

Databases and ontologies

WTFgenes: What's The Function of these genes? Static sites for model-based gene set analysis

Christopher J. Mungall¹ and Ian H. Holmes^{2,3}

¹Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Rd, Berkeley, CA 94720, USA.

²Department of Bioengineering, University of California, Berkeley, CA 94720, USA.

³Molecular Biophysics and Integrated Bioimaging Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Rd, Berkeley, CA 94720, USA.

Abstract

Motivation. A common technique for interpreting experimentally-identified lists of genes is to look for enrichment of genes associated with particular ontology terms. The most common test uses the hypergeometric distribution; more recently, a model-based test was proposed. These approaches must typically be run using downloaded software, or on a server. **Results.** We develop a collapsed likelihood for model-based gene set analysis and present WTFgenes, an implementation of both hypergeometric and model-based approaches, that can be published as a static site with computation run in JavaScript on the user's web browser client. Apart from hosting files, zero server resources are required: the site can (for example) be served directly from Amazon S3 or GitHub Pages. A C++11 implementation yielding identical results runs roughly twice as fast as the JavaScript version. **Availability and Implementation.** WTFgenes is available from <https://github.com/evoldoers/wtfgenes> under the BSD3 license. A demonstration for the Gene Ontology is usable at <https://evoldoers.github.io/wtfgo>. **Contact.** Ian Holmes ihholmes+wtfgenes@gmail.com. **Supplementary Information.** None.

Introduction

Term Enrichment Analysis (TEA) is a common technique for finding functional patterns, specifically over-represented ontology terms, in a set of experimentally identified genes (Boyle *et al.*, 2004). The most common approach, which we refer to as *Frequentist TEA*, is a one-tailed Fisher's Exact Test (based on the hypergeometric distribution, which models the number of term-associations if the gene set was chosen by chance), with a suitable correction for multiple hypothesis testing. Frequentist TEA has been implemented many times on various platforms (Robinson *et al.*, 2002; Khatri *et al.*, 2002; Zeeberg *et al.*, 2003; Boyle *et al.*, 2004; Bauer *et al.*, 2008; Jiao *et al.*, 2012; Mi *et al.*, 2013; Chen *et al.*, 2013).

A model-based alternative to Frequentist TEA, which more directly addresses multiple testing issues (for example, by modeling the partitioning of an observed gene set by complementary terms), is *Bayesian TEA*. In contrast to Frequentist TEA, which just rejects a null hypothesis that genes are chosen by chance, Bayesian TEA explicitly models the alternative hypothesis that the gene set was generated by a few random ontology terms. This approach was introduced by Lu *et al.* (2008), developed by Bauer *et al.* (2010), and implemented in Java and R (Bauer

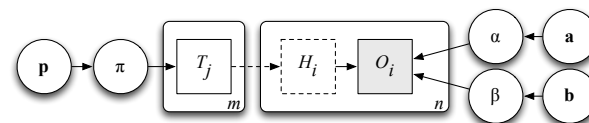


Fig. 1. Model-based explanation of observed genes (O_i) using ontology terms (T_j), following Bauer *et al.* (2010). Other variables and hyperparameters are defined in the text. Circular nodes indicate continuous-valued variables or hyperparameters; square nodes indicate discrete-valued (boolean) variables. Dashed lines indicate deterministic relationships; shaded nodes indicate observations. Plates (rounded rectangles) indicate replicated subgraph structures.

et al., 2011). However, the model-based approach remains significantly less well-explored than Frequentist TEA.

The graphical model underpinning Bayesian TEA is sketched in Figure 1. For each of the m terms there is a boolean random variable T_j ("term j is activated"). For each of the n genes there is a directly-observed boolean random variable O_i ("gene i is observed in the gene set"), and one deterministic boolean variable H_i ("gene i is activated") defined by $H_i = 1 - \prod_{j \in G_i} (1 - T_j)$ where G_i is the set of terms associated

with gene i (including directly annotated terms, as well as ancestral terms implied by transitive closure of the directly annotated terms). The probability parameters are π (term activation), α (false positive) and β (false negative), and the respective hyperparameters are $\mathbf{p} = (p_0, p_1)$, $\mathbf{a} = (a_0, a_1)$ and $\mathbf{b} = (b_0, b_1)$. The model is

$$\begin{aligned} P(T_j = 1 | \pi) &= \pi \\ P(O_i = 1 | H_i = 0, \alpha) &= \alpha \\ P(O_i = 1 | H_i = 1, \beta) &= 1 - \beta \end{aligned}$$

with $\pi \sim \text{Beta}(\mathbf{p})$, $\alpha \sim \text{Beta}(\mathbf{a})$ and $\beta \sim \text{Beta}(\mathbf{b})$. The model of Bauer *et al.* (2010) is similar but used an *ad hoc* discretized prior for π , α and β .

Most Bayesian and Frequentist TEA implementations are designed for desktop use. Several Frequentist TEA implementations are designed for the web, such as DAVID-WS (Jiao *et al.*, 2012) and Enrichr (Chen *et al.*, 2013) which has a rich dynamic web front-end. However, these generally require a server-hosted back end that executes code. Further, there are no web-based Bayesian TEA implementations.

Results

In the Supplementary Information, we describe two refinements to model-based TEA: (i) a collapsed version of Figure 1 with α , β and π integrated out; (ii) additional MCMC sampling moves, which improve mixing.

Our software, WTFgenes, includes JavaScript and C++11 implementations of Bayesian and Frequentist TEA. As reported in the SI, the C++ version is about twice as fast. We also report simulations suggesting Bayesian TEA has greater statistical power than Frequentist TEA, particularly for gene sets explained by multiple terms.

Our JavaScript software, when used as a web application, offers a “quick report” view using Frequentist TEA. For the slower-running but more powerful Bayesian TEA, the software plots the log-likelihood during an MCMC sampling run, for visual feedback. The repository includes setup scripts allowing the tool to be deployed as a “static site”, i.e. consisting only of static files (HTML, CSS, JSON, and JavaScript) that can be hosted via a minimal web server with no need for dynamic code execution.

An example WTFgenes static site, configured for the GO-basic ontology and GO-annotated genomes from the Gene Ontology website, can be found at <https://evoldoers.github.io/wtfgo>.

Discussion

We have provided a static site generator for ontological term enrichment analysis, offering both Bayesian and frequentist tests. Static sites are, in general, cheaper and more secure than running code on a server.

Model-based TEA is versatile: it can readily be extended to allow for datasets that are structured temporally (Hejblum *et al.*, 2015), spatially (Lin *et al.*, 2015), or by genomic region (McLean *et al.*, 2010); to use domain-specific biological knowledge (Szcurek and Beerenwinkel, 2014); or to incorporate additional lines of evidence such as quantitative data (Kalaitzis and Lawrence, 2011). We hope our development of a collapsed likelihood, and evaluation of different MCMC kernels, will assist these efforts.

Coincidentally, Fisher’s Exact Test—which we call Frequentist TEA—was originally motivated by a blind tea-tasting challenge (Fisher, 1935).

Funding

IHH was partially supported by NHGRI grant HG004483. CJM was partially supported by Office of the Director R24-OD011883 and by the

Director, Office of Science, Office of Basic Energy Sciences, of the US Department of Energy under Contract No. DE-AC02-05CH11231.

References

- Bauer, S., Grossmann, S., Vingron, M., and Robinson, P. N. (2008). Ontologizer 2.0—a multifunctional tool for GO term enrichment analysis and data exploration. *Bioinformatics*, **24**(14), 1650–1651.
- Bauer, S., Gagneur, J., and Robinson, P. N. (2010). GOing Bayesian: model-based gene set analysis of genome-scale data. *Nucleic Acids Res.*, **38**(11), 3523–3532.
- Bauer, S., Robinson, P. N., and Gagneur, J. (2011). Model-based gene set analysis for Bioconductor. *Bioinformatics*, **27**(13), 1882–1883.
- Boyle, E. I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J. M., and Sherlock, G. (2004). GO::TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, **20**(18), 3710–3715.
- Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., Clark, N. R., and Ma’ayan, A. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*, **14**, 128.
- Fisher, R. A. (1935). Mathematics of a lady tasting tea. In *The Design of Experiments*. Oliver and Boyd, Edinburgh.
- Hejblum, B. P., Skinner, J., and Thiebaut, R. (2015). Time-Course Gene Set Analysis for Longitudinal Gene Expression Data. *PLoS Comput. Biol.*, **11**(6), e1004310.
- Jiao, X., Sherman, B. T., Huang, d. a. W., Stephens, R., Baseler, M. W., Lane, H. C., and Lempicki, R. A. (2012). DAVID-WS: a stateful web service to facilitate gene/protein list analysis. *Bioinformatics*, **28**(13), 1805–1806.
- Kalaitzis, A. A. and Lawrence, N. D. (2011). A simple approach to ranking differentially expressed gene expression time courses through Gaussian process regression. *BMC Bioinformatics*, **12**, 180.
- Khatri, P., Draghici, S., Ostermeier, G. C., and Krawetz, S. A. (2002). Profiling gene expression using onto-express. *Genomics*, **79**(2), 266–270.
- Lin, Z., Sanders, S. J., Li, M., Sestan, N., State, M. W., and Zhao, H. (2015). A Markov random field-based approach to characterizing human brain development using spatial-temporal transcriptome data. *Ann Appl Stat.*, **9**(1), 429–451.
- Lu, Y., Rosenfeld, R., Simon, I., Nau, G. J., and Bar-Joseph, Z. (2008). A probabilistic generative model for GO enrichment analysis. *Nucleic Acids Res.*, **36**(17), e109.
- McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., Wenger, A. M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.*, **28**(5), 495–501.
- Mi, H., Muruganujan, A., Casagrande, J. T., and Thomas, P. D. (2013). Large-scale gene function analysis with the PANTHER classification system. *Nat Protoc.*, **8**(8), 1551–1566.
- Robinson, M. D., Grigull, J., Mohammad, N., and Hughes, T. R. (2002). FunSpec: a web-based cluster interpreter for yeast. *BMC Bioinformatics*, **3**, 35.
- Szcurek, E. and Beerenwinkel, N. (2014). Modeling mutual exclusivity of cancer mutations. *PLoS Comput. Biol.*, **10**(3), e1003503.
- Zeeberg, B. R., Feng, W., Wang, G., Wang, M. D., Fojo, A. T., Sunshine, M., Narasimhan, S., Kane, D. W., Reinhold, W. C., Lababidi, S., Bussey, K. J., Riss, J., Barrett, J. C., and Weinstein, J. N. (2003). GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol.*, **4**(4), R28.