance is selected by the user, where $\alpha = 1\%$ ($z = 2.58$) or $\alpha = 5\%$ ($z = 1.96$). If the original affirmation is the alternative hypothesis and the statistical test is located in a critical region ($H_1$), $H_0$ is rejected, concluding that there is enough evidence to guarantee the original affirmation. However, if the statistical test is located in $H_0$, $H_0$ is accepted, concluding that there is insufficient evidence to support the original affirmation. The confidence interval is calculated as follows:

$$p_1^{'} - p_2^{'} \pm z_{\alpha/2} \cdot \sqrt{\frac{p_1^{'}(1-p_1^{'})}{n_1} + \frac{p_2^{'}(1-p_2^{'})}{n_2}}$$

The confidence interval (or estimated interspace) is an amplitude (or an interval) of values that has the probability of containing the true value of the population (Triola, 1998) used in each category, for SAGE tags having GO annotation in that category.

## RESULTS AND DISCUSSION

The association of functional classification in the GO database is established for two SAGE tag lists with frequency, as can be seen in Figure 1.

Figure 1 shows the input user interface from Gene Class. The input identified by Gene Class is SAGE tags. The user can choose organism, GO term, GO level, and levels of significance for the statistical tests. For interesting SAGE tag set *vs* reference SAGE tag set, the user needs to upload the file (or copy/paste) and choose select fields. The result will be the classification for both SAGE tag sets, identified in the GO category (Figure 2). The input file should be a plain text file, including the frequency of tags, separated by tabs or spaces in the format of one per line, and the analysis is limited by the number of genes and the size of file.

The result shows the output, with the list of SAGE tags, associated symbol, organism, and frequency, by GO category, a display of the percentage of SAGE tags by category, along with the test of hypothesis and confidence interval (Figure 2B). A list of SAGE tags not found is

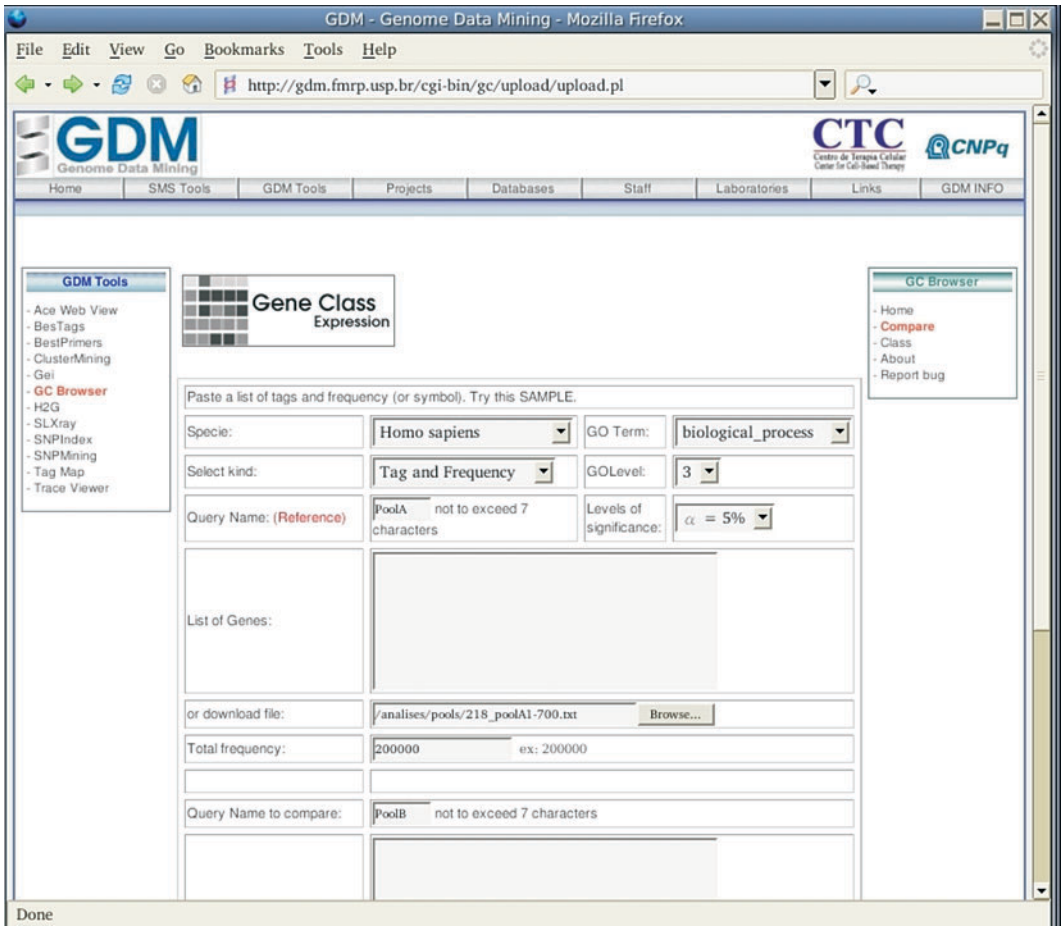


**Figure 1.** Input user interface from Gene Class Expression for uploading analysis, parameters and data.

also available. The Gene Class tool provides a comprehensive classification of SAGE tags in GO structure. This tool makes the normalization of frequency of tags sent, showing an expression interval by frequency (Figure 2A).

## CONCLUSION

The Gene Class tool complements and extends the functionality of similar data-mining tools for GO hierarchy, and it offers additional information through statistical analysis that helps users to analyze the classification for a set of SAGE tags in the GO category. The application of Gene Class is still limited by the number of genes that have GO annotation. However, with the bioinformatics effort to automatically predict protein functions based on the literature, gene expression data and protein sequence information, and a rapid growth in GO, it is expected that this new tool will become more useful with the improvement of GO.

Homepage Gene Class: http://gdm.fmrp.usp.br/cgi-bin/gc/upload/upload.pl.

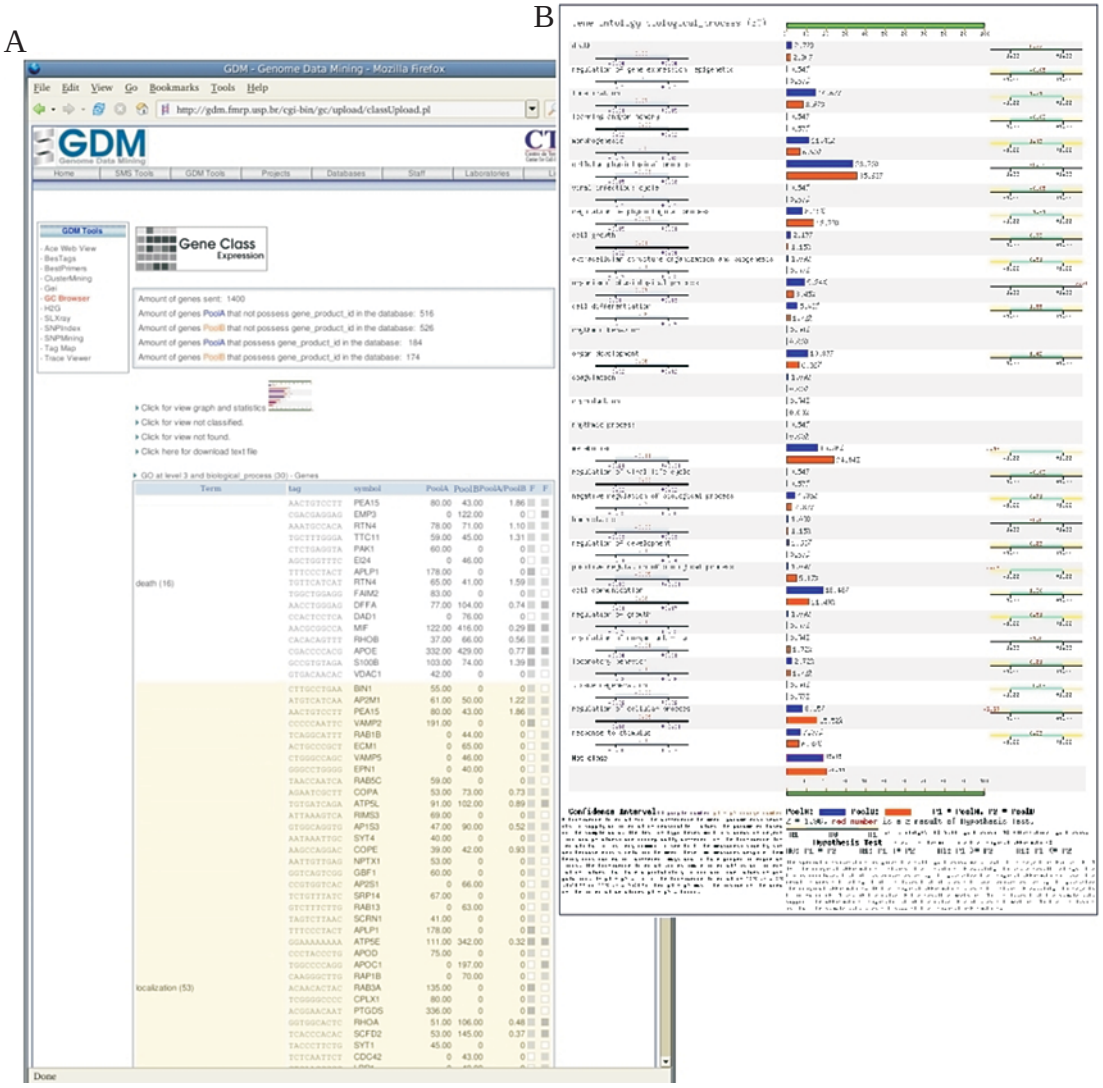


**Figure 2.** Result from submission to Gene Class. **A.** The window displays the Gene Ontology (GO) database category and the tags classified by category. **B.** The display shows the percentage of tags, the test of hypothesis and the confidence interval by GO category.

## ACKNOWLEDGMENTS

We thank Daniel Pinheiro, Israel Tojal, Marco Valtas, and Roberto Focosi for helpful comments. Research supported by the Center for Cell-Based Therapy/FAPESP, CNPq.

## REFERENCES

Al-Shahrour F, Diaz-Uriarte R and Dopazo J (2004). FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics* 20: 578-580.

Arciero C, Somiari SB, Shriver CD, Brzeski H, et al. (2003). Functional relationship and gene ontology classification of breast cancer biomarkers. *Int. J. Biol. Markers* 18: 241-272.

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

Chen JJ, Rowley JD and Wang SM (2000). Generation of longer cDNA fragments from serial analysis of gene expression tags for gene identification. *Proc. Natl. Acad. Sci. USA* 97: 349-353.

Cheng J, Sun S, Tracy A, Hubbell E, et al. (2004). NetAffx Gene Ontology mining tool: a visual approach for microarray data analysis. *Bioinformatics* 20: 1462-1463.

Cunliffe HE, Ringner M, Bilke S, Walker RL, et al. (2003). The gene expression response of breast cancer to growth regulators: patterns and correlation with tumor expression profiles. *Cancer Res.* 63: 7158-7166.

Doniger SW, Salomonis N, Dahlquist KD, Vranizan K, et al. (2003). MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biol*. 4: R7.

Harris MA, Clark J, Ireland A, Lomax J, et al. (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*. 32: D258-D261.

Hosack DA, Dennis Jr G, Sherman BT, Lane HC, et al. (2003). Identifying biological themes within lists of genes with EASE. *Genome Biol*. 4: R70.

Khatri P, Bhavsar P, Bawa G and Draghici S (2004). Onto-Tools: an ensemble of web-accessible, ontology-based tools for the functional design and interpretation of high-throughput gene expression experiments. *Nucleic Acids Res*. 32: W449-W456.

Lee V, Camon E, Dimmer E, Barrell D, et al. (2005). Who tangos with GOA? - Use of Gene Ontology Annotation (GOA) for biological interpretation of '-omics' data and for validation of automatic annotation tools. *In Silico Biol*. 5: 5-8.

Philip-Couderc P, Pathak A, Smih F, Dambrin C, et al. (2004). Uncomplicated human obesity is associated with a specific cardiac transcriptome: involvement of the Wnt pathway. *FASEB J*. 18: 1539-1540.

Prabakaran S, Swatton JE, Ryan MM, Huffaker SJ, et al. (2004). Mitochondrial dysfunction in schizophrenia: evidence for compromised brain metabolism and oxidative stress. *Mol. Psychiatry* 9: 684-697.

Triola MF (1998). Introdução à estatística. 7ª edn. Editora LTC, Rio de Janeiro, RJ, Brazil.

Velculescu VE, Zhang L, Vogelstein B and Kinzler KW (1995). Serial analysis of gene expression. *Science* 270: 484-487.

Zeeberg BR, Feng W, Wang G, Wang MD, et al. (2003). GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol.* 4: R28.

Zhang B, Schmoyer D, Kirov S and Snoddy J (2004). GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. *BMC. Bioinformatics* 5: 16.

Zhang L, Zhou W, Velculescu VE, Kern SE, et al. (1997). Gene expression profiles in normal and cancer cells. *Science* 276: 1268-1272.