ToppGene Suite for gene list enrichment analysis and candidate gene prioritization

Jing Chen¹, Eric E. Bardes², Bruce J. Aronow²,³ and Anil G. Jegga²,³,*

¹Department of Environmental Health, University of Cincinnati, ²Division of Biomedical Informatics, Cincinnati Children’s Hospital Medical Center and ³Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH, USA

Received January 28, 2009; Revised April 24, 2009; Accepted May 11, 2009

ABSTRACT

ToppGene Suite (http://toppgene.cchmc.org; this web site is free and open to all users and does not require a login to access) is a one-stop portal for (i) gene list functional enrichment, (ii) candidate gene prioritization using either functional annotations or network analysis and (iii) identification and prioritization of novel disease candidate genes in the interactome. Functional annotation-based disease candidate gene prioritization uses a fuzzy-based similarity measure to compute the similarity between any two genes based on semantic annotations. The similarity scores from individual features are combined into an overall score using statistical meta-analysis. A P-value of each annotation of a test gene is derived by random sampling of the whole genome. The protein–protein interaction network (PPIN)-based disease candidate gene prioritization uses social and Web networks analysis algorithms (extended versions of the PageRank and HITS algorithms, and the K-Step Markov method). We demonstrated the utility of ToppGene Suite using 20 recently reported GWAS-based gene–disease associations (including novel disease genes) representing five diseases. ToppGene ranked 19 of 20 (95%) candidate genes within the top 20%, while ToppNet ranked 12 of 16 (75%) candidate genes among the top 20%.

INTRODUCTION

High-throughput genome-wide studies like linkage analysis and gene expression profiling, although useful for classification and characterization, do not provide sufficient information to identify specific disease causal genes. Both of these approaches typically result in hundreds of potential candidate genes, often failing to help researchers in reducing the target genes to a manageable number for further validation. To overcome these limitations, several gene prioritization methods have been developed (1–10). While all of these tools are based on the assumption that similar phenotypes are caused by genes with similar or related functions (2,11–13), they differ by the strategy they adopt in calculating similarity and by the data sources they use (14). Except for ENDEAVOUR (5,14) and ToppGene (10), most of the existing approaches mainly focus on the combination of few data sources. Interestingly, none of these approaches utilize mouse phenotype data explicitly in their prioritization approaches even though the mouse is the key model organism for the analysis of mammalian developmental, physiological and disease processes (15). Additionally, previous reports (16,17) have shown that a direct comparison of human and mouse phenotypes allowed rapid recognition of disease causal genes. In an earlier study (10), we have demonstrated that employing mouse phenotype data in fact improves candidate gene prioritization. Through various examples, we also demonstrated (10) that ToppGene performs better than SUSPECTS (9), PROSPECTR (3) and ENDEAVOUR (5), three commonly used methods in candidate gene prioritization.

Most of the current computational disease candidate gene prioritization methods (1–10) rely on functional annotations, gene-expression data or sequence-based features. The coverage of the gene functional annotations, however, is a limiting factor. Currently, only a fraction of the genome is annotated with pathways and phenotypes (10). While two-thirds of all the genes are annotated by at least one functional annotation, the remaining one-third is yet to be annotated. Recent biotechnological advances such as the high-throughput yeast two-hybrid screen have facilitated building proteome-wide protein–protein interaction networks (PPINs) or ‘interactome’ maps in humans (18,19). The shift in focus to systems biology in the post-genomic era has generated further interest in PPINs and biological pathways. While protein–protein interactions (PPI) have been used widely to identify novel disease candidate genes (20–24), several recent studies...
(22,23,25–27) report also using them for candidate gene prioritization.

Since biological networks have been found to be comparable to communication and social networks (28) through commonalities such as scale-freeness and small-world properties, we reasoned that the algorithms used for social and Web networks should be equally applicable to biological networks and developed ToppNet (27). One of the earliest efforts (24) uses a classifier based on several topological features, including degree (number of links to the protein), 1N index (proportion of links to disease-related proteins), 2N index (average 1N index in the neighbors), average distance to disease genes and positive topology coefficient (average neighborhood overlapping with disease genes). A more recent application, Genes2Networks (29), identifies important genes based on functional similarity to training gene list.

The database is updated periodically, and the current status of the data (versions and coverage) can also be accessed from the homepage (‘Database details’). Additionally, several examples with stepwise instructions are provided to demonstrate the utility of these applications (see ‘Supplementary' section from ToppGene homepage).

### TOPPFUN: GENE LIST FUNCTIONAL ENRICHMENT

ToppFun can be used for gene list functional enrichment analysis. It uses as many as 14 annotation categories including GO terms, pathways, protein–protein interactions, protein functional domains, transcription factor-binding sites, microRNAs, gene tissue expressions and literatures. Flexible options are provided to either download results as a tab-delimited file or display as a chart. Hypergeometric distribution with Bonferroni correction is used as the standard method for determining statistical significance.

### TOPPGENE: FUNCTIONAL ANNOTATIONS-BASED CANDIDATE GENE PRIORITIZATION

ToppGene works by generating a representative profile of the training genes using as many as 14 features and identifies over-representative terms from the training genes. This forms the first step and is done by using ToppFun (see previous section). The test set genes are compared to this representative profile of the training set or the overrepresented terms from the training genes for all categorical annotations and the average vector for the expression values (Figure 1). For a test gene, a similarity score to the training profile for each of the 14 features

---

**Table 1. Summary of ToppGene suite applications**

<table>
<thead>
<tr>
<th>Application</th>
<th>Description</th>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>ToppFun</td>
<td>Detects functional enrichment of input gene list based on Transcriptome (gene expression), Proteome (protein domains and interactions), Regulome (TFBS and miRNA), Ontologies (GO, Pathway), Phenotype (human disease and mouse phenotype), Pharmacoome (Drug-Gene associations) and Bibliome (literature co-citation).</td>
<td>Supported identifiers include NCBI Entrez gene IDs, approved human gene symbols, NCBI Reference Sequence accession numbers; single gene list.</td>
<td>HTML output; Tab-delimited downloadable text file; graphical charts</td>
</tr>
<tr>
<td>ToppGene</td>
<td>Prioritize or rank candidate genes based on functional similarity to training gene list.</td>
<td>Same as above but with two gene lists (training and test)</td>
<td>HTML output</td>
</tr>
<tr>
<td>ToppNet</td>
<td>Prioritize or rank candidate genes based on topological features in protein–protein interaction network.</td>
<td>Same as above</td>
<td>HTML output; Cytoscape-compatible input file; graphical networks</td>
</tr>
<tr>
<td>ToppGeNet</td>
<td>Identify and prioritize the neighboring genes of the ‘seeds’ in protein–protein interaction network based on functional similarity to the ‘seed’ list (ToppGene) or topological features in protein–protein interaction network (ToppNet).</td>
<td>Single gene list</td>
<td>Same as above</td>
</tr>
</tbody>
</table>
is derived and summarized by the 14 similarity scores. In the case of a missing value (for instance, lack of one or more annotations for a test gene), the score is set to \( \frac{1}{C_0} \). Otherwise, it is a real value in \([0, 1]\). Different methods are used for similarity measures of categorical (e.g. GO annotations) and numeric (i.e. gene expression) annotations. While a fuzzy-based similarity measure is applied for categorical terms [see Popescu et al. (30) for additional details], for numeric annotation, i.e. the microarray expression values, the similarity score is calculated as the Pearson correlation of the two expression vectors of the two genes. The 14 similarity scores are combined into an overall score using statistical meta-analysis. A \( P \)-value of similarity score \( S_i \) is defined as:

\[
p(S_i) = \frac{\text{Count of genes having higher than } G \text{ in the random sample}}{\text{Count of genes in the random sample containing annotation}}
\]

Fisher’s inverse chi-square method, which states that \( -2 \sum_{i=1}^{n} \log p_i \to \chi^2(2n) \) (assuming \( p_i \) values come from independent tests) is then applied to combine the \( P \)-values from multiple annotations into an overall \( P \)-value. The final similarity score of the test gene is then obtained by 1 minus the combined \( P \)-value. For more
details, validation and comparison with other related applications; the readers are referred to our previous study (10).

ToppNet gene prioritization is based on protein–protein interaction network (PPIN) analyses. Based on the observation that biological networks share many properties with Web and social networks (28), ToppNet uses extended versions of three algorithms from White and Smyth (31)—PageRank with Priors, HITS with Priors and K-step Markov—to prioritize disease candidate genes by estimating their relative importance in the PPIN to the disease-related genes. For more details about the protein interaction datasets used, algorithmic details and validation, see our recently published study (27).

ToppGeNet differs from ToppGene and ToppNet in that the test set is derived from the protein interactome. In other words, for a training set of known disease genes, the test set is generated by mining the protein interactome and compiling the genes either directly or indirectly interacting (based on user input) with the training set. After any overlapping or common genes between test and training sets are removed, interactome-based test set genes can be prioritized using either a functional annotation-based method (ToppGene) or PPIN-based method (ToppNet). The human protein interaction dataset (file ‘interactions.gz’), a compilation of PPIs from BIND(32), BioGRID (33) and HPRD (34), is downloaded from NCBI Entrez Gene FTP site (ftp://ftp.ncbi.nih.gov/gene/).

The programs of our enrichment and prioritization methods are implemented in JAVA. An open-source JAVA package, FtpBean by Calvin Tai (http://www.geocities.com/SiliconValley/Code/9129), is used to automatically download data and annotation files from FTP servers. BioJava packages are used to process UniProt records and extract related protein domain information. GOLEM (http://function.princeton.edu/GOLEM/download.html) source code is adapted and modified for dealing with ontology annotations. Colt (http://dsd.lbl.gov/~hoschek/colt) and Jakarta Commons-Math libraries (http://jakarta.apache.org/commons/math) are used for statistical analysis. The fuzzy similarity measure and related functions are implemented locally. The user front end of ToppGene Suite is a web application written in JAVA. The application server is Sun GlassFish Enterprise Server v2.1 running on OpenSUSE 10.3 Linux. Speed is a key consideration in the design choices of the ToppGene Suite front end. When the web server is started, most of the data is loaded from a relational database and kept in memory.

ToppGene Suite uses two different relational databases for persistence of data: (i) Oracle Database 10g Enterprise Edition Release 10.2.0.3.0 – 64 bit; and (ii) Apache Derby Server - 10.4.2.0. The two databases are used differently. The ‘production data’ are stored in a Derby database on the same computer as the web server, which gives Derby the advantage that it does not have to fetch large data sets across a network and therefore eliminates network latency for small queries. The Oracle Database, on the other hand, is used for data collection and refresh. The data schemas in Oracle are highly structured according to the generally accepted database practice of Third Normal Form. ToppGene Suite uses Hibernate (http://www.hibernate.org/) for updating and retrieving data to and from the databases. The back end of ToppGene Suite is a scripted process that automatically downloads data from publicly available data sources [see (10,27) for more details]. The process, also written in JAVA, is launched using a common JAVA utility called Ant (provided by the Apache Foundation).

The gene information, annotation and the interactions data is updated automatically except for pathways (see the ‘Database details’ section from the homepage of ToppGene Suite for a list of data resources, coverage and version details and dates of last updates). The ‘Database details’ is a dynamic web page that reads the in-memory data structures and displays the counts and statistics of the live data. As the data are refreshed, the counts and statistics are automatically updated. Users can enter the training and test sets of genes of interest as queries from the interface, and the application will display enriched themes in the training set genes along with annotated prioritized test genes. Alternately, users can enter training sets and use the extended gene list from the PPIN as a test set to rank the genes in the interactome using either functional annotations or network features.

For a more detailed validation study using ToppGene, the readers are referred to our previous study (10). In the present study, to demonstrate the utility of ToppGene Suite, we focused on recently reported GWAS. The aim was to test whether ToppGene and ToppNet are capable of retrieving or prioritizing the GWAS-discovered novel disease genes in a training-test type of analysis. We used 20 gene–disease associations (including novel disease genes) representing five diseases (Bipolar Disorder, Cardiomyopathy, Celiac Disease, Crohn’s Disease and Obesity; Table 2). For each of these five disorders, we built a training set containing all the genes already known to play a role in that disorder according to the OMIM and NCBI’s Entrez Gene records (limiting the search field to ‘Disease/Phenotype’ and organism ‘Homo sapiens’) (See ‘Supplementary’ section from ToppGene homepage). The test set consisted of the GWAS-reported disease gene plus 99 nearest neighboring genes based on
their location on the same chromosome. ToppGene and ToppNet prioritization results are presented in Table 2. ToppGene ranked 19 of 20 (95%) candidate genes within the top 20%, while ToppNet ranked 12 of 16 (75%) candidate genes among the top 20%. The mean ranks for ToppGene- and ToppNet-based prioritization were 6.8 and 11.75, respectively (excluding four disease genes that lacked interaction data).

**LIMITATIONS**

ToppGene or any functional annotation-based prioritization method has some limitations. First, when using a training set of genes, the assumption is that the disease genes we have yet to discover will be consistent with what is already known about a disease and/or its genetic basis, which may not always be the case. Second, the annotations and analyses, as well as the prioritization, can only be as accurate as the underlying open sources from which the annotations are retrieved. Similar to functional annotation-based methods, the performance of network-based prioritization methods (ToppNet) is also dependent on the quality of interaction data, which currently suffers from incompleteness and unreliability with missing interactions and false positives.

**CONCLUSIONS**

Existing disease candidate gene prioritization methodologies mine biological and functional information about candidate genes, and we believe that our ToppGene Suite can complement these existing approaches by applying novel methods that mine mouse phenotype data and PPIN. Through various examples, we demonstrate that ToppGene Suite is capable of identifying true candidate genes. However, it needs to be emphasized that our aim is not to prove that ToppGene Suite-prioritized genes are true disease genes but rather to aid in selection of a subset of most likely disease gene candidates from larger sets of disease-implicated genes identified by high-throughput genome-wide techniques like linkage analysis, microarray analysis. As the functional annotations of human and mouse genes and the quality of PPIN improves, we envisage a proportional increase in the performance of ToppGene Suite and strongly believe that it will be a valuable adjunct to wet lab experiments in human genetics and disease research. We further hypothesize that integrating the rankings obtained using functional annotations and PPIN-based approaches may improve the prioritization of disease genes.

**ACKNOWLEDGEMENTS**

We acknowledge the help of Ron Bryson, Technical Writer, Division of Biomedical Informatics, CCHMC, Ohio, USA, in editing the manuscript.

**FUNDING**

State of Ohio Computational Medicine Center (ODD TECH 04-042); National Institutes of Health/National Institute of Diabetes and Digestive and Kidney Diseases (NIH/NIDDK) 1U01 DK70219 (Murine Atlas of a Genitourinary Smooth Muscle Development); PHS Grant P30 DK078392 (Cincinnati Digestive Health...

