A model-based analysis to infer the functional content of a gene list

Michael A. Newton[1 2 3], Qiuling He[1], & Christina Kendziorski[2]

June 7, 2011

Technical Report #220, UW Biostatistics & Medical Informatics.

Abstract

An important challenge in statistical genomics concerns integrating experimental data with exogenous information about gene function. A number of statistical methods are available to address this challenge, but most do not accommodate complexities in the functional record. To infer activity of a functional category (e.g., a gene ontology term), most methods use gene-level data on that category, but do not use other functional properties of the same genes. Not doing so creates undue errors in inference. Recent developments in model-based category analysis aim to overcome this difficulty, but in attempting to do so they are faced with serious computational prob- lems. This paper investigates statistical properties and the structure of posterior computation in one such model for the analysis of functional category data. We examine the graphical structures underlying posterior computation in the original parameterization and in a new parameterization aimed at leveraging elements of the model. We characterize identifiability of the underlying acti- vation states, describe a new prior distribution, and introduce approximations that aim to support numerical methods for posterior inference.

[1]Department of Statistics; University of Wisconsin, Madison; 1300 University Avenue; Madison, WI 53706

[2]Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison

[3]To whom correspondence should be addressed: newton@stat.wisc.edu

1 Introduction

A common problem in statistical genomics concerns the points of contact between genomic data generated experimentally and exogenous functional information that has been accumulated by bioinformatics projects like GO and KEGG (The Gene Ontology Consortium, 2000; Kanehisa and Goto, 2000). In this rather extensive domain of data integration, functional information is used in two complementary ways. One mode is about data reduction. The experimentalist is faced with hundreds of genes that exhibit some interesting property in her experiment, and the inferential problem is to summarize the functional content of the identified genes. As a prime example, enrichment analysis seeks to identify functional categories that are over-represented in the experimentally identified gene list. Alternatively, functional categories are used to boost the signal-to-noise ratio. A weak gene-level signal differentiating two cellular states is easier to detect if it is consistent over a set of genes having some shared function. In either mode of application, the integration of experimental and functional data is a central component of the genomic data analysis.

A number of useful statistical methods and software tools have been developed to address the challenge. Fisher's exact test and related random-set enrichment methods operate conditionally on the experimental data and aim to detect over-representation of a category among experimentally interesting genes (*e.g.,* Drăghici *et al.*, 2003; Beiβbarth and Speed, 2004; Grossman *et al.* 2007; Newton *et al.*, 2007; Jiang and Gentleman, 2007; Bauer *et al.*, 2008; Sartor *et al.* 2009). Other approaches test category differential expression from replicated microarray data (*e.g.*, Barry *et al.*, 2005; Subramanian *et al.*, 2005; Efron *et al.*, 2007), while others develop models of gene-level results using functional categories (Lu *et al.* 2008; Bauer, *et al.* 2010). Careful comparisons among selected methods have helped to clarify their relative advantages and disadvantages (*e.g.*, Goeman and Bühlmann 2007; Barry *et al.* 2008). This list of citations hardly does justice to the field, and a detailed evaluation of the state-of-the-art is beyond the present scope. Suffice it to say that all methodological contributions in this domain have made simplifying assumptions on how the functional information relates to the experimental data on test. The continued expansion of the functional record makes some of these simplifications ever-more problematic.

Variation in category size makes it difficult to infer a prioritized list of significant functional categories. Methods that test either over-representation or category differential expression suffer from a power imbalance across categories owing to this variation. Power is related to size of both effect and category; large categories may deliver a small p-value by virtue of large size and small effect, while scientific relevance is linked more to the size of the effect. Thus ranking categories

by p-value tends to inflate the importance of large ones; while ranking them by an estimated effect tends to inflate the importance of small categories, since in these chance variation will more easily place them in a high ranking position.

As the functional record is complex and extensive, it necessarily encodes a substantial amount of overlapping information. GO organizes functional information in three directed acyclic graphs (biological process, molecular function, cellular component), wherein each graphical node is a functional category and directed edges convey proper-subset information. For example, the category *response to hydroperoxide* (GO:0033194) is a subset of *response to oxidative stress* (GO:0006979). It is less well appreciated that functional categories in GO overlap to a much greater extent than is suggested by any of the GO graphs. Of course overlaps among categories from different graphs are not immediately indicated, but there is also the issue that many pairs of categories share genes without one category being a proper subset of the other. A consequence of this phenomenon is that overlapping categories have positively correlated test results, often resulting in lists of significant functional categories that are unduly long (sometimes longer than an input list of significant genes!). An investigator may find that results of a statistical analysis have added relatively little insight because these results are muddied by complexities in the functional record that have been poorly accounted for.

Category overlap is related to the fact that many genes are multi-functional. The concept is called *pleitropy* in genetics, and it may be more the rule than the exception. For example, the PCNA1 gene (proliferating cellular nuclear antigen, 1) is involved in DNA mismatch repair; it plays another role in cell cycle regulation. At writing, 5056 human genes were annotated to 220 KEGG pathways, with over half these genes (2631) annotated to 2 or more pathways. Similarly, 14047 human genes were annotated to 13026 GO categories that contained between 1 and 500 genes, with a median number of 11 recorded functional properties per gene. (R package `org.Hs.eg.db`, version 2.4.6).

Category-differential-expression methods assert that a category is non-null if *any* of its contained genes is non-null. This basic premise is groundwork for the construction of test statistics and inference procedures, but it is at odds with the multi-functionality of genes. In the cellular state under experimentation, a gene may be non-null by virtue of one (or perhaps a subset) of its functions. A method which finds another of that genes' functions to be non-null may have inferred a spurious association. The presence of spurious associations unduly limits and complicates inference about the functional content of gene-level data. By way of analogy, suppose that we're watching a movie featuring an actor (*e.g.*, Mike Meyers in *The Spy Who Shagged Me*) who plays more than one character (Dr. Evil, Austin Powers, & Fat Bastard). And suppose further that our movie-watching skills are so limited that rather than being able to recognize what characters are in a given scene, we only

recognize the actors involved. Then of course the recognition that Mike Meyers is doing something interesting in the scene does not imply, for example, that Austin Powers is doing something interesting (maybe it is actually Dr. Evil)! In genomics we know that a gene can have different functional *roles* depending on the biological *scene* in which it plays a part. We may get closer to understanding that biology if our analytical methods are more in line with this fact.

Experimental data are measured on genes, while inference is required at the level of functional categories. Any legitimate method designed to infer something about a given functional category surely needs to use the experimental data on genes in that category. At issue is what other information ought to be used, and how that information should be incorporated into the calculations. Most category-inference methods are *global*: if they use any data beyond the gene-level data from the category on test, it is information from genome-wide summaries or summaries computed across the collection of categories. Basic enrichment methods, for example, use a genome-wide statistic on the proportion of genes that show some significant feature of interest. Many methods obtain category-specific p-values and then use the collection of p-values to get a false-discovery-rate correction. Global methods do not use specific information on category assignments of the genes in the category on test. We call a category-inference method *local* if, by contrast, it does use this functional information. Several local testing methods have been developed to utilize some overlap information (Jiang and Gentleman 2007; Grossman *et al.* 2007). Although useful, they suffer from inherent difficulties with sequential testing and they do not consider the full extent of category overlaps. Recently there has been a development of local category inference methods based on probability models of genetic and functional data (Lu *et al.* 2008; Bauer *et al.* 2010). These approaches are compelling because they address the overlap problem head-on and may provide an accurate representation of the the multivariate functional signal underlying observed data. They too, however, are limited by their computational complexity, by the nature of reported inferences, and by undue restrictions on gene-level data.

In the Bauer *et al.* (2010) model, non-null behavior starts with the functional category rather than the gene. Each gene inherits non-null behavior from non-null categories to which it is annotated. This is in contrast to the category-differential-expression methods, where a category is non-null if any of its contained genes is non-null. The apparently simple switch completely transforms the statistical problem. Inference on a given category relies on gene-level data on that category, but it also requires information on the other functional properties of these same genes, since any non-null behavior may be attributable to a different function than the one on test. This suggests that gene-level data from genes in overlapping categories are also relevant, but again their behavior may be affected by yet other categories to which they are assigned. We find ourselves in a complicated regress to infer the

state of a given functional category. Approximate inference is possible via Markov chain Monte Carlo sampling (MCMC). We appreciate the transformative effect of MCMC, but we also recognize limits on the ability to assess Monte Carlo error; one cannot be confident in inferences derived from slowly mixing chains operating in high dimensions. Even if convergence is assured, there are limitations in what can be inferred using marginal posterior summaries as in Bauer *et al.* (2010). Aspects of functional-category inference suggest that the posterior mode would also be useful to compute, though this is beyond the reach of MCMC in high-dimensions. We discuss the point further in Section 5.

The present paper initiates the development of probabilistic graphical modeling for functional-category inference. Probabilistic graphical modeling is a highly active field at the interface of statistics and machine learning (*e.g.*, Koller and Friedman, 2009). It considers how to organize and deploy inference computations derived from generative probability models for data using graphical structures and algorithms. Belief propagation algorithms (*e.g.*, the junction-tree algorithm) use message-passing schemes to represent the results of inferential calculations on subproblems. New algorithms that leverage advances in high-throughput computing enable message passing on large and complicated graphs (*e.g.*, Mendiburu *et al.* 2007; Gonzalez *et al.* 2009). In this paper, we examine the graphical structures underlying posterior computation, both in the original parameterization of Bauer *et al.* (2010), and in a new paramaterization that is designed to leverage simplifying elements of the model. We develop some theory to represent mappings between parameterizations; this has implications for posterior computation and it also clarifies identifiability and consistency issues. We introduce a new prior distribution designed to operate more naturally in the new parameterization. Finally, we investigate approximation schemes for reducing graph complexity and we present model extensions aimed at improving the performance of model-based local category inference. Our numerical experiments use functional-category information made available through the Bioconductor project (Gentleman *et al.* 2004).

2 The role model and the category intersection graph

The *role model* has potential in a number of domains, so it is described here using generic terminology. We have a number of different *parts* $p = 1, 2, \ldots, P$, and from these are formed a number of *wholes* $w = 1, 2, \ldots, W$. The parts comprising each whole are known in advance and recorded in a $P \times W$ incidence matrix $I = (I_{p,w})$, where $I_{p,w} = 1$ if part p is in whole w, else it is 0. We'll also say $p \in w$ if $I_{p,w} = 1$. Each whole is comprised of at least one part, and each part can be present in more than one whole. (In our case, parts are genes and wholes are functional categories.)

Experimental data are available on the parts, say $x = \{x_p\}$. Depending on the particular application the data may take various forms; either a vector of measurements across multiple samples, or a summary statistic of some kind. The simplest case has x_p the binary indicator of whether or not part p is reported on a short list of interesting parts. Observed data x are viewed as the realization of a random element X whose joint distribution depends on latent activation states $Z = \{Z_w\}$ of the wholes, which indicate whether each w is null ($Z_w = 0$), or non-null ($Z_w = 1$). We also use the language *active* and *inactive* to express $Z_w = 1$ or $Z_w = 0$, respectively. The simplest role model is:

$$Z_w \sim_{i.i.d.} \text{Bernoulli}(\pi) \tag{1}$$

$$X_p|\{Z_w = z_w\} \sim \text{Bernoulli}\left\{\alpha + (\beta - \alpha) \max_{w:p\in w} z_w\right\} \tag{2}$$

where α, β, and π are unknown parameters all in $(0,1)$, with $\alpha < \beta$. Additionally, the model asserts conditional independence among $\{X_p\}$ given $\{Z_w\}$. The statement in (2) says that X_p has rate β if any of the wholes to which it contributes is activated; otherwise it has rate α. Bauer *et al.* (2010) described this model for genes and categories, and proposed to rank what we call the wholes by MCMC-approximated marginal posteriors $P(Z_w = 1|X = x)$.

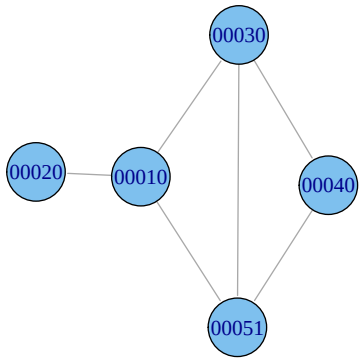
The prior (1) requires amendment in order to cope with general collections of wholes. For example, if a whole w' is fully contained in another w (as happens routinely in GO) then activities Z_w and $Z_{w'}$ ought to be related. In category inference, w' corresponds to a property that is more specific than w. To say "property w is activated" is to say "genes with property w are activated" from which it follows that "genes with property w' are activated", and thus "property w' is activated." Notice that the implication is not symmetric. If a subset is activated it does not follow that a containing set is activated. Indeed a goal of the inference is to assess the proper level of granularity regarding the activity states of the categories as evidenced by the apparent activity states of the genes. As inference considers activity as a property both of individual parts and of sets of parts, we require a clear definition of their relationship. The following assumption is key.

Activation hypothesis: A set of parts is active if and only if all parts in the set are active.

Various implications follow. A single part p is active if $p \in w$ for any w such that $Z_w = 1$. This precisely expresses model (2) and the interpretation of an active part as one delivering a higher success probability on Bernoulli data than a non-active part. Also, if the whole w is the union of various subsets; then all those subsets being active is equivalent to $Z_w = 1$. The activation hypothesis is equivalent

Figure 1: Category intersection graph from 5 KEGG pathways (the first five by ID order). Sets are of size 62, 32, 26, 25, and 33 genes, respectively. Edges in the graph indicate set overlap.

ID	functional category
00010	Glycolysis/Gluconeogenesis
00020	Citrate cycle (TCA cycle)
00030	Pentose phosphate pathway
00040	Pentose and glucuronate interconversions
00051	Fructose and mannose metabolism



to asserting that any subset of an active set is itself active. The hypothesis is related to the *true path rule* used in GO, to the extent that both convey logical constraints on collections of related categories. However, it seems not to have been expressed clearly in prior work. One might object to the activation hypothesis on the grounds that it is too strict, perhaps because it does not allow wholes to be activated by a subset of their parts. However, a sufficiently rich collection of wholes ought to include this relevant subset, and so if data point to activation of this subset, it is this subset that the inference procedure ought to detect (rather than the larger set). Furthermore, our language could get unduly complicated if we allow activated sets that contain no activated genes. A more important issue, however, concerns what we could ever hope to estimate about the whole-level activation states from part-level data. We take up the issue again in the next section and also in Section 5. For now, consider the set $\mathscr{Z}$ of valid activation-state vectors across the wholes

$$\mathscr{Z} = \left\{ z = (z_1, z_2, \ldots, z_W) \in \{0,1\}^W : z \text{ satisfies the activation hypothesis} \right\}.$$

Although the *i.i.d.* prior on $\{Z_w\}$ gives positive probability to vectors outside of $\mathscr{Z}$, certainly we can amend the prior by conditioning to enforce the activation hypothesis.

Returning to the statistical inference problem, we aim to develop Bayesian posterior computations over activation states $\{Z_w\}$ in order to express concisely the functional content of our gene-level data. We could apply the MCMC approach of Bauer *et al.* (2010), but we are concerned about Monte Carlo error and also the restriction to marginal posterior summaries. Sometimes, limitations of MCMC can be overcome by numerical methods from probabilistic graphical modeling. From this

perspective we start with a factorization of the joint posterior distribution into factors that have arguments localized on a certain undirected graph. Here we consider parameters α, β, and π in (1, 2) as fixed in order to simplify discussion. (Ultimately, we would like to estimate these from the data, and thus deploy empirical Bayesian computations, or possibly integrate them out.) The posterior distribution over whole-level activation states in the role model introduced above is:

$$\begin{aligned} p(z|x) &\propto p(z)\, p(x|z) \\ &= p(z) \prod_{p=1}^{P} p(x_p|z) \\ &= p(z) \prod_{p=1}^{P} p(x_p| \max_{w:p\in w} z_w) \\ &= p(z) \prod_{p=1}^{P} \left[\alpha^{x_p}(1-\alpha)^{1-x_p}\right]^{1-\max_{w:p\in w} z_w} \left[\beta^{x_p}(1-\beta)^{1-x_p}\right]^{\max_{w:p\in w} z_w} \end{aligned} \tag{3}$$

where $p(z)$ is the suitably conditioned *i.i.d.* Bernoulli(π) prior distribution. Although expressed as a product over parts, $p(z|x)$ also can be expressed as a product of data-dependent factors that are local functions on the *intersection graph* of the wholes. (The intersection graph has nodes equal to the wholes and edges between wholes that share parts.)

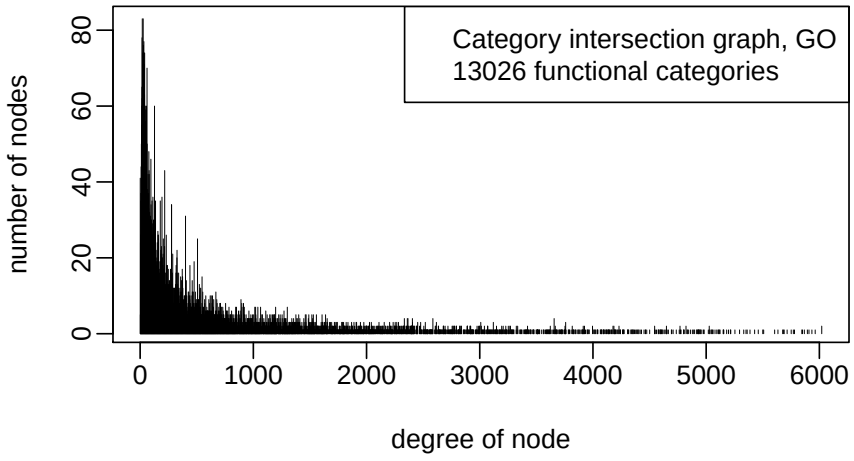
Proposition 1 *The role-model posterior in (3) satisfies:*

$$\mathrm{p}(z|x) \propto \prod_{w=1}^{W} \psi_w \left[z_w, z_{nb(w)}\right]. \tag{4}$$

where ψ_w is a data-dependent function of both z_w and neighboring states $z_{nb(w)} = \{z_{w'} : w \cap w' \neq \phi\}$.

A proof is in Section 5. Because the joint posterior factorizes into local functions over the intersection graph, this graph can be used, in principle, to support various inference computations implied by the role model. Figure 1 gives a simple example of the category intersection graph. Ideally, one would like to utilize the entirety of GO or KEGG. However the associated intersection graphs are highly complex and prohibit exact numerical methods (*e.g.*, Figure 2). Fortunately, inference in large-scale problems can proceed using alternative formulations or approximations, as we now discuss.

Figure 2: Degree distribution of intersection graph of GO (categories holding between 1 and 500 human genes). It is somewhat remarkable that so many overlaps are possible. The most extreme case is the category *cell motility* (GO:0048870), which annotates 495 human genes and shares genes with 6160 other categories among the 13026 GO categories that annotate between 1 and 500 human genes. These 13026 categories annotate 14047 genes. The median number of other category assignments per cell-motility gene is 64, and one gene happens to be in 631 other categories.



3 Reparameterization and the function profile graph

A reparameterization of the role model (1,2) offers another route to approximate inference. This reparameterization supports the same sampling model and it continues to rely on graphs to organize posterior computation, but in many cases it delivers simpler overall graph structure. Recall the incidence matrix I indicating which parts are in which wholes. Nothing so far disallows the possibility that different parts have the same rows in I. To proceed further it is helpful to consider the distinct rows of I, which we call *atoms*, following Boca *et al.* (2010). In category inference an atom corresponds to a particular profile of 0's and 1's across the functional record; it is the set of genes (parts) that have the same profile of category inclusions and exclusions. Each part p is an element of some atom. We say that a whole w is *assigned* to an atom v, and express this $w \to v$, if and only if $I_{p,w} = 1$ for all $p \in v$. Similarly $w \not\to v$ if and only if $I_{p,w} = 0$ for all $p \in v$. Indeed, the atom v is the intersection of wholes assigned to it and whole complements for wholes not so assigned. Thus, rather cryptically,

$$v = \left(\bigcap_{w:w\to v} w \right) \cap \left(\bigcap_{w:w\not\to v} w^c \right).$$

While wholes (categories) can overlap, atoms cannot. Furthermore, every whole is the union of atoms to which it is assigned:

$$w = \bigcup_{v:w\to v} v,$$

and in this way the atoms form a sort of basis for the collection of wholes. Table 1 shows an example. Boca *et al.* (2010) introduced atoms in a decision-theoretic analysis of the same basic data-integration problem. Their aim was somewhat different from ours, in that they sought a subset of atoms (rather than functional categories) whose activation could explain gene-level data.

While the functional record is becoming ever more complex, the number of atoms is bounded by the number of genes, and this is far smaller than the theoretical maximum 2^W. In other words, the vast majority of functional profiles do not manifest themselves. This feature is one reason why considering the role model from the atom perspective has potential advantages. To pursue this, we first construct atom-specific activation Bernoulli variables from the activation states of the wholes:

$$A_v = 1 - \prod_{w:w\to v} (1 - Z_w) = \max_{w:w\to v} Z_w \tag{5}$$

Table 1: Eleven atoms from the example shown in Figure 1 where there are 5 wholes (each a KEGG pathway). The atom entry gives the unique row of the incidence matrix I associated with the involved genes. For example, there are 4 genes involved in both *Glycolysis/Gluconeogenesis* and *Fructose and mannose metabolism* (the first and last pathways) and not involved in the other three.

atom	# genes	atom	# genes
00011	1	10100	3
00110	1	01000	25
10101	8	00001	20
11000	7	00100	14
00010	23	10000	40
10001	4		

Again, the atom is activated if any of the wholes to which its parts are assigned is activated. The range of mapping (5) is

$$\mathscr{A} = \{a = (a_1, a_2, \ldots, a_N) : a = a(z), z \in \mathscr{Z}\}, \tag{6}$$

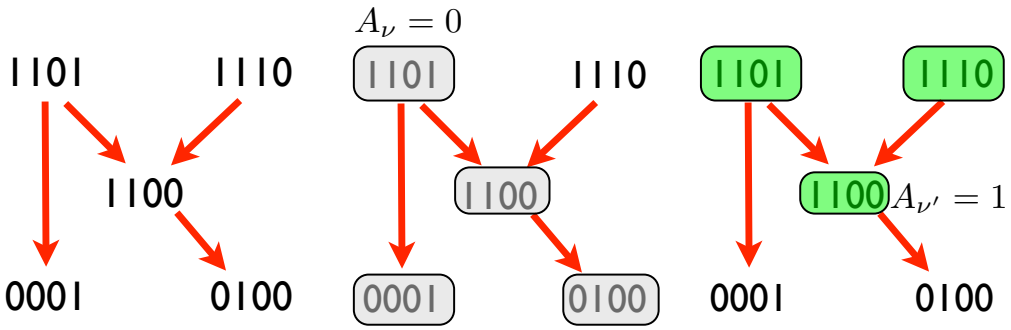
where N is the number of atoms. The notation is intended to convey the set of all atom-level activation vectors a that could have been produced from whole-level activation vectors z which satisfy the activation hypothesis. Indeed this property is convenient, because, as we prove in Section 5:

Proposition 2 *The mapping (5) from $\mathscr{Z}$ to $\mathscr{A}$ is one-to-one, and has inverse*

$$Z_w = \prod_{v:w\to v} A_v = \min_{v:w\to v} A_v.$$

A computational strategy is supported by this finding. We perform posterior computations over atom-level activations in $\mathscr{A}$, and then transform findings back to the whole-level of interest. The finding also supports the identifiability of whole-level activation states from part-level data. If part-level data were to increase, then we would consistently estimate the atom-level activation states. Thus we would consistently estimate the whole-level activation states by Proposition 2. Without the activation hypothesis, there could be states that are beyond our ability to estimate, regardless of the amount of part-level data.

Figure 3: Reparameterizing the role model with a *function profile graph*: The nodes in each panel represent five atoms. Each atom shows a profile of assignments (1) or not (0) to four wholes w. A directed edge goes from v to v' if the assignments at v include those at v' (except we omit redundant edges *e.g.*, no edge from 1110 to 0100.) The middle and right panels show logical dependencies on activity variables. *E.g.,* in the middle panel, knowing $A_v = 0$ implies $A_{v'} = 0$ for all downstream atoms, and knowing $A_{v'} = 1$ on the right panel implies $A_v = 1$ for all upstream atoms.



Rather conveniently, the role-model posterior distribution (3) can be re-expressed on the transformed scale as:

$$p(a|x) \propto p(a) \prod_{v=1}^{N} p(x_v|a_v) \tag{7}$$

where $x_v = \sum_{p \in v} x_p$ summarizes the part-level data at atom v, and where $p(a)$ is a prior distribution. Conditionally upon the activation states, x_v is the realization of a Binomial random variable, based on $n_v = \sum_{p \in v} 1$ trials (*i.e.*, the atom size). Thus (7) simplifies further

$$p(a|x) \propto p(a) \left(\frac{1-\beta}{1-\alpha}\right)^{\sum_v n_v a_v} \left[\frac{\beta(1-\alpha)}{\alpha(1-\beta)}\right]^{\sum_v x_v a_v}. \tag{8}$$

Just as the intersection graph of the wholes is the data structure supporting posterior inference in the original parameterization, there is another graph – we call it the *function profile graph* – that supports atom-level computations. Its nodes are the atoms. One might try having an edge between v and v' if a common whole w is assigned to both, but this is more than we need. Instead, we create a directed edge from v to v' if: (1) the assignments at v' are a proper subset of the assignments

at ν, and also (2) there is no other atom ν^* with assignments that are a subset of assignments at ν and a superset of assignments at ν' (Figure 3). We say ν is a parent of ν' and ν' is a child of ν.

The relevance of the function profile graph becomes more apparent when we occupy the nodes with atom-level activity variables A_ν. We see that the edges of the function profile graph express role-model information. For example, knowing $A_\nu = 0$ implies that for no w assigned to ν do we have $Z_w = 1$. Naturally this forces $A_{\nu'} = 0$, when ν' is a child of ν, since assignments to ν' are a subset of those going to ν. By the same token, knowing that $A_{\nu'} = 1$ is equivalent to knowing that at least one w assigned to ν' has $Z_w = 1$, which forces $A_\nu = 1$ when ν is a parent of ν'. Essentially, the logic of atom-level activations is encoded by the function profile graph. Let $\mathscr{A}^*$ denote all possible binary activation vectors $a = (a_1, a_2, \ldots, a_N)$ that respect the function profile graph in the sense above; *i.e.*,

$$\begin{aligned} a_\nu = 0 &\implies a_{\nu'} = 0 \qquad \text{for all children } \nu' \text{ of } \nu \\ a_{\nu'} = 1 &\implies a_\nu = 1 \qquad \text{for all parents } \nu \text{ of } \nu' \,. \end{aligned} \tag{9}$$

Curiously, the collection $\mathscr{A}$ in (6) does not necessarily constitute all of $\mathscr{A}^*$, though we do have $\mathscr{A} \subset \mathscr{A}^*$. (See Section 5.) Importantly, the mapping $a \to \left\{ z_w = \min_{\nu : w \to \nu} a_\nu \right\}$ from $\mathscr{A}^*$ does map onto the original set of activity vectors $\mathscr{Z}$.

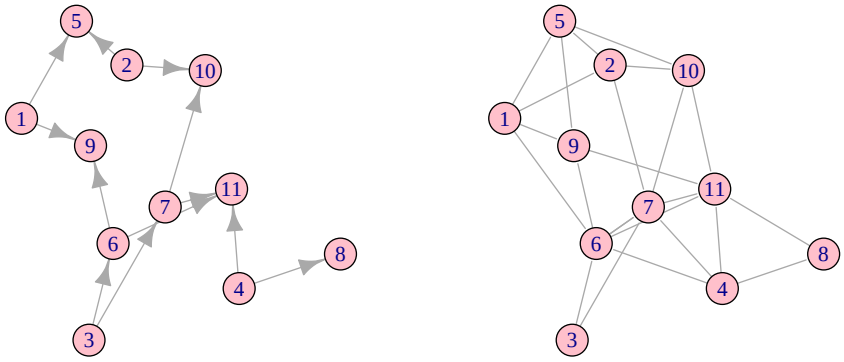
Part of the computational complexity in the original parameterization stems from the fact that the category intersection graph allows an arbitrary function of Z_w on neighboring nodes to affect the state at a given node (*i.e.,* the ψ_w in (4)). But model (2) encodes a very specific function (through max), which is used to advantage in the proposed reparameterization. There is an effect on graph properties, which in some cases leads to simpler posterior computations.

To support inference we need an undirected version of the function profile graph, which we obtain by a form of *moralization* used in graphical models analysis. Specifically, we include an undirected edge between any two nodes ν and ν' that are both parents of a common child. We also include an undirected edge between any two nodes ν and ν' that are children of a common parent. (This two-way moralization comes from the fact that information flows both ways along a given directed edge.) Finally we make all remaining directed edges undirected. The resulting graph is the undirected function profile graph. An example is given in Figure 4.

Proposition 3 *For a suitable prior $p(a)$ over $\mathscr{A}^*$, the posterior distribution in (8) is the product of functions $\tilde{\psi}_\nu$ that are local in the undirected function profile graph:*

$$p(a|x) \propto \prod_{\nu=1}^{N} \tilde{\psi}_\nu \left[a_\nu, a_{\mathrm{nb}(\nu)} \right] \tag{10}$$

Figure 4: Function profile graphs for the small KEGG example shown in Figure 1, with 11 atoms as listed in Table 1.



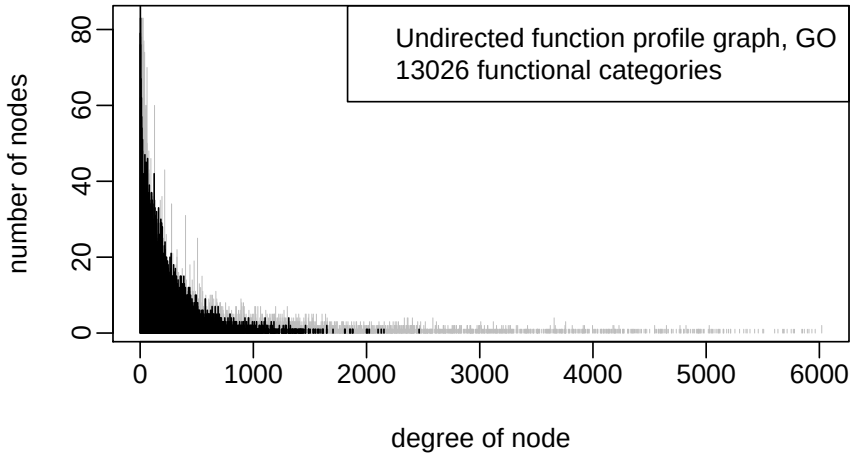
Coupled with Proposition 2, the above result indicates that we can perform inference computations on the function profile graph, and then transform back as needed to get inference on whole-level activation states. In GO, for example, the transformation provides a much simpler graph (Figure 5). Unfortunately this simpler graph is still too complicated for exact numerical methods. Approximation methods discussed in the next section offer several approaches to address this challenge.

4 Approximations and graph-based computations

Filtering categories: Instead of including the entirety of GO or KEGG in a role-model computation, we could select a smaller set of categories based on an initial filter. For example, we could filter by marginal p-value from an enrichment test. We investigated this approach using three gene lists obtained by Keller *et al.* (2008) in a murine study of diabetes. Using microarrays, this study profiled genome-wide expression of of islet (Data A), adipose (Data B), and gastrocnemius (Data C) cells, among others not shown. Of interest were genes exhibiting co-expression within each tissue; co-expression modules holding 85, 150, and 114 genes, respectively, were identified for followup. Role-model computations address the functional content of these lists. We use the lists here simply to demonstrate how much graph simplification can be achieved by filtering.

Using a normal approximation to Fisher's exact test, as implemented in the R package *allez* (Newton *et al.* 2007), we considered GO and KEGG categories holding no more than 500 genes, and two p-value cut-offs ($p = 0.01, p = 0.001$). Table 2 summarizes the complexity of the category intersection graph and the function

Figure 5: Degree distribution of the undirected function profile graph of GO (categories holding between 1 and 500 human genes). The maximal degree is 2464; the graph itself has 10366 nodes (atoms). The corresponding results for the category intersection graph (from Figure 2) are repeated here in grey. Not shown are results for the directed function profile graph, which is much simpler, having maximal degree 268.



profile graph derived from these data-dependent category collections. The maximal degree of the function profile graph is usually smaller than the maximal degree of the category intersection graph, with graph complexity substantially reduced compared to the case of no filtering. Even so, the graphs remain too complex for the deployment of exact numerical methods. One solution strategy is to approximate the functional record itself, as we discuss next.

Ablating annotations: There is a class of approximation schemes that work by modifying the incidence matrix I to have fewer non-zero entries. We describe one such *ablation* scheme that retains a fraction of each part's column assignments, preferentially retaining assignments to small wholes. The rationale is that a small whole is more proximal to a part than a large one, and so its data, on the average, may be more relevant to the state of that part than the data from a larger whole. Without loss of generality, suppose that the columns of I (*i.e.*, the wholes or categories) are organized in increasing order of size. Fix a retention parameter $\rho \in (0,1]$. Create a new incidence matrix $\tilde{I}$ of the same dimension as I, initially with $\tilde{I} = I$. Working one part p (*i.e.,* row) at a time to update $\tilde{I}$, let $n_p = \sum_{w=1}^{W} I_{p,w}$

Table 2: Graph properties when filtering GO and KEGG according to marginal enrichment p-value. For each data set, the rows correspond to filtering at $p = 0.01$ and $p = 0.001$, respectively. The number of nodes in the intersection graph equals the number of categories that are significant by the marginal test. max D and mean D refer to the maximal and mean degree of the graph. The function profile graph has more nodes (atoms) than the intersection graph and also fewer edges per node, on the average, but similar maximal degree. Data sets are discussed in the text.

	intersection graph			function profile graph				
					undirected		directed	
Data	# nodes	max D	mean D	# nodes	max D	mean D	max D	mean D
A	465	394	149.3	2190	409	72.1	85	7.9
	348	285	106.9	1553	284	55.8	46	7.4
B	313	234	69.7	1190	195	32.4	44	6.1
	264	196	58.9	922	132	29.3	38	5.9
C	398	328	106.4	2015	355	55.1	56	7.1
	280	197	61.1	1123	187	37.2	34	6.4

denote the number of wholes containing part p. If $I_{p,w^*} = 1$, set $\tilde{I}_{p,w^*} = 0$ if

$$\sum_{w=1}^{w^*} I_{p,w} > \rho n_p \tag{11}$$

with the caveat that every p be retained to at least one whole. The resulting incidence matrix is more sparse than the original, and it produces ever simpler graphs as the retention rate ρ is reduced. Table 3 shows the effects of ablation on the function profile graph of KEGG. Results for GO are similar (not shown).

This ablation scheme increases the average size of atoms and reduces the complexity of the function profile graph. Ablation does not remove part-level data from the system, nor does it remove functional categories. Rather, heavily annotated parts are simplified and they convey their effect directly to the smallest wholes containing them. Consider, for example, two overlapping wholes w_1 and a larger one w_2. With $\rho < 1$, the atom $w_1 \cap w_2$ is affected. The annotation of those parts to the larger whole is ablated, and the parts (and their data) are delivered to the smaller whole. In this way the ablated incidence matrix delivers posterior computations over a reduced collection of atom activations; from results of that inference, we can trace back to the real atoms and infer activations over the real functional categories.

Table 3: Ablating annotations in KEGG. We start with the 5052 genes annotated to 219 KEGG pathways containing no more than 500 genes. The category-defining incidence matrix is ablated as in (11), and properties of the undirected function profile graph are obtained. The right-most column shows the size of the largest clique in a triangulated version of the undirected function profile graph.

ρ	# atoms	genes per atom	max D	mean D	max clique
1	1093	4.6	214	34.3	436
1/2	735	6.9	110	20.4	249
1/4	395	12.8	41	7.8	59
1/8	260	19.4	24	2.9	13
1/16	220	23.0	8	1.3	5

Graph-based computations: Our premise is that numerical methods from probabilistic graphical modeling can support posterior computation for the role model. We are motivated partly by the discrete nature of functional-category inference and partly by advances in this domain of statistical computing. Our theoretical considerations suggest what graphs might be used, and our numerical experiments provide some insight into the properties of these graphs for GO and KEGG. Very little has been said so far about the actual calculations and how these need to be organized. We make a few brief remarks here.

There are several ways to organize exact *belief propagation* algorithms. By one route, the supporting undirected graph is the conditional independence graph associated with the joint posterior under consideration. This graph is triangulated (every cycle of four or more nodes has an edge between non-adjacent nodes) by adding edges if necessary, and then its cliques (maximal complete subgraphs) are found. A *junction tree* is formed, with nodes equal to these cliques, and with edges between these nodes that satisfy the running intersection property. That is to say, if a node from the original graph is in any two cliques (nodes in the tree), then it is in every clique-node on the unique intervening path in the tree. This property is key for subsequent algorithms to properly marginalize activation states inside the graph. A number of technical issues affect this computational sequence, but they are routinely addressed using graphical algorithms. Inference proceeds via *message passing*. In a simple approach to computing the marginal posterior distribution of a variable in some particular tree node, we make that node the root of the tree and we send messages towards that root from any ready nodes. A node is ready after it has received messages from all its neighbors that are distal from the root. The messages themselves are vectors holding conditional probabilities of data in the distal nodes conditional on the activation states at the node receiving the message. By

the rules of probability, outgoing messages are computed by summing over certain latent activation states, and it is this component of the computation that is very sensitive to graph complexity. At some point the exact posterior computation requires manipulating 2^M sums, where M is the size of the largest clique represented in the junction tree: hence our interest in the maximal clique size of the triangulated graph (Table 3). Evidently, exact computations are feasible using the atom transform in an approximate version of the problem in which we ablate weakly informative annotations.

In *loopy belief propagation* we give up on exact posterior computation. We do not attempt to triangulate the original graph, find cliques, or form a junction tree. One approach uses *factor graphs*, which are bipartite graphs having nodes for factors and other nodes for arguments of those factors (Kschischang *et al.* 2001). We emphasized the factor structure of posterior distributions in both Proposition 1 and Proposition 3 because it is relevant to loopy belief propagation on the factor graph. Edges go between arguments and any factors in which they participate, and so the degree structure of the factor graph is essentially the same as the degree structure of the undirected graphs we have thus far considered. Approximate posterior computation proceeds by transmitting conditional probability messages along edges of the factor graph. Without further intervention, the complexity of these computations is exponential in the maximal degree of the graph (rather than clique size). Advances such as in Mendiburu *et al.* (2007) and Gonzalez *et al.* (2009) indicate that accurate and computationally efficient algorithms may be feasible on large and complex graphs, such as those we have with model-based inference and functional categories.

5 Proofs and notes

Proposition 1: The posterior is proportional to the prior $p(z)$ times the likelihood $p(x|z) = \prod_{p=1}^{P} p(x_p|z)$. Taking the likelihood first, factors $p(x_p|z) = p(x_p|\max_{w^*:p\in w^*} z_{w^*})$ that involve a given whole w are from all those parts $p \in w$, and thus depend on the activation states z_{w^*} for any other wholes w^* that also contain those p's; that is $z_{nb(w)}$. There is not a unique assignment of these part-based factors to whole-based factors ψ_w in (4), but any such assignment must allow the possibility that at most z_w and the neighboring activation states contribute to ψ_w.

By independence, Bauer's *i.i.d.* Bernoulli(π) prior for the Z_w's factorizes over the category intersection graph. It remains to confirm that such factorization continues when we condition each realization z to satisfy the activation hypothesis $1[z \in \mathscr{Z}]$. The activation hypothesis is equivalent to saying that any subset of

an activated set is active, which is a combination of properties of sets and their neighboring subsets.

Proposition 2: Let $a \in \mathscr{A}$ denote a vector $a = (a_1, a_2, \ldots, a_N)$ of atom-level activation states. This vector results from mapping, through (5), some vector $z = (z_1, z_2, \ldots, z_W) \in \mathscr{Z}$ of whole-level activation states satisfying the activation hypothesis. Suppose we have another point $z^* \in \mathscr{Z}$ for which $z^* \neq z$ and z^* also maps to the same vector a. If we reach a contradiction then no such z^* exists, and the mapping is one-to-one.

As we are fixing the vector a, we can partition the atoms into those ν for which $a_\nu = 0$ and those for which $a_\nu = 1$. Call these respective index sets V_0 and V_1. First consider $\nu \in V_0$. By supposition and definition of a_ν,

$$\prod_{w:w\to\nu} (1 - z_w) = \prod_{w:\to\nu} (1 - z^*_w) = 1. \tag{12}$$

Thus all sets w assigned to ν must have $z_w = z^*_w = 0$. That is, $z_w = z^*_w$ at all wholes w assigned to any ν for which $a_\nu = 0$.

Next consider some $\nu \in V_1$. In contrast to (12), we have

$$\prod_{w:w\to\nu} (1 - z_w) = \prod_{w:\to\nu} (1 - z^*_w) = 0. \tag{13}$$

Either side can be zero by virtue of any one of the factors, and so we do not immediately get $z_w = z^*_w$. However, we can eliminate from both sides of (13) any factors $(1 - z_w) = (1 - z^*_w) = 1$ corresponding to sets w already considered above that map to some other atom ν' with $a_{\nu'} = 0$. Then (13) reduces to

$$\prod_{\{w:w\to\nu, w\not\to\nu'\in V_0\}} (1 - z_w) = \prod_{\{w:w\to\nu, w\not\to\nu'\in V_0\}} (1 - z^*_w) = 0. \tag{14}$$

Any w in this set $\{w : w \to \nu, w \not\to \nu' \in V_0\}$ may be comprised of multiple atoms, but all of them are in V_1 and thus are activated, like ν itself ($a_\nu = 1$). Since w is equals a union of activated atoms, it must be activated, by the activation hypothesis. That is, for all w in (14), $z_w = z^*_w = 1$. By applying this argument to all $\nu \in V_1$, we complete the proof that $z_w = z^*_w$ for all wholes, and thus mapping (5) is one-to-one. The inversion formula encodes the rule that any subset of an activated set of parts is activated.

Proposition 3: From (8) the posterior $p(a|x)$ is proportional to a prior $p(a)$ times a product of atom-specific (likelihood) factors. Thus it suffices to find a prior $p(a)$ that is local on the undirected function profile graph. We have restricted the domain

to vectors $a = (a_1, a_2, \ldots, a_N)$ in $\mathscr{A}^*$, according to (9). Note that this restriction can be presented as a local function on the function profile graph:

$$1[a \in \mathscr{A}^*] = \prod_{v=1}^{N} B_{1,v}\left[a_v, a_{\text{children}(v)}\right] B_{2,v}\left[a_v, a_{\text{parents}(v)}\right]$$

where $B_{1,v}$ and $B_{2,v}$ encode the two constraints in (9). Various priors are possible. A simple one entails *i.i.d.* Bernoulli(π) atom activities A_v that are then conditioned to be in $\mathscr{A}^*$.

There are problems in trying to use Bauer's *i.i.d.* Bernoulli(π) prior on the Z_w's to induce a prior over $\mathscr{A}$. For one, we needed to amend this prior so that the Z_w's satisfy the subset inheritance problem. A larger issue is that the induced distribution may not be local on the function profile graph. For example, a given Z_w might be assigned to two atoms that are unconnected in the function profile graph. Choice of prior has an effect on the computations.

We mentioned in Section 3 that for some collections of wholes, $\mathscr{A}^* \neq \mathscr{A}$. As an example, consider three wholes made from three parts, with incidence matrix

$$I = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}.$$

Every pair of wholes (columns) overlaps, so the category intersection graph is complete. But no profile of assignments is a subset of any other, so the function profile graph has no edges. The atom-level activation vector $(1,0,0)$ is not a possible result of any whole-level activations, since activating any sets would activate two or three atoms.

Marginal posterior inference versus MAP inference: Even if MCMC convergence is assured, there is a problem in using it to drive inference about non-null functional categories. Suppose that $\mathscr{C}_{\text{true}}$ holds all the truly non-null categories, and $\mathscr{C}_{\alpha,\text{marg}}$ holds a list of estimated non-null categories, estimated by *marginal* computations set up to target a false discovery rate of $\alpha = 5\%$, say. Typically, this would go by calling a category non-null if its MCMC-estimated *marginal* posterior probability of being null is less than α. Considering the positive association of related GO categories (in terms of gene content), and considering a potential sparsity in the true signal $\mathscr{C}_{\text{true}}$, it is quite likely that related categories will be *negatively* associated in the joint posterior distribution, given experimental data. Simply put, if taking one category to be non-null explains the non-null'ness of some gene-level data, then there is no incentive for a related category to be non-null. As a consequence of this negative posterior association, there will be a discordance between

marginal findings in an FDR-controlled list and the actual state of $\mathscr{C}_{\text{true}}$. The true *joint* state may be much simpler (*i.e.*, many fewer non-null categories), but measuring this state is not within the reach of MCMC for even a moderately-sized problem. Arguably, the joint state is better estimated in this case by the maximum a posteriori (MAP) estimate, which is the Bayes estimate under 0-1 loss. The MAP estimate may be associated with a high level of posterior uncertainty, as Bauer *et al.* (2010) argue, but its relative simplicity may be useful for providing a concise summary of functional content. Ideally the analyst is able to compute both marginal posterior summaries and MAP summaries to make the most informed inferences.

Extending the role model: Bauer's model is limited by its restriction to binary gene-level data and by an assumed homogeneity of responses within the activated and inactivated classes. Inactivated states all deliver conditionally independent responses with a common success probability α, and activated states similarly deliver responses with success probability $\beta > \alpha$. This constrains the atom level counts $x_v = \sum_{p \in v} x_p$ to be Binomially distributed given the activation states. A more flexible approach within the same general framework allows each part p to have its own Beta distributed success probability; then atom counts x_v are more broadly distributed as Beta-binomial counts. In place of two basic parameters α and β we need four parameters to encode the activated and inactivated Beta distributions; this seems to be a small price for the added flexibility. Posterior computations may also benefit from the flattening out of the posterior distribution over activation states.

Software: Tools in `R` (version 2.12.1) were used throughout. For annotation information, we used Bioconductor packages `org.Hs.eg.db` and `org.Mm.eg.db`, both versions 2.4.6. For graph computations we used `igraph` version 0.5.5-1 (Csardi and Nepusz, 2006), `RBGL` version 1.26.0 (Carey *et al.* 2010), and `gRbase` version 1.3.4 (Ren *et al.* 2010).

Acknowledgements

The authors thank Giovanni Parmigiani, Simina Boca and their co-authors for sharing a preprint of Boca *et al.* (2010). This research was funded in part by R01 grants ES017400 and GM076274 from the National Institutes of Health, and by a fellowship from the Morgridge Institute of Research.

References

Barry, WT, Nobel, AB, and Wright, FA (2005). Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics*, 21, 1943-1949.

Barry, WT, Nobel, AB, and Wright, FA (2008). A statistical framework for testing functional categories in microarray data. *Annals of Applied Statistics*, 2, 286-315.

Bauer, S, Grossman, S, Vingron, M, Robinson, PN (2008). Ontologizer 2.0– a multifunctional tool for GO term enrichment analysis and data exploration. *Bioinformatics*, 24, 1650-1651.

Bauer, S, Gagneur, J, and Robinson, PN (2010). GOing Bayesian: model-based gene set analysis of genomic-scale data. *Nucleic Acids Research*, 38, 3523-3532.

Beiβbarth, T and Speed, TP (2004). GOstat: Find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics*, 20, 1464-1465.

Boca, S, Corrada Bravo, H, Leek, JT, and Parmigiani, G (2010). A decision-theory approach to set-level inference for high-dimensional data. Johns Hopkins University Biostat Working Paper 211.

Carey, V, Long, Li, and Gentleman, R. RBGL: An interface to the BOOST graph library, R package version 1.26.0, *www.bioconductor.org*.

Csardi, G and Nepusz, T (2006). The igraph package for complex network research. *InterJournal, Complex Systems*, 1695, *igraph.sf.net*.

Drăghici, S, Khatri, P, Martins, RP, Ostermeier, GC, and Krawetz, SA (2003). Global functional profiling of gene expression. *Genomics*, 81, 98-104.

Efron, B, and Tibshirani, R (2007). On testing the significance of a set of genes. *Annals of Applied Statistics*, 1, 107–129.

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5(10), R80.

The Gene Ontology Consortium (2000). Gene ontology: tool for the unification of biology. *Nature Genetics*, 25, 25-29.

Goeman, JJ, and Bühlmann, P (2007). Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23, 980–987.

Gonzalez, J, Low, Y, and Guestrin (2009). Residual Splash for optimally parallelizing belief propagation. In, *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS)*, Clearwater Beach Florida, USA, Volume 5 of JMLR.

Grossman, S, Bauer, S, Robinson, PN, and Vingron, M (2007). Improved detection of overrepresentation of Gene-Ontology annotations with parent child analysis.

Bioinformatics, 23, 3024-3031.

Jiang, Z and Gentleman, R (2007). Extensions to gene set enrichment. *Bioinformatics*, 23, 306-313.

Kanehisa, M and Goto, S (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28, 27-30.

Keller, MP, Choi, Y, Wang, P, Belt Davis, D, Rabaglia, ME, Oler, AT, Stapleton, DS, Argmann, C, Schueler, KL, Edwards, S, Steinberg, HA, Chaibub Neto, E, Kleinhanz, R, Turner, S, Hellerstein, MK, Schadt, EE, Yandell, BS, Kendziorski, C, and Attie, AD (2008). A gene expression network model of type 2 diabetes links cell cycle regulation in islets with diabetes susceptibility. *Genome Research*, 18, 706-716.

Koller, D and Friedman, N. *Probabilistic graphical models*. MIT Press, 2009.

Kschischang, FR, Frey, BJ, and Loeliger, HA (2001). Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47, 498-519.

Lu, Y, Rosenfeld, R, Simon, I, Nau, GJ, and Bar-Joseph, Z (2008). A probabilistic generative model for GO enrichment analysis. *Nucleic Acids Research*, 36, e109.

Mendiburu, A., Santana, R, Lozano, JA, Bengoetxea, E (2007). A parallel framework for loopy belief propagation. In, *GECCO '07: Proceedings of the 2007 GECCO conference companion on genetic and evolutionary computation*, pages 2843–2850.

Newton, MA, Quintana, FA, den Boon, JA, Sengupta, S, and Ahlquist, P (2007). Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. *Annals of Applied Statistics*, 1, 85-106.

Ren, SA, Jsgaard, HA, Dethlefsen, C, and Bowsher (2010). gRbase: A package for graphical modelling in R. `R package version 1.3.4 CRAN.R-project.org/package=gRbase`

Sartor, MA, Leikauf, GD, Medvedovic, M (2009). LRpath: a logistic regression approach for identifying enriched biological groups in gene expression data. *Bioinformatics*, 25, 211-217.

Subramanian, A, Tamayo, P, Mootha, VK, Mukherjee, S, Ebert, BL, Gillette, MA, Paulovich, A, Pomeroy, SL, Golub, TR, Lander, ES, and Mesirov, JP (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102, 15545–15550.