

File ID 110284
Filename Chapter 2 T-profiler : scoring the activity of pre-defined groups of genes
using gene expression data

SOURCE (OR PART OF THE FOLLOWING SOURCE):

Type Dissertation
Title Dissection of transcriptional regulation networks and prediction of gene
functions in *Saccharomyces cerevisiae*
Author A. Boorsma
Faculty Faculty of Science
Year 2008
Pages 136
ISBN 9078675303

FULL BIBLIOGRAPHIC DETAILS:

<http://dare.uva.nl/record/271379>

Copyright

*It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or
copyright holder(s), other than for strictly personal, individual use.*

DNA micro-array T-profiler mRNA Aneuploidy
unpaired t-test cluster free post-genomic
era HSF1 *Candida albicans*
gene groups **Chapter 2** REDUCE
chromosome Bonferonni ChIP-chip Environmental
Stress Response Gene Ontology con-
sensus motif *Saccharomyces*

T-profiler: Scoring the activity of pre-defined groups of genes using gene expression data

André Boorsma, Barrett C. Foat, Daniel Vis, Frans M. Klis
& Harmen J. Bussemaker

This chapter was published in *Nucleic Acids Research* **33**: W592-595 (2005).

Abstract

One of the key challenges in the analysis of gene expression data is how to relate the expression level of individual genes to the underlying transcriptional programs and cellular state. Here we describe T-profiler, a tool that uses the t-test to score changes in the average activity of pre-defined groups of genes. The gene groups are derived from Gene Ontology, ChIP-chip experiments, and consensus transcription factor binding motifs; by scoring entire chromosomes, T-profiler can also detect aneuploidy. If desired, an iterative procedure can be used to select a single, optimal representative from sets of overlapping gene groups, and a jack-knife procedure makes calculations more robust against outliers. T-profiler makes it possible to interpret microarray data in a way that is both intuitive and statistically rigorous, without the need to combine experiments or choose parameters. Currently, gene expression data from *Saccharomyces cerevisiae* and *Candida albicans* are supported. Users can upload their microarray data for analysis on the web at <http://www.T-profiler.org>.

Introduction

An important technique in the post-genomic era is the simultaneous measurement of the transcript levels of all genes from a genome by microarray experiments [21, 22]. In recent years, the amount of data from such experiments has rapidly increased [92, 93]. Furthermore, the combination of chromatin-immunoprecipitation and microarray technology (“ChIP-chip”) has made it possible to measure the physical binding of transcription factors to the upstream DNA in a global way [61, 62].

There has also been an explosion in the number of computational methods for analyzing microarray data. Among the most popular are algorithms such as hierarchical clustering [72], K-means clustering [94], and self-organizing maps [95]. A limitation of these clustering methods is the need to have gene expression profiles across multiple experiments. Alternative methods have been developed that can take a single genomewide expression pattern as input, such as motif-based correlation or regression [88, 96, 97].

To obtain easily interpretable information on changes in the cellular state in terms of functional annotation, methods such as Funspec [79], GO term finder [98], GOAL [99], and GeneXpress (<http://genexpress.stanford.edu>) score the significance of overlap between pre-defined gene groups – from Gene Ontology [100] or the MIPS database [101] – and the subset of induced or repressed genes. These methods are based on the cumulative hypergeometric distribution, also referred to as Fisher’s exact test. A disadvantage of these methods is that they require genes to be significantly up or down-regulated on an individual basis in order to contribute.

We previously developed a method that can score Gene Ontology categories without the need to apply cut-offs to the expression level of individual genes [81]. This algorithm, now named T-profiler, uses the standard t-test to score the difference between the mean expression level of pre-defined groups of genes and that of all other genes on the microarray (see Methods). A similar approach was independently pioneered by Pavlidis *et al.* [102]. T-profiler is currently suitable for the analysis of *Saccharomyces cerevisiae* and *Candida albicans* gene expression profiles, and in the near future will be extended to other organisms.

Methodology

For a given gene group G , the t -value is given by the following formula:

$$t_G = \frac{\mu_G - \mu_{G'}}{s \sqrt{\frac{1}{N_G} + \frac{1}{N_{G'}}}}$$

where

$$s = \sqrt{\frac{(N_G - 1) * S_G^2 + (N_{G'} - 1) * S_{G'}^2}{N_G + N_{G'} - 2}}$$

Here μ_G is the mean expression log-ratio of the N_G genes in gene group G ; $\mu_{G'}$ is the mean expression log-ratio of the remaining $N_{G'}$ genes; and s is the pooled standard deviation, as obtained from the estimated variances for groups G and G' . The associated two-tailed p -value can be calculated from t using the t -distribution with $N-2$ degrees of freedom, and is corrected for multiple testing by multiplying it by the number of gene groups that is being tested in parallel (Bonferroni correction). All groups with a corrected p -value of 0.05 or smaller are considered to be significantly regulated. To reduce the influence of outliers, which may result in false positives or false negatives, we discard the highest and lowest expression value in each gene group. This method is similar to the jack-knife procedure [103].

Gene groups sharing a common motif in their upstream region

Motif groups are defined as genes with a match to a particular consensus motif within 600 base pairs of upstream sequence [104], allowing no overlap with neighboring ORFs. The consensus motifs used in T-profiler are derived from three different sources. First, motifs were extracted from the SCPD database (<http://cgsigma.cshl.org/jian/>). Next, motifs found by comparing the genome sequence of highly related yeast species [40, 105], and motifs discovered from various microarray experiments by the REDUCE algorithm [88, 106], were added. Most of these motifs are similar or identical to motifs described in literature. In total, 153 motif groups are included in the T-profiler calculation. Far less information is available about regulatory sequences of *C. albicans*. It was recently reported that about one third of *S. cerevisiae* regulatory elements are conserved in *C. albicans* [107]. T-profiler therefore uses the same list of *S. cerevisiae* motifs, supplemented with some newly discovered *C. albicans* regulatory motifs from the same study, to score *C. albicans* expression data.

Gene groups bound by a common transcription factor based on ChIP-chip data

Binding of transcription factors to their global DNA targets can be measured by so-called ChIP-chip experiments. In *S. cerevisiae* this technique has been explored on a large scale by Lee *et al.* [61]. In this study the targets of more than 100 transcription factors were identified in mid-

log phase cells growing in rich media. Since many transcription factors are not expressed under these conditions, this study was recently extended to 200 transcription factors whose DNA binding was quantified under various stress conditions [62]. We used the transcription factor binding (TFB) data from Harbison *et al.* [62] as input in T-profiler. A gene was considered to be part of a TFB group if the p-value reported by the authors was smaller than 0.001. In addition, TFB groups were required to have at least 7 gene members (before Jack-knife procedure). This resulted in 252 TFB groups that were used for T-profiler analysis.

Gene Ontology Categories

The third type of gene group is based on membership of a specific Gene Ontology (GO) category [100]. In GO, each gene is classified according to biological process, molecular function and cellular component. The GO gene group contains the genes associated with a specific GO category as well as all its child categories. Only Gene Ontology groups with more than 6 members were used for calculation (before Jack-knife procedure). This resulted in 1389 Gene Ontology-derived gene groups, which were used for T-profiler analysis. Positive scores of Gene Ontology groups give direct information about which functions or cellular processes are expected to have changed as result of the altered gene expression.

Iterative removal of redundant gene groups

Several of the pre-defined gene groups scored by T-profiler show strong mutual overlap: the Gene Ontology categories used by T-profiler are hierarchically organized; consensus motifs can match to similar sequences; and ChIP-chip experiments can reveal similar sets of genes to be bound by different transcription factors and/or under different conditions. The t-values for overlapping gene groups are strongly correlated, and therefore mutually redundant. Following the idea of forward selection of non-redundant motifs in REDUCE [88], we implemented an iterative procedure to select a non-redundant set of gene groups among those that have t-values significantly different from zero. At each step, we subtract the mean expression level of the genes in the gene group with the highest absolute t-value from all genes in that gene group. The t-values are then recalculated for all other gene groups, and the procedure is repeated until even the most significantly regulated gene group has a p-value larger than 0.05. In the case of nested Gene Ontology categories at different levels in the hierarchy in particular, this procedure will naturally select the most appropriate level for a given type of pathway.

Aneuploidy test

Hughes *et al.* [108] described the discovery of chromosomal aberrations in yeast deletion mutants based on gene expression profiles. These are often duplications or deletions of an entire chromosome. By applying T-profiler on the level of whole chromosomes, where gene groups are defined as the set of all genes on a specific chromosome, it is possible to detect such aneuploidy.

An Example

Gene expression datasets can be uploaded as a tab-delimited text file with the systematic ORF name in the first column and the expression data in the second column as a log-ratio base two. After uploading, the user is presented with some basic information about the dataset, including the number of genes, the average and standard deviation (Figure 1A). Importantly, no cut-offs are applied; all values are used for calculation. An expression profile comparing cells 80 minutes after a heat shift from 30 to 37°C from the Environmental Stress Response data set of Gasch *et al.* [40] will serve as an example.

Typically, only a small subset of the gene groups considered will be scored as differentially expressed (Figure 1). The statistical parameters that are output by T-profiler are: (i) a t-value measuring the upregulation ($t > 0$) or downregulation ($t < 0$) in units of the standard error of the difference, and (ii) a p-value that is Bonferroni corrected for the parallel testing of the large number of categories, which represents the probability that the t-value would be observed by chance. Four different types of pre-defined gene groups can be scored: genes whose promoter region matches a specific consensus motif (Figure 1B); genes whose promoter is significantly bound by a specific TF according to a ChIP-chip experiment (Figure 1C), genes that lie on a specific chromosome (Figure 1D); and genes that belong to a specific Gene Ontology category (Figure 1E).

Figure 1 (page 40) Screenshots of the various T-profiler analysis results: (a) statistics of the uploaded gene expression dataset (cells assayed 80 minutes after temperature shift from 30° to 37°C)[40]. The type of analysis can be selected from the panels at the right; (b) scoring consensus motifs. Only significantly scoring motifs are shown (p -value < 0.05). By selecting the motifs in the left column, information about the genes containing this motif and their expression levels can be obtained; (c) scoring ChIP-chip based gene groups; (d) graph showing the t-value for each chromosome, obtained from the gene expression profile of the mutant *pdf2Δ*, in which chromosome 14 is duplicated; (e) scoring of Gene Ontology categories; only a subset of the 50 significant (p -value < 0.05) categories are shown; (f) the same result, but now with redundant gene groups removed by our iterative procedure.

Figure 1B shows consensus motifs associated with differential regulation. The heat shock response motif (HSF1) and the general stress response motif (MSN2/4) score positively, whereas the PAC and rRPE motifs, both over-represented in genes involved in rRNA biosynthesis [43] score negatively. The upregulation of genes under control of the HSF1 motif is specific for heat-shocked cells, while the downregulation of genes involved in rRNA biosynthesis and genes containing MSN2/4 motifs is typical for the environmental stress response [40]. **Figure 1C** shows which transcription factors and corresponding ChIP-chip conditions are associated with differential regulation. The fact that genes bound by the transcription factor Hsf1p score positively whereas the genes bound by the transcription factors Rap1p, Sfp1p and Fhl1p, which are all involved in the regulation of ribosomal genes, score negatively is consistent with the motif-based results. **Figure 1E** shows the results of T-profiler analysis based on Gene Ontology; in total, 50 categories have a significant t-value. Most of the positively scoring categories are involved in heat shock and stress response whereas most of the negatively scoring categories are comprised mainly of ribosomal genes. Again, the results compare well to the results obtained by T-profiler using motif and ChIP-chip based gene groups. However, the large number of similar GO categories reported makes it harder to interpret the results. **Figure 1F** shows how this problem is resolved by the iterative removal of redundant categories. Finally, **Figure 1D** clearly shows that the deletion mutant *pdf2Δ* contains a duplication of chromosome 14. By testing several microarray datasets obtained from aneuploid cells we found that as a rule-of-thumb an absolute t-value larger than 10 is a good indicator of aneuploidy.

Conclusion

T-profiler analyzes genomewide expression patterns one experiment at a time, without the need to tune any parameters. Our use of the t-test to score gene groups eliminates the need to impose a threshold on the expression level of individual genes. A group can be scored as significantly induced or repressed even if the expression of none of its member genes changes significantly on a single-gene basis. This feature greatly increases the sensitivity to small-amplitude, coordinate changes in the expression of groups of genes. Representing a transcriptome by a relatively small set of statistically robust and easily interpretable t-values allows for seamless comparison between experiments, even across different platforms and laboratories. We plan to extend the functionality of T-profiler to multiple experiments in the near future.

Acknowledgments

We would like to thank Merijn Schuurmans and Ania Zakrzewska for helpful discussions and for testing T-profiler, and Reka Letso for a critical reading of the manuscript. This work was supported by grants from the Netherlands Foundation for Technical Research (STW) to F.K. (APB.5504) and from the National Institutes of Health to H.J.B. (R01HG003008).

