W186–W192 *Nucleic Acids Research, 2007, Vol. 35, Web Server issue*
doi:10.1093/nar/gkm323

GeneTrail—advanced gene set enrichment analysis

Christina Backes[1,*], Andreas Keller[1], Jan Kuentzer[1], Benny Kneissl[1], Nicole Comtesse[2], Yasser A. Elnakady[3], Rolf Müller[3], Eckart Meese[2] and Hans-Peter Lenhof[1]

[1]Center for Bioinformatics, Saarland University, Building E1.1, 66041 Saarbrücken, [2]Department of Human Genetics, Medical School, Building 60, 66421 Homburg/Saar and [3]Department of Pharmaceutical Biotechnology, Saarland University, Building A4.1, 66041 Saarbrücken, Germany

Received January 31, 2007; Revised April 6, 2007; Accepted April 17, 2007

ABSTRACT

We present a comprehensive and efficient gene set analysis tool, called 'GeneTrail' that offers a rich functionality and is easy to use. Our web-based application facilitates the statistical evaluation of high-throughput genomic or proteomic data sets with respect to enrichment of functional categories. GeneTrail covers a wide variety of biological categories and pathways, among others KEGG, TRANSPATH, TRANSFAC, and GO. Our web server provides two common statistical approaches, 'Over-Representation Analysis' (ORA) comparing a reference set of genes to a test set, and 'Gene Set Enrichment Analysis' (GSEA) scoring sorted lists of genes. Besides other newly developed features, GeneTrail's statistics module includes a novel dynamic-programming algorithm that improves the *P*-value computation of GSEA methods considerably. GeneTrail is freely accessible at http://genetrail.bioinf.uni-sb.de

INTRODUCTION

Modern high-throughput methods generate large sets of genes and proteins that cannot be analyzed manually. Therefore, computer-aided gene set analysis tools that try to identify significantly enriched functional categories in these sets have gained increasing importance.

For the statistical evaluation of gene sets two basic approaches have been developed. The first method, the so called 'Over-Representation Analysis' (ORA), compares the set of interest to a reference set. When considering a certain functional category, i.e. a GO term, this method tries to detect if this category is over-represented or under-represented in the respective set and estimates how likely this is due to chance. The second method is called 'Gene Set Enrichment Analysis' (GSEA). Here, the input set is sorted by some specific criteria (e.g. gene expression values). When considering an arbitrary functional category, GSEA tests if the genes in the set that belong to the category are uniformly distributed or accumulated on top or on bottom of the sorted input list.

Some of the developed tools focus on the analysis of only one type of functional categories for example various Gene Ontology (GO) (1) based tools, among them FatiGO (2), BiNGO (3), and GOstat (4). Other tools focus on certain types of high-throughput data as microarray gene expression data [ErmineJ (5), CRSD (6), GSEA-P (7)] or offer only one type of statistical analysis, as the GSEA-P tool (7), that is designed for GSEA only. Furthermore, some tools, like Catmap (8), do not include biochemical categories and it is left to the user to define these categories. A few tool packages, however, allow for the analysis of different types of functional categories, e.g. WebGestalt (9) and Babelomics (10).

Here, we present GeneTrail, a web-based application, allowing for the identification of enriched functional categories in protein or gene sets. GeneTrail supports the ORA as well as the GSEA approach. In addition, our implementation of the GSEA analysis includes a novel algorithm that computes the correct *p*-value instead of estimating it by permutation tests. Since our tool is based on the comprehensive integrative system BN++ (11), GeneTrail allows the evaluation of a broad range of functional categories. The advantage of using the BN++ database BNDB is that we are able to find cross-links (e.g. using the data of different protein–protein interaction databases), which remain undetected using single databases. The current version of BN++ integrates for example the following biological data sources: RefSeq (12), KEGG (13), TRANSPATH (14), TRANSFAC (15), DIP (16), MINT (17), HPRD (18), IntAct (19). Besides the categories mentioned above, GeneTrail also offers amino acid sequence analyses as motif search, coiled-coil prediction or granzyme B cleavage site prediction,

*To whom correspondence should be addressed. Tel: +49-681-302-68608; Fax: +49-681-302-64719; Email: cbackes@bioinf.uni-sb.de

The authors wish to be known that, in their opinion, this first two authors should be regarded as joint First Authors.

a chromosomal location analysis and a protein–protein interaction analysis.

With GeneTrail, we developed a user friendly web-based application, which can be easily extended concerning new functional categories or statistical methods for the evaluation of arbitrary high-throughput data. GeneTrail is freely accessible (http://genetrail.bioinf.uni-sb.de).

MATERIALS AND METHODS

Information resources and databases

GeneTrail imports KEGG (13) and TRANSPATH pathways (14), TRANSFAC transcription factors (15) and protein–protein interactions from DIP (16), MINT (17), HPRD (18) and IntAct (19) from the biochemical network database (BNDB). BNDB is a powerful relational database platform, allowing a complete semantic integration of an extensive collection of external databases. It is built upon a comprehensive and extensible object model called BioCore, which is powerful enough to model most known biochemical processes and at the same time it is easily extensible to be adapted to new biological concepts. This database is part of BN++ (11), the biochemical network library, which is freely available at http://www.bnplusplus.org. Additionally, GeneTrail uses a local copy of the GO database (1) that includes electronically inferred annotations (IEAs) and manually curated annotations. GeneTrail provides the user the option to analyze the complete data set or to exclude the IEAs. To complement the above mentioned data sources, our application imports different flat files from the NCBI containing current versions of gene identifiers and amino acid sequences. By using BN++ together with the described annotation files, GeneTrail allows for using many identifier types. A survey on supported gene identifiers is provided in Table 1. The listed identifiers are initially mapped to NCBI Gene IDs. These can directly be associated with most of the integrated categories, for example Gene Ontology terms. Thereby, a minimal loss of information is guaranteed.

Table 1. Overview of the identifier types currently supported by GeneTrail

Identifier	Examples
NCBI GeneID	5894, 11186, 11848
NCBI NP/XP number (Protein RefSeq)	NP_006261, XP_941900, NP_872606
NCBI Protein GI	28201876, 113431221, 121114292
NCBI NM/XM number (RNA RefSeq)	NM_018993, NM_008284, NM_021168
NCBI RNA GI	54792783, 51093847, 91105420
SwissProt/UniProt	Q9NZD4, P55008, O15155
UniGene	Hs.652097, Hs.652094, Hs.652089
Ensembl Gene ID	ENSG00000003147, ENSG00000005801
SGD yeast ORF ID	YCR024C-B, YCR108C, YLR157W-E
Amersham Human Whole Genome	GE200018, GE897528, GE519380
Affymetrix HG-U133A	1487_at, 1320_at, 1316_at
Affymetrix HG-U95A	1014_at, 1015_s_at, 1017_at
Affymetrix HG-U133 Plus 2.0	1552258_at, 1487_at, 1438_at
Affymetrix HG-U133B	200017_at, 200018_at, 200013_at

We have developed update routines ensuring that GeneTrail imports the most recent versions of flat files. The underlying database and all flat files are updated monthly.

Supported functional categories

Besides the functional categories from the KEGG, TRANSPATH and GO databases, GeneTrail also offers the possibility to study amino acid sequence properties, e.g. the presence of specific amino acid motifs, coiled–coiled structures [as described by Lupas *et al.* (20)] and granzyme B cleavage sites (21). Additionally, GeneTrail studies the enrichment of genes regulated by certain transcription factors from the TRANSFAC database or significant protein–protein interactions in the test set. Moreover, GeneTrail allows for studying the distribution of the genes in the test set on the chromosomes and chromosome arms.

GeneTrail can also be used to perform statistical analyses for self-defined functional categories. To use this option, a category file can be uploaded by the user. This file has to contain the category name with a leading '#' symbol and the identifiers belonging to this category each separated by a line break.

To provide insight into tested categories, GeneTrail offers a comprehensive log file that can be accessed and downloaded from the web page. For each category, the log file contains the source of that category, and the number of genes in the considered category.

Statistical methods

GeneTrail provides two different types of statistical approaches. First, genes of a test set can be compared to a reference set (ORA). Second, a sorted test set can be analyzed without a reference set (GSEA). For each biological category, a significance value (*P*-value) is computed. Since many categories are usually tested, the raw P-values need to be adjusted for multiple testing.

Over-representation analysis (*ORA*). Suppose that we are given a test set of n genes of which k belong to a certain category C and a reference set of m genes of which l belong to C. Since l elements of the reference set belong to C, we expect to find $k' = l \cdot n/m$ elements in the test set. If k is larger than k', C is said to be enriched, if k is smaller than k', C is said to be depleted. To estimate the statistical significance, *P*-values are computed. If the test set is a subset of the reference set, the hypergeometric test is applied to compute a one tailed *P*-value for C:

$$P_C = \begin{cases} \sum_{i=k}^{n} \dfrac{\binom{l}{i}\binom{m-l}{n-i}}{\binom{m}{n}} & \text{if } k' < k \\ \sum_{i=0}^{k} \dfrac{\binom{l}{i}\binom{m-l}{n-i}}{\binom{m}{n}} & \text{if } k' \geqslant k \end{cases}$$

If test and reference set are disjoint, Fisher's exact test is used instead:

$$P_C = \begin{cases} \sum_{i=k}^{n} \dfrac{\binom{n}{i}\binom{m}{l+k-i}}{\binom{m+n}{l+k}} & \text{if } k' < k \\ \sum_{i=0}^{k} \dfrac{\binom{n}{i}\binom{m}{l+k-i}}{\binom{m+n}{l+k}} & \text{if } k' \geqslant k \end{cases}$$

Gene Set Enrichment Analysis (*GSEA*). If the genes in the test set are sorted, e.g. by expression values, a running sum statistic is computed for each category C. This statistic shows whether the genes of C are accumulated on top (Figure 1A, B) or bottom (Figure 1C) of the sorted test set, or if they are randomly distributed (Figure 1D). Given a set of m genes of which l belong to C, the sorted list is processed from top to bottom. Whenever a gene of C is found, the running sum is increased by a certain amount, otherwise it is decreased. We consider the unweighted case where the running sum is increased by $m-l$ or decreased by l, corresponding to a Kolmogorov–Smirnov non-parametric rank statistic [as described in (22) and (23)]. The minimum and maximum of the running sum statistic are used to estimate the significance of the enrichment. The P-value is computed as the probability that a random running sum reaches a value as high as the maximum of the running sum (accumulation on top) or as low as the minimum of the running sum (accumulation on bottom). Usually, these probabilities are computed by so-called permutation tests that are time consuming and provide only an estimation of the correct P-value. We developed a dynamic programming algorithm that

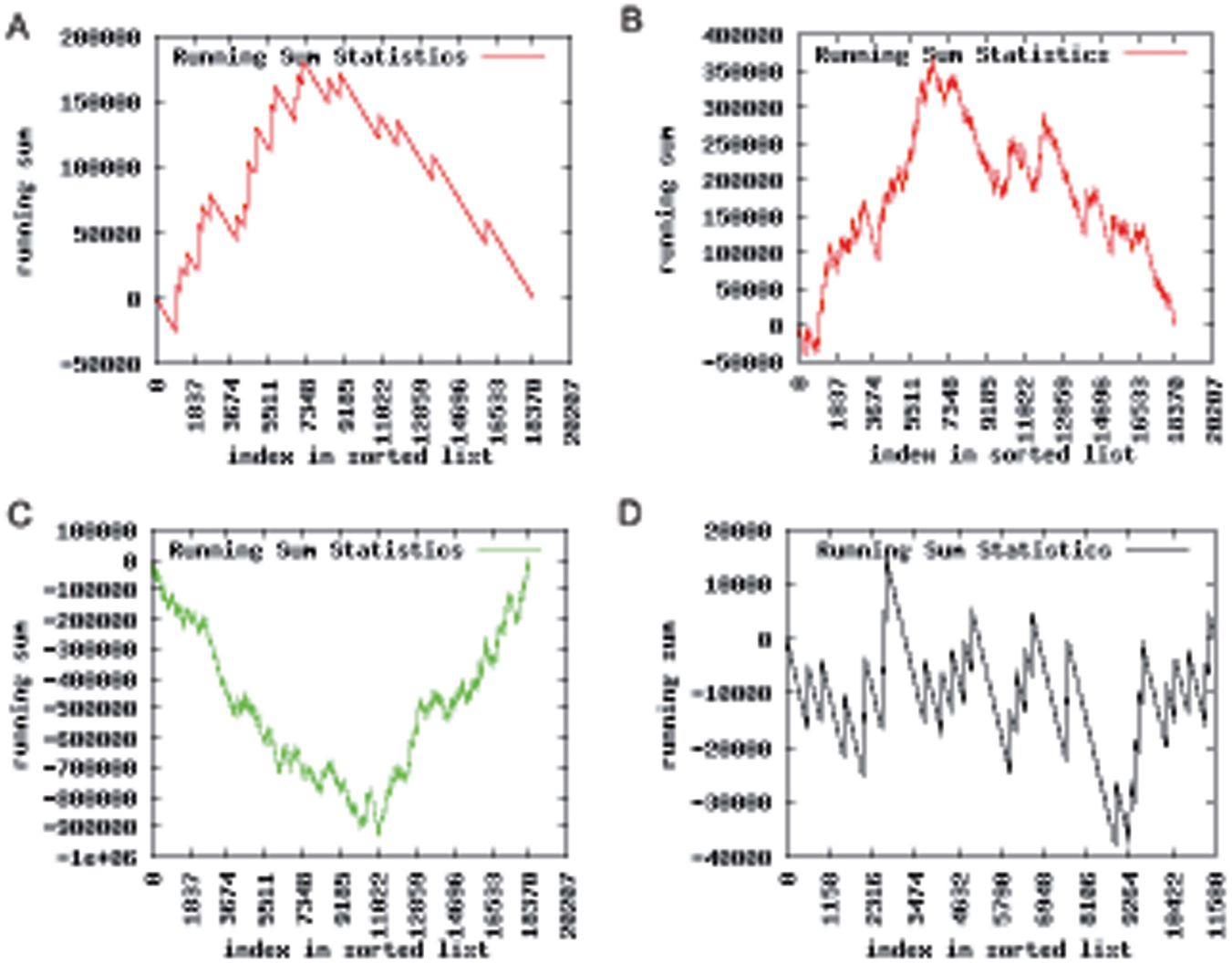

Figure 1. Visualization of different running sum statistics when applying a 'Gene Set Enrichment Analysis'. The running sum (y-axis) is shown as function of the index in the sorted list (x-axis). Part **A** and **B** of the figure illustrate a 'mountain-like shape' for top ranked genes. In part **C**, a 'valley-like shape' for bottom ranked genes is shown. Part **D** illustrates a 'zigzag' shape which is not statistically significant; the genes are randomly distributed.

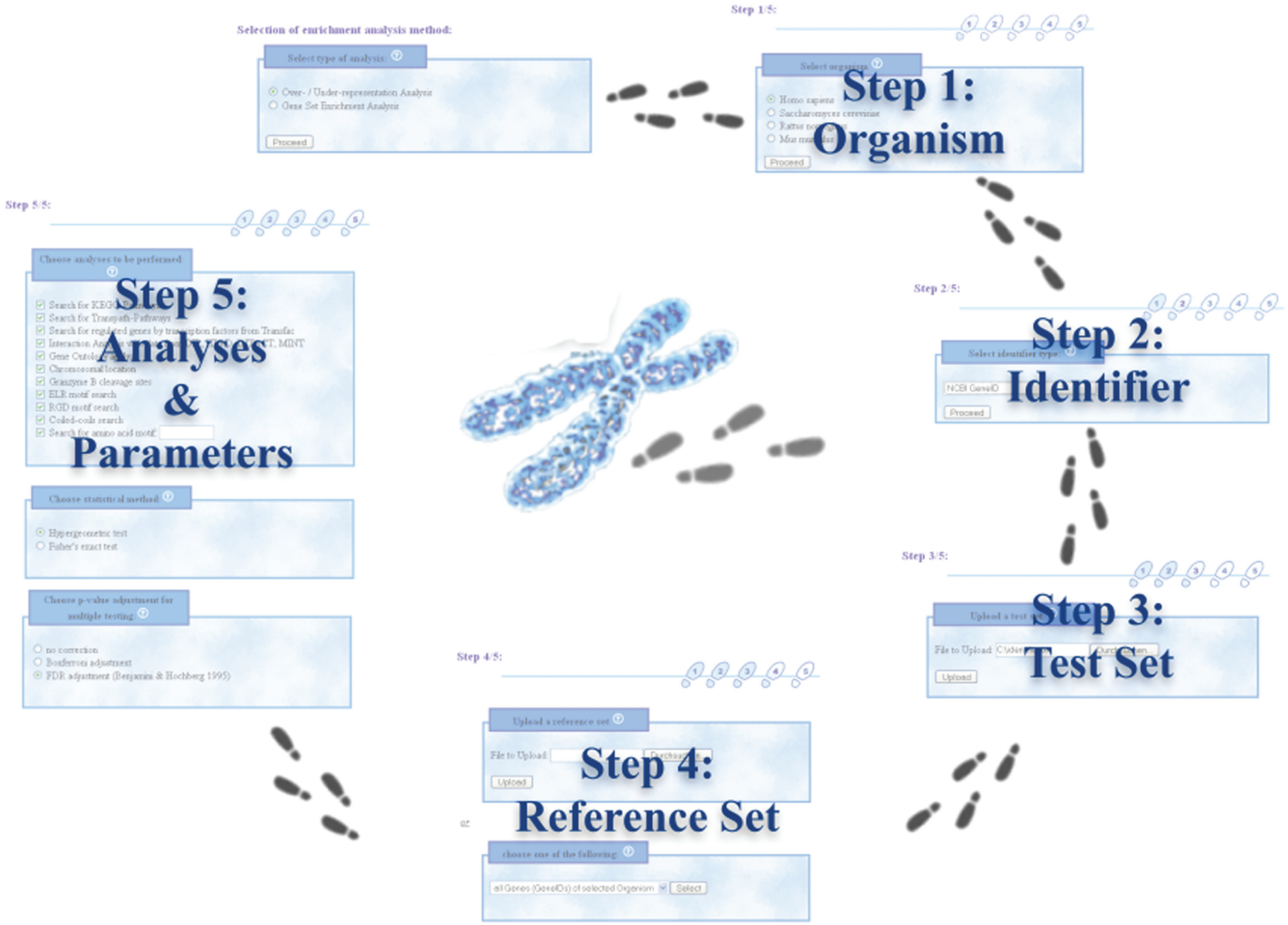

Figure 2. This figure exemplifies the workflow of the GeneTrail server. The five steps needed to perform an 'Over-Representation Analysis' are shown in consecutive order. First, the organism and the identifier type have to be selected. Afterwards, a test set should be uploaded and a reference set can be uploaded or selected from a pre-defined list. Finally, the user can specify the desired analysis methods and the required parameters. For each step, we show small screenshots in the background taken from the GeneTrail user interface.

computes the correct *P*-value time efficient (manuscript in preparation).

Multiple testing adjustment. Since many categories are tested simultaneously, we are facing the well-known multiple testing problem. Therefore, GeneTrail offers two adjustment methods, the conservative Bonferroni adjustment and the control of the false discovery rate (FDR) according to Benjamini and Hochberg (24). Please note that the adjustment is performed separately for different category types like KEGG or TRANSPATH pathways.

Handling of replicates. For example in microarray experiments, genes are frequently represented by several transcripts of the microarray. Therefore, the uploaded sorted lists may include several data points for one gene. GeneTrail offers three possibilities to select a transcript, or in general a unique position for each gene in the sorted list: the first occurrence, the last occurrence or the median position of each gene can be selected. For the third option, in the case of an even number of positions for a given gene, we calculate the average of the two 'middle' positions in the sorted list. The new calculated 'position' must not be a natural number. If several genes have the same position, the one with the first occurrence in the original list is placed first etc.

GENETRAIL WEB SERVER

Workflow

The intuitive user interface of GeneTrail guides the user through several steps as illustrated for the ORA in Figure 2. First, the user has to select the organism and the identifier types (see also Table 1) for the genes or proteins. After uploading a test set the user can also upload a reference set or select a set from a pre-defined list. Before starting the analysis, the user has to specify the functional categories to be evaluated and the required parameters. The results of the computation are presented in two different ways. An HTML result page containing significant categories sorted by the calculated *P*-values and an interactive graph visualization. Figure 3 shows an excerpt of the HTML result page of the example set

Figure 3. HTML view excerpt of the output of the ORA performed on the example set provided on the GeneTrail web server homepage. The illustration shows the two significant KEGG pathway categories with the highest *P*-value. The red arrows denote the over-representation of these two categories. If available, the categories and the genes are connected via weblink to their external data sources.

provided on the GeneTrail homepage. Figure 4 presents the graph visualization of the same example. The following two sections discuss the output of the web server in detail.

Representation of the results

HTML-output. The results page summarizes the statistical significant functional categories that are enriched with respect to the test set. For each type of category we provide an overview, which can be extended to see details. The details comprise the functional category name (e.g. the name of a pathway or GO term), the computed *P*-value, a red or green arrow illustrating over- or under-representation with respect to ORA, the number of observed and expected genes, and, in the case of GSEA, an image of the running sum statistic. If possible, the functional categories are also linked to their original sources, e.g. KEGG or NCBI. For a better overview, we additionally visualize the results of the GO term analysis using the GO graph representation. Likewise, the protein–protein interactions are provided as a static graph. Images for GO and protein–protein interactions are, however, only available for sparse graphs.

GeneTrail also illustrates the distribution of genes along chromosomes. Usually, the genes are only represented by points or crosses at their genomic localization. To prevent problems with the visualization of genes that are located very closely to each other, we additionally describe the location of each gene using Gaussian normal distributions with user selectable variance and compute their joint distribution. Since the normal distributions of neighboring genes overlap, we get a more interpretable view of the chromosomal distribution of genes.

All generated files are compressed in a zip file that can be downloaded. This archive contains the HTML result page as well as a comprehensive PDF file.

Graph representation of the results. For a concise overview of the computed results, GeneTrail provides an interactive graph visualization implemented as a Java applet (see Figure 4).

There are two types of nodes in the graph. Oval nodes represent genes and logos categories. For the categories, the nodes' shape corresponds to the data source, the nodes' label to the category. In addition, we indicate an over- or under-representation by a red up-arrow or green down-arrow, respectively. Each gene node is labeled with its gene symbol. The edges are divided in four different classes. Blue edges connect genes and the categories they are found in, black edges denote interactions of gene products, green edges represent activation and red edges indicate inhibition.

We used a circular layout for the graph where genes that belong to the same category are located next to each other. This representation allows the user to easily identify the significant categories for each gene and the associated genes for each category. The user can stepwise include the categories or can include them all at once. This greatly facilitates the analysis of the categories and the corresponding genes by the user. Another useful feature permits to highlight significant categories for a given gene or the genes that belong to a significant category.

DISCUSSION

With GeneTrail we present a web-based application facilitating the statistical evaluation of high-throughput genomic or proteomic data sets. The statistical analysis takes into account the identification of a variety of functional categories that are 'enriched'. The analysis is based on two different statistical approaches, namely ORA and GSEA.

The selection of the analysis method depends on the performed experiment. GSEA is only applicable for sorted gene sets, whereas ORA can be applied to the detection of over- or under-representations of categories in any data set compared to a reference set. In contrast to comparable methods, GSEA represents a threshold free approach. Frequently in ORA of microrarray data, the user determines which genes are considered as upregulated by choosing an expression threshold X. In contrast, GSEA does not rely on a chosen parameter but considers the entire sorted list of transcripts.

The relation of the data and the reference set is crucial. If the data set is a subset of the reference set, the hypergeometric distribution is used to compute *P*-values. For disjoint data and reference sets, Fisher's exact test is applied instead. In all other cases, GeneTrail offers the possibility to download an appropriate reference set disjoint to the data set.

One important feature of GeneTrail is a novel algorithm for computing the exact *P*-value in GSEA. So far, the statistical significance of GSEA is approximated by so-called permutation tests that usually consider only a small amount of all possible permutations. For example, regarding a microarray containing 20 000 genes and a category containing about 2000 genes, the number of possible arrangements is given as $\binom{20000}{2000}$, approximately $4{\cdot}10^{2821}$. Even if thousands of permutations are computed,

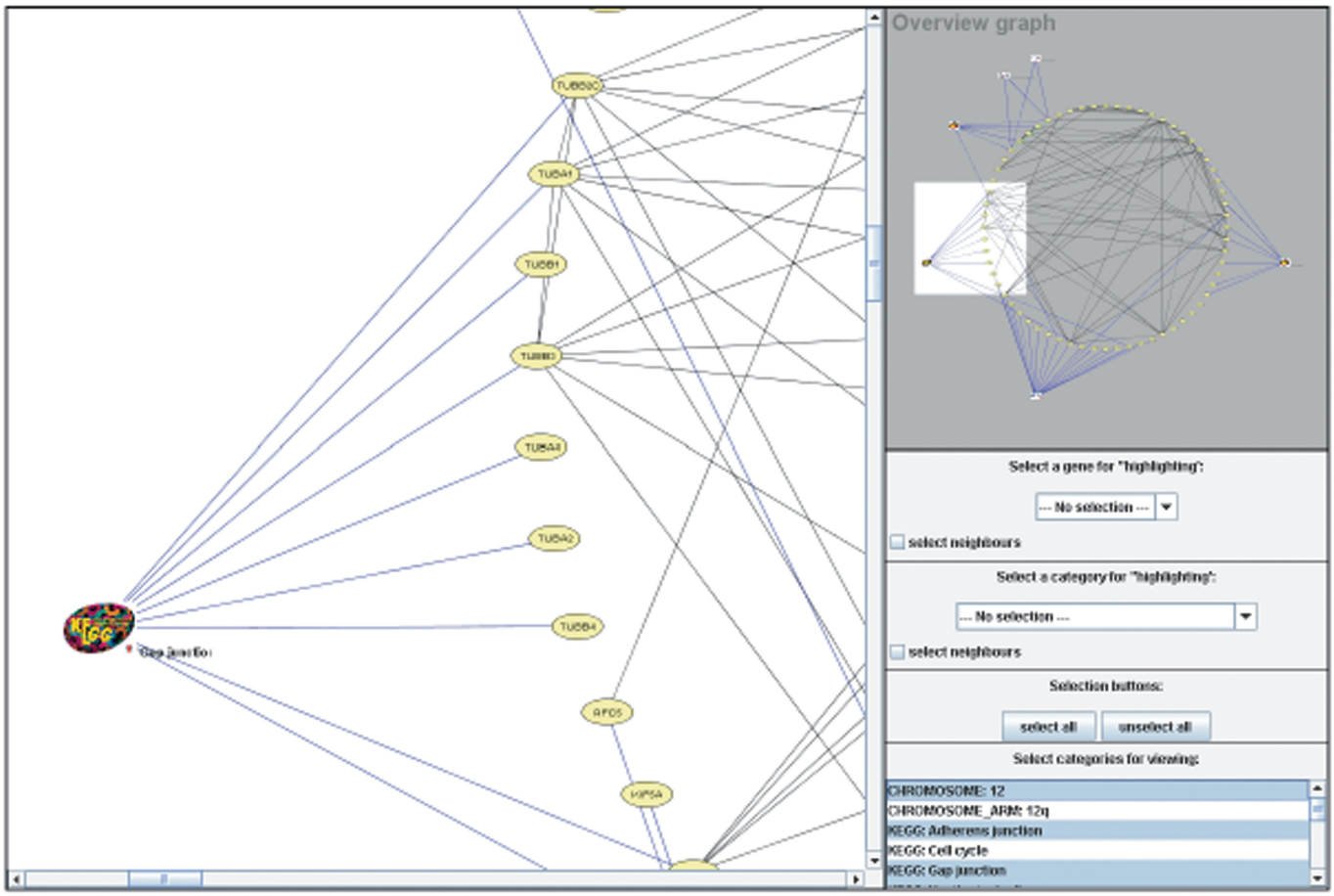

Figure 4. Graph visualization of the output of the ORA performed on the example set provided on the GeneTrail web server homepage. The left hand side shows an excerpt of the complete overview graph presented on the upper right. There are two types of nodes: oval nodes representing the genes in the example set and logos representing the categories. Blue edges connect the genes and the categories they are found in, black edges denote interactions of gene products.

this will probably not yield a good estimation of the *P*-value. However, with our approach, we are able to compute the correct *P*-value efficiently.

The offered statistical approaches treat the genes to work individually. This of course does not reflect the reality, where genes act together. The significance of findings, especially their sensitivity, may be improved by integrating additional information concerning the genes and their interactions, especially the topology of biological pathways. However, we have to carefully select the *a priori* biological information to be included. Since this information could be related to biological coherences we want to detect, we would introduce a bias into the data set.

A common problem of biological data management is the usage of appropriate identifier for genes or proteins. External identifiers need to be mapped to the identifiers used internally. In our case, the NCBI Gene IDs are the internal identifiers. If a provided data set does not contain NCBI Gene IDs, GeneTrail needs to convert these IDs. Therefore, we recommend the usage of NCBI Gene IDs to avoid possible mismatches.

The interactive graph visualization offers the possibility to grasp the interactions between genes of a data set and computed significant categories. However, large gene sets lead to complex graphs that are hard to visualize and to interpret. Currently, we are developing improved methods for online visualization of large graphs that we plan to integrate into future versions of GeneTrail.

GeneTrail offers the possibility to extract information from complex proteome data, microarray data or data generated by other high-throughput methods with minimal effort. Two examples of GeneTrail analyses can be found in the Supplementary Data. In conclusion, GeneTrail complements the conventional evaluation of experimental data and offers new starting points for further experimental investigation.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We appreciate the assistance of Tobias Zimmer in designing the web pages for the GeneTrail server. This work was supported in parts by the 'Deutsche Forschungsgemeinschaft', grant BIZ 4:1-4, the 'Deutsche Krebshilfe', grant 107342 and the 'Klaus-Tschira-Stiftung'. Funding to pay the Open Access publication charges for this article was provided by internal funds of the Saarland University.

Conflict of interest statement. None declared.

REFERENCES

1. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S. *et al*. (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet*., **25**, 25–29.
2. 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.
3. Maere,S., Heymans,K. and Kuiper,M. (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, **21**, 3448–3449.
4. Beissbarth,T. and Speed,T.P. (2004) GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics*, **20**, 1464–1465.
5. Lee,H.K., Braynen,W., Keshav,K. and Pavlidis,P. (2005) ErmineJ: tool for functional analysis of gene expression data sets. *BMC Bioinformatics*, **6**, 269.
6. Liu,C.C., Lin,C.C., Chen,W.S., Chen,H.Y., Chang,P.C., Chen,J.J. and Yang,P.C. (2006) CRSD: a comprehensive web server for composite regulatory signature discovery. *Nucleic Acids Res*., **34**, W571–W577.
7. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R. *et al*. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15278–15279.
8. Breslin,T., Eden,P. and Krogh,M. (2004) Comparing functional annotation analyses with Catmap. *BMC Bioinformatics*, **5**, 193.
9. Zhang,B., Kirov,S. and Snoddy,J. (2005) WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Res*., **33**, W741–W748.
10. Al-Shahrour,F., Minguez,P., Vaquerizas,J.M., Conde,L. and Dopazo,J. (2005) BABELOMICS: a suite of web tools for functional annotation and analysis of groups of genes in high-throughput experiments. *Nucleic Acids Res*., **33**, W460–W464.
11. Küntzer,J., Blum,T., Gerasch,A., Backes,C., Hildebrandt,A., Kaufmann,M., Kohlbacher,O. and Lenhof,H.-P. (2006) BN++ – A biological information system. *J. Integr. Bioinformatics*, **3**, 34.
12. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2005) NCBI reference sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*., **33**, D501–D504, Database Issue.
13. Kanehisa,M., Goto,S., Hattori,M., Aoki-Kinoshita,K.F., Itoh,M., Kawashima,S., Katayama,T., Araki,M. and Hirakawa,M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res*., **34**, D354–D357, Database Issue.
14. Krull,M., Pistor,S., Voss,N., Kel,A., Reuter,I., Kronenberg,D., Michael,H., Schwarzer,K., Potapov,A. *et al*. (2006) TRANSPATH: an information resource for storing and visualizing signaling pathways and their pathological aberrations. *Nucleic Acids Res*., **34**, D546–D551, Database Issue.
15. Matys,V., Kel-Margoulis,O., Fricke,E., Liebich,I., Land,S., Barre-Dirrie,A., Reuter,I., Chekmenev,D., Krull,M. *et al*. (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res*., **34**, D108–D110, Database Issue.
16. Salwinski,L., Miller,C.S., Smith,A.J., Pettit,F.K., Bowie,J.U. and Eisenberg,D. (2004) The database of interacting proteins: 2004 update. *Nucleic Acids Res*., **32**, D449–D451, Database Issue.
17. Zanzoni,A., Montecchi-Palazzi,L., Quondam,M., Ausiello,G., Helmer-Citterich,M. and Cesareni,G. (2002) MINT: a Molecular INTeraction database. *FEBS Lett*., **513**, 135–140.
18. Peri,S., Navarro,J.D., Amanchy,R., Kristiansen,T.Z., Jonnalagadda,C.K., Surendranath,V., Niranjan,V., Muthusamy,B., Gandhi,T.K. *et al*. (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res*., **13**, 2363–2371.
19. Hermjakob,H., Montecchi-Palazzi,L., Lewington,C., Mudali,S., Kerrien,S., Orchard,S., Vingron,M., Roechert,B., Roepstorff,P. *et al*. (2004) IntAct – an open source molecular interaction database. *Nucleic Acids Res*., **32**, D452–D455, Database Issue.
20. Lupas,A., Van Dyke,M. and Stock,J. (1991) Predicting coiled coils from protein sequences. *Science*, **252**, 1162–1164.
21. Backes,C., Küntzer,J., Lenhof,H.P., Comtesse,N. and Meese,E. (2005) GraBCas: a bioinformatics tool for score-based prediction of Caspase- and Granzyme B-cleavage sites in protein sequences. *Nucleic Acids Res*., **33**, W208–W213.
22. Hollander,M. and Wolfe,D. (1999) *Nonparametric Statistical Methods*, 2nd edn. Wiley, New York, USA.
23. Lamb,J., Ramaswamy,S., Ford,H.L., Contreras,B., Martinez,R.V., Kittrell,F.S., Zahnow,C.A., Patterson,N., Golub,T.R. *et al*. (2003) A mechanism of cyclin D1 action encoded in the patterns of gene expression in human cancer. *Cell*, **114**, 323–332.
24. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B*, **57**, 289–300.