frontiers in
GENETICS

ORIGINAL RESEARCH ARTICLE
published: 23 February 2012
doi: 10.3389/fgene.2012.00024

Methods for determining the statistical significance of enrichment or depletion of gene ontology classifications under weighted membership

Ernesto Iacucci[1,2]*, **Hans H. Zingg**[3] **and Theodore J. Perkins**[4,5]

[1] Department of Electrical Engineering ESAT-SCD, Katholieke Universiteit Leuven, Leuven, Belgium
[2] IBBT-KU Leuven Future Health Department, Katholieke Universiteit Leuven, Leuven, Belgium
[3] Department of Physiology, McGill University, Montreal, QC, Canada
[4] Department of Biochemistry, Immunology and Microbiology, University of Ottawa, Ottawa, ON, Canada
[5] Ottawa Hospital Research Institute, Ottawa, ON, Canada

Edited by:
Radhakrishnan Nagarajan, University of Arkansas for Medical Sciences, USA

Reviewed by:
Chi Zhang, University of Nebraska Lincoln, USA
Preetam Ghosh, Virginia Commonwealth University, USA

****Correspondence:***
Ernesto Iacucci, IBBT-KU Leuven Future Health Department, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, Box 2446, 3001 Leuven, Belgium.
e-mail: ernesto.iacucci@gmail.com

High-throughput molecular biology studies, such as microarray assays of gene expression, two-hybrid experiments for detecting protein interactions, or ChIP-Seq experiments for transcription factor binding, often result in an "interesting" set of genes – say, genes that are co-expressed or bound by the same factor. One way of understanding the biological meaning of such a set is to consider what processes or functions, as defined in an ontology, are over-represented (enriched) or under-represented (depleted) among genes in the set. Usually, the significance of enrichment or depletion scores is based on simple statistical models and on the membership of genes in different classifications. We consider the more general problem of computing *p*-values for arbitrary integer additive statistics, or weighted membership functions. Such membership functions can be used to represent, for example, prior knowledge on the role of certain genes or classifications, differential importance of different classifications or genes to the experimenter, hierarchical relationships between classifications, or different degrees of interestingness or evidence for specific genes. We describe a generic dynamic programming algorithm that can compute exact *p*-values for arbitrary integer additive statistics. We also describe several optimizations for important special cases, which can provide orders-of-magnitude speed up in the computations. We apply our methods to datasets describing oxidative phosphorylation and parturition and compare *p*-values based on computations of several different statistics for measuring enrichment. We find major differences between *p*-values resulting from these statistics, and that some statistics recover "gold standard" annotations of the data better than others. Our work establishes a theoretical and algorithmic basis for far richer notions of enrichment or depletion of gene sets with respect to gene ontologies than has previously been available.

Keywords: gene ontology, weighted membership, enrichment, depletion, dynamic programming

INTRODUCTION

High-throughput studies often generate large sets of "interesting" genes that must be further researched from available annotation resources. This research may encompass literature reviews and querying various databases. This process leads to a multitude of information to consider for each gene as well as inconsistent gene descriptions from one resource to another. The application of a statistic to a particular annotation resource can identify the statistically significant classifications common to the set of interesting genes (Man et al., 2000; Berriz et al., 2003; Castillo-Davis and Hartl, 2003; Diaz-Uriarte and Dopazo, 2003; Doniger et al., 2003; Draghici et al., 2003; Hosack et al., 2003; King et al., 2003; Al-Shahrour et al., 2004; Zhang et al., 2004; Khatri and Draghici, 2005). In this paper we generalize previous work by allowing genes to have weighted memberships in classifications.

Weighted membership functions offer far more expressive power than traditional binary membership functions, as we demonstrate below. They allow us to represent complicated relationships within an ontology, distinguish good from weak representatives of a classification, account for the strength of evidence from a high-throughput study, or more generally account for relationships between different data sets. Thus, the extension of gene ontology (GO) analysis to weighted membership is an important one.

The problem of determining statistical significance under weighted membership is more complicated than under binary membership. There is a computational "cost" to the generality we gain. In this sense, it is reminiscent of the well-known knapsack problem. The knapsack problem is trivially solvable if all items have unit weight. If items have arbitrary weights, it is NP-complete, although a pseudopolynomial-time dynamic program can usually solve realistic instances quite efficiently. In similar fashion, we derive a pseudopolynomial-time dynamic programming algorithm for computing exact *p*-values for the statistical significance of enrichment or depletion of classifications under weighted membership. Moreover, we show that for several specific weighted membership functions, the algorithm can be specialized to gain greater computational efficiency. We apply these methods to two test data sets, where we show that weighted membership functions are better able to extract biologically meaningful classifications than traditional binary analysis.

GENE ONTOLOGY

Our methods are not restricted to any particular ontology, but because of its popularity, we focus on the well-known GO. GO (The Gene Ontology Consortium, 2000) is a controlled, hierarchically organized vocabulary that aims to provide a consistent description of genes to the biological community, along with relationships between these descriptions. GO includes three principal classifications: *molecular function*, *biological process*, and *cellular component*, as well as many more specific classifications. Molecular functions are the biochemical roles of genes, for example, *GTPase activity*. Biological processes involve one or more molecular functions over multiple distinguished *steps*. An example of this would be *mitosis*, where each interaction of proteins in the progression of the cell cycle would be considered a step. Cellular components are described as the localization of a protein in the cell or its membership in a complex. An example would be the proteins that are members of the ribosomal complex.

The relationships between classifications define a directed acyclic graph (DAG), in which vertices in the DAG correspond to classifications and a directed edge from classification X to classification Y means that Y is a direct specialization of X. Some authors assume that there is no more general classification than the three principal classifications. We assume a root vertex *gene ontology*, with the principal classifications as children. Having defined a root, we therefore discuss the GO DRAG (Directed Rooted Acyclic Graph) in the rest of this paper. All vertices except the root have at least one parent, and may have more. *Hexose biosynthesis*, for example, has two parents, *hexose metabolism* and *monosaccharide biosynthesis* (The Gene Ontology Consortium, 2000).

Genes may be *mapped* to one or more classifications. A mapping is the association of a gene to a classification. *Cytochrome c* is one such example; it is mapped to the molecular function classification *electron transporter activity*, the biological process classifications *oxidative phosphorylation* and *induction of cell death*, and cellular component classifications *mitochondrial matrix* and *mitochondrial inner membrane.* As parent vertices are less specific classifications than their child vertices, genes mapped to the child vertices are understood to be also described by the parent vertices. We say that gene X is a *member* of classification Y if X is directly mapped to Y or if it is mapped to any classification that specializes Y – that is, to any classification which is reachable in the DRAG from Y.

While GO provides classifications that can be used to describe the genes in an interesting set, it does not provide guidance on which of the classifications are of interest to researchers for further investigation. The number of classifications to consider can be reduced by focusing on classifications that are *enriched* or *depleted* as defined in a statistical framework (Draghici et al., 2003; Zeeberg et al., 2003).

WEIGHTED MEMBERSHIP IN CLASSIFICATIONS

Suppose a gene g has been flagged as interesting based on a laboratory experiment. For example, gene g may be differentially expressed in a microarray experiment or may have a specific post-transcriptional modification. Now consider a classification v. Does g provide evidence, and if so, how much, for the involvement of classification v in the condition under study? The simplest heuristic would be to say that g provides evidence for v if and only if g is a member of v (Draghici et al., 2003; Zeeberg et al., 2003). However, one may also consider weighted membership scores, which might account for relationships between classifications, degrees of differential expression, or other knowledge. For example, consider the DRAG and gene mappings depicted in **Figure 1**. Gene g_2 is mapped to classifications v_7 and v_8. If g_2 is flagged as interesting, we do not know if it is because of the involvement of v_7, v_8, or both. In any case, however, the flagging of g_2 indicates the involvement of classification v_4. So, we might interpret g_2 as providing stronger evidence, or greater weight, for v_4 than either of v_7 and v_8. Then again, if g_5 is also flagged, and has much greater differential expression than g_2, we might decide the strongest evidence is for classification v_6. Conversely, some genes in a classification many never be differentially expressed, and that should not be taken as evidence against the classification. For example, in signaling or metabolic pathways, it is common for many proteins to be ubiquitously expressed at a basal rate, whether the pathway is active or not. One might want to give such genes small or even zero membership weight in that pathway – for the purpose of determining which pathways are represented in the interesting set of genes – so that when the genes that are regulated change their expression, the activity in the pathway is more apparent. Thus, we wish to be able to determine significance for more general notions of evidence, or weighted membership functions.

Given a DRAG and a population set of genes G, we define a weighted membership function $\Phi : G \times V \rightarrow Z^+$, where V is the set of all classifications and Z^+ is the set of non-negative integers. $\Phi(g, v)$ is always zero if g is not a member of v, but otherwise $\Phi(g, v)$ is unconstrained. The standard membership function is an important special case, which we will denote by $\Phi_1(g, v) = \{1$ if $g \in v$; 0 if $g \notin v\}$. We will consider two other weighted membership functions, although many interesting choices are possible. We define $\Phi_2(g, v)$ as the number of paths in the ontology DRAG from v to any vertex v' to which g is mapped. This is intended to account for situations such as for g_2 in the example of the previous paragraph. In that example, $\Phi_2(g_2, v_7) = \Phi_2(g_2, v_8) = 1$, but $\Phi_2(g_2, v_4) = 2$, indicating that, all other things being equal, we have greater confidence that g_2 is representative of classification v_4 than it is of either v_7 or v_8 under any particular circumstance. For Φ_3 we assume access to an integer-valued vector X, where $X(g)$ might represent, for example, a discretized measure of differential expression from a high-throughput gene expression experiment. We define $\Phi_3(g, v) = \{X(g)$ if $g \in v$; 0 if $g \notin v\}$.

Ultimately, we want to assess the enrichment or depletion of a set of genes with respect to a classification. For this purpose, any weighted membership function Φ can be generalized to take as input a set of genes H and a classification, simply by summing the scores of the individual genes. That is, $\Phi(H, v) = \Sigma_g \in H \Phi(g, v)$.

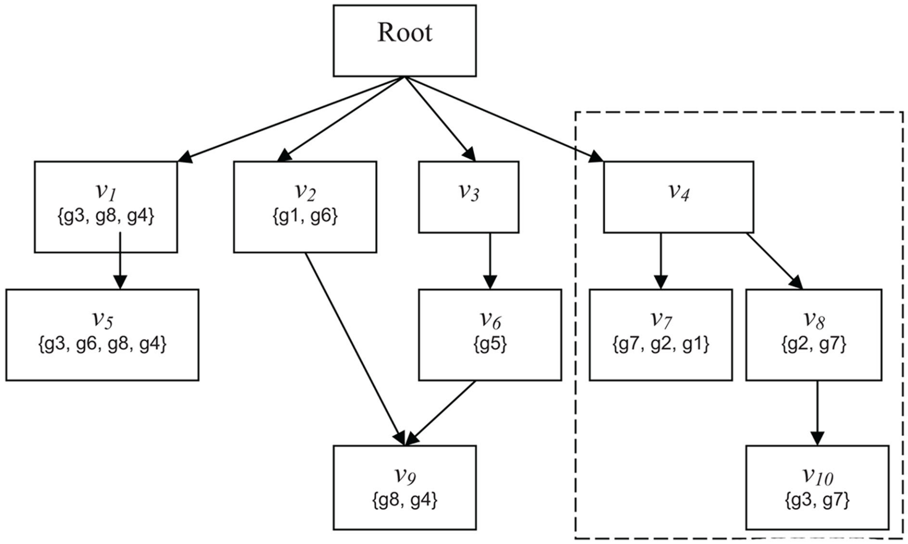


FIGURE 1 | Schematic of GO. Abstract example of an ontology illustrating the principles of the GO DRAG and the scoring functions Boxes labeled with *Root* or n correspond to classifications Sets represent the genes directly mapped to each classification. The broken line indicates the subgraph rooted at *v – i*.

P-VALUES FOR THE SIGNIFICANCE OF ENRICHMENT OR DEPLETION

We assume now that a laboratory experiment has flagged a set of interesting genes *H*, and we want to know whether the set *H* provides evidence for a classification *v*. We choose a weighted membership function Φ_A, and want to know whether $\Phi_A(H, v)$ is significantly greater than would be expected by chance. To define this precisely, we must account for properties of the interesting set, *H*. If *H* is large, for example, we generally expect higher classification scores, and thus the threshold for significant enrichment ought to be higher. More generally, we allow conditioning not just on the size of *H*, but on the score of the set *H* under any weighted membership function Φ_B evaluated at the root classification. Thus, let H' represent a random subset of *G*, which we take to be uniformly distributed over all subsets of *G* satisfying $\Phi_B(H', r) = \Phi_B(H, r)$, where *r* is the root classification. In general, we consider *p*-values for enrichment of classification *v* of the form:

$$\text{Prob}\left(\Phi_\text{A}\left(H', v\right) \geq \Phi_\text{A}\left(H, v\right) \mid \Phi_\text{B}\left(H', r\right) = \Phi_\text{B}\left(H, r\right)\right) \qquad (1)$$

where Φ_A and Φ_B can be any two weighted membership functions and *r* is the root classification. We use the shorthand notation $\Phi_A \mid \Phi_B$ to denote this *p*-value. (See **Figure A1** in Appendix for some worked examples based on **Figure 1**.) The *p*-value for depletion simply has the inequality reversed.

In the special case that $\Phi_A = \Phi_B = \Phi_1$, the condition $\Phi_B(H', r) = \Phi_B(H, r)$ implies that H' is restricted to sets of the same size as *H* and $\Phi_A(H', v)$ is just the number of genes in H' that are members of *v*. The *p*-value is then just the widely used membership *p*-value, which can be computed via the hypergeometric equation as (Draghici et al., 2003; Zeeberg et al., 2003):

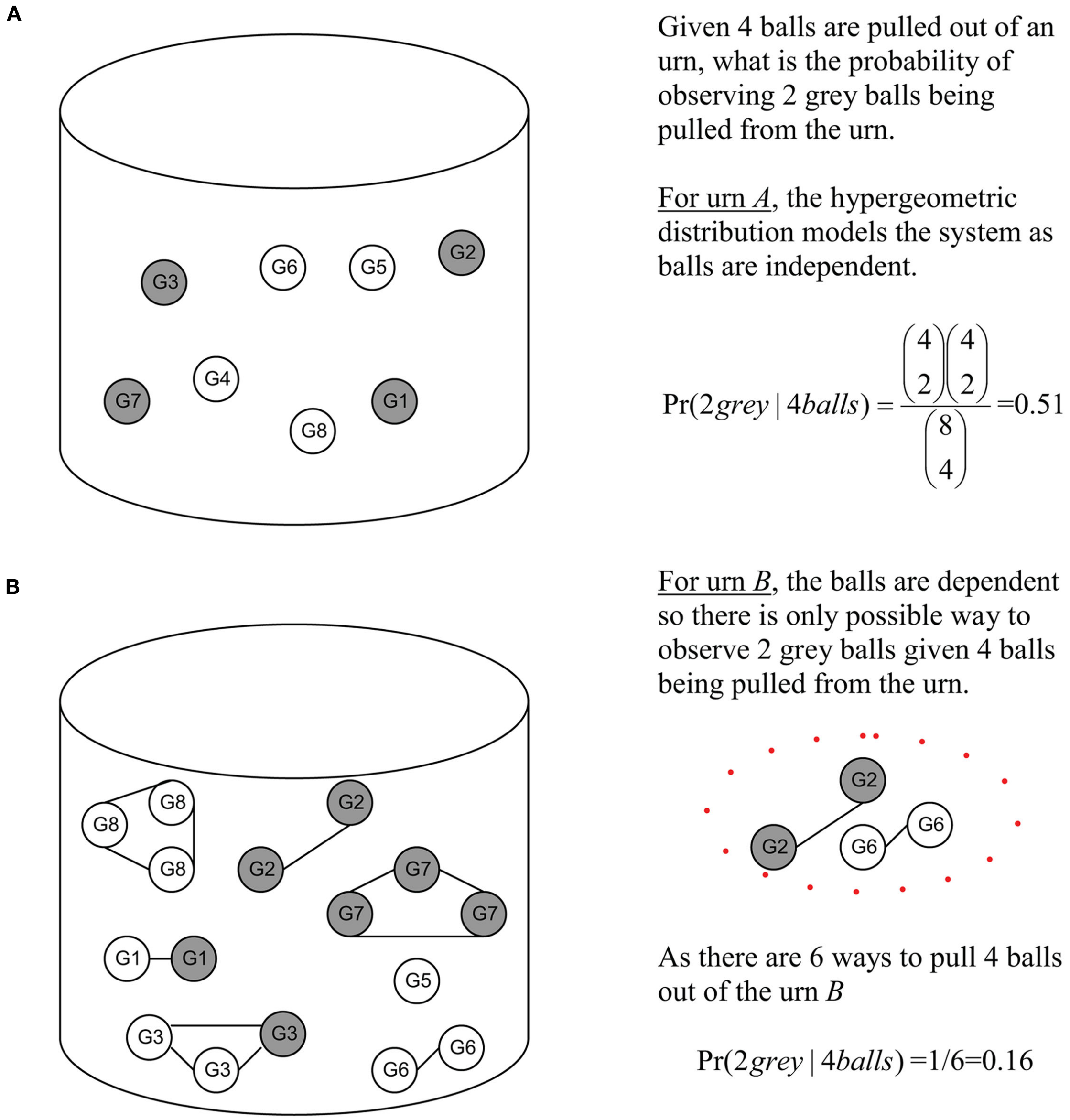
$$\sum_{n=\Phi_1(H,v)}^{c} \frac{\binom{N}{n}\binom{M}{m}}{\binom{M+N}{c}} \qquad (2)$$

where *N* is the number of genes that are members of *v*, *M* is the number of genes in *G* that are not members of *v*, $c = |H|$, and $m = c - n$. We can also write this without reference to *N*, *M*, or *c* using the substitutions $N = \Phi_1(G, v)$, $M = \Phi_1(G, r) - \Phi_1(G, v)$ and $c = \Phi_1(H, r)$. More generally, if $\Phi_B = \Phi_1$ but Φ_A is not equal to Φ_1, then the *p*-value represents the chance of getting a score $\Phi_A(H', v) \geq \Phi_A(H, v)$ if H' is uniformly random from all sets of genes of the same size as *H*.

The simple formula for $\Phi_1 \mid \Phi_1$ is often introduced in the context of ball-and-urn models. One imagines *N* red balls and *M* blue balls mixed together in an urn, and then *c* balls are drawn from the urn randomly but without replacement. This metaphor fails for general $\Phi_A \mid \Phi_B$ for two reasons. First, because membership is weighted, the balls, corresponding to genes, do not have equal significance. Second, because a gene *g* may have $\Phi_A(g, v) > 0$ and $\Phi_A(g, v) \neq \Phi_B(g, r) > 0$, the balls may be not just strictly red or blue but some combination of both. There is no simple formula for $\Phi_A \mid \Phi_B$ in general, but in the next section we describe dynamic programming algorithms for computing this *p*-value.

Due to representation of GO as a DRAG and the mapping of genes to GO, an exact method for measuring such *p*-values has never been proposed. More precisely, the main challenges presented by the GO model are the lack of independent and unique genes. As genes can be mapped to one or more classifications, they are not *unique*. For example, in **Figure 1**, g_6 is mapped to both v_2 and v_5. Due to this feature of the model, genes can be reachable in both a subgraph rooted at *v* and in its complement $\bar{v}$. Thus, such genes are not *independent* as selecting them in *v* will also require that they be selected in $\bar{v}$. For example, in **Figure 1**, g_1 is reachable in both the subgraph rooted at v_4 (mapped to v_7) and at its complement (mapped to v_2).

Looking at **Figure 2**, one can better appreciate the difference in an unweighted versus a weighted model. Unweighted values



Given 4 balls are pulled out of an urn, what is the probability of observing 2 grey balls being pulled from the urn.

For urn *A*, the hypergeometric distribution models the system as balls are independent.

$$\Pr(2grey \mid 4balls) = \frac{\binom{4}{2}\binom{4}{2}}{\binom{8}{4}} = 0.51$$

For urn *B*, the balls are dependent so there is only possible way to observe 2 grey balls given 4 balls being pulled from the urn.

As there are 6 ways to pull 4 balls out of the urn *B*

$$\Pr(2grey \mid 4balls) = 1/6 = 0.16$$

FIGURE 2 | Urn representation of Figure 1 diagram illustrating the principles of independent versus interrelated models. Gray balls (genes) are those found at or below v_4 and white balls (white) are those found in the rest of the DRAG. **(A)** Gray and white balls (genes) in urn arc independent. **(B)** Gray and while (genes) in urn are interrelated, the GO DRAG and the scoring functions. The solid lines indicate that balls (genes) are related.

can be seen as the balls in urn *A*. As these balls are independent, the possible observations can be easily modeled by the hypergeometric. Under a weighted model, this is not possible due to the interrelated nature of the balls. The balls in urn *B* are dependent as they are attached by a string, hence the hypergeometric model cannot be used to model observations of randomly picking balls from urn *B*. Thus, the probability of observing two gray balls given four balls were selected from urn *B* is 0.16.

Our primary technical contribution is to describe how to determine *p*-vales given a model where one cannot assume gene weights to be independent or unique. Precise computation of the *p*-values allows greater confidence in biological assertions made based on the statistical modeling.

MATERIALS AND METHODS

Exact computation of the *p*-value (Eq. 1) is NP-complete, for example, by reduction from the 0 to 1 knapsack problem. Like the 0–1 knapsack problem, however, it can be solved in pseudopolynomial-time by a dynamic programming method. We first describe a generic dynamic programming method for computing the *p*-value, and then describe several improvements and special cases that can significantly speed the computation.

NAÏVE DYNAMIC PROGRAMMING APPROACH

We construct a two-dimensional table *T*, where $T(a, b)$ represents the number of subsets H' of G for which $\Phi_A(H', v) = a$ and $\Phi_B(H', r) = b$. The main loop is over genes g in G. The update rule for the table (see line five below), $T(a+\Phi_A(g, v), b+\Phi_B(g, r)) \leftarrow T(a, b)+T(a+\Phi_A(g, v), b+\Phi_B(g, r))$, is based on the observation that any subset H' satisfying the constraints $\Phi_A(H', v)=a$ and $\Phi_B(H', r)=b$ either includes the gene g or does not include g. Values for a range from 0 to $A_{\max}=\Phi_A(G, v)$, and values for b range from 0 to $B_{\max}=\Phi_B(G, r)$. Line 8 of the code below accounts for a set H' made up of only the gene g. To simplify the presentation of the code, we have allowed it to access T outside of these ranges; such values should always be taken to be zero.

```
(1) Initialize: $T(a, b) \leftarrow 0$ for all $a$, $b$
(2) For each $g \in G$
(3)     For $a = 0$ to $A_{\max}$
(4)         For $b = 0$ to $B_{\max}$
(5)             $T(a+\Phi_A(g, v), b+\Phi_B(g, r)) \leftarrow T(a, b) +$
                    $T(a+\Phi_A(g, v), b+\Phi_B(g, r))$
(6)         End
(7)     End
(8)     $T(\Phi_A(g, v),\Phi_B(g, r)) \leftarrow T(\Phi_A(g, v),\Phi_B(g, r)) + 1$
(9) End
```

The desired p-value, $\Phi_A \mid \Phi_B$, is the computed as:

$$\frac{\sum_{a=\Phi_A(H,v)}^{\Phi_A(G,v)} T\left(a, \Phi_B(H, r)\right)}{\sum_{a=0}^{\Phi_A(G,v)} T\left(a, \Phi_B(H, r)\right)} \tag{3}$$

The time complexity of this algorithm, measured by the number of table updates, is $O(|G| A_{\max}\ B_{\max})$. The table entries can easily become larger than integer or long integer precision, and so it is necessary to use arbitrary precision arithmetic. Nevertheless, this computation is feasible for typical problems in which $|G|$ may be of the order 10^3 or 10^4 and $A_{\max}$ and $B_{\max}$ may be of the order 10^3–10^5. This gives the p-value for enrichment of only a single classification. If p-values for multiple, even all, of the classifications are sought, the time complexity is multiplied by the number of classifications being tested. For particular choices of Φ_A and Φ_B, more efficient dynamic programs can be developed. This is the subject of the following subsections.

IMPROVEMENTS TO THE NAÏVE DYNAMIC PROGRAM

Several changes to the naïve dynamic program can drastically increase its computational and storage efficiencies. First, one only needs the first $\Phi_B(H, r)$ columns of the table T, as it is the last column that is used in the computation of the p-value. Although a simple observation, this can be a dramatic improvement if $\Phi_B(H, r)$ is small. For example, if $\Phi_B=\Phi_1$ and considering that typically $|H|=|G|$, one could expect to save at least an order of magnitude computation and storage.

Second, many genes g may have the same weights, $\Phi_A(g, v)$ and $\Phi_B(g, r)$. We define the weight count W_i,j to be the number of genes g with $\Phi_A(g, v)=i$ and $\Phi_B(g, r)=j$. These groups of genes can be treated simultaneously – if k of these genes are included in a subset H', then the contribution to $\Phi_A(H', v)$ and $\Phi_B(H', r)$ is the same regardless of precisely which k are chosen. If we add an innermost loop over k to our dynamic program, the main update rule above (step 5) can thus be restated as:

$$T\left(a+i^{*}k,\ b+j^{*}k\right) \leftarrow T\left(a, b\right)^{*}\binom{W_{i,j}}{k}+T\left(a+i^{*}k,\ b+j^{*}k\right) \tag{4}$$

A third optimization is possible in the case that $\Phi_A(g, v)=0$ for many genes g. We can split the dynamic program into two tables, T^1 and T^2. $T^1(a, b)$ is the number of subsets H' of genes with $\Phi_A(H', v)=a$ and $\Phi_B(H', r)=b$, where H' is restricted to contain only genes for which $\Phi_A(g, v)>0$. T^2 is a one-dimensional table in which $T^2(b)$ is the number of subsets H' of genes with $\Phi_B(H', r)=\text{b}$, where H' is restricted to contain only genes for which $\Phi_A(g, v)=0$. From these two tables, the original $T(a, b)$ can be reconstructed as $T(a, b)=\Sigma_b{}'\ T^1(a, b')*T^2(b-b')$.

These improvements can be used independently or combined for the greatest effect, and may even be further improved in some particular cases. In the next subsection we discuss three particular p-values we use in our experiments in relation to these improvements to the dynamic program.

DYNAMIC PROGRAMS FOR $\Phi_2 \mid \Phi_1$, $\Phi_2 \mid \Phi_2$ AND $\Phi_3 \mid \Phi_3$

There are many possible choices for Φ, and many combinations, resulting in different p-values for measuring the statistical significance of enrichment or depletion of GO classifications. Recall that in Section "*P*-Values for the Significance of Enrichment or Depletion" we defined several possible choices of Φ (see also **Table 1** for a notational reminder): Φ_1 is the standard binary membership function, $\Phi_2(g, v)$ measured the number of paths in the DRAG from v to any classification v' to which g is mapped, and $\Phi_3(g, v)$ is equal to some global "score" (e.g., differential expression) assigned to g if g is mapped to v, and otherwise is equal to zero. Here, we derived specialized, more efficient dynamic programs for computing p-values involving these three choices of Φ. The first

Table 1 | Summary of notation used in this paper.

r	Root of the ontology
v	Vertex (classification) in the ontology
V	Set of all vertices
$\bar{v}$	The complement of v
g	Gene
G	Set of all genes
H	Set of interesting genes (Subset of G)
H'	Random subset of G
$\Phi(g, v)$	Weighted membership of gene g in classification v (any non-negative integer)
$\Phi_1(g, v)$	Counts the membership in the DRAG from v to any classification v' to which g is mapped
$\Phi_2(g, v)$	Measures the number of paths in the DRAG from v to any classification v' to which g is mapped
$\Phi_3(g, v)$	Is equal to some global "score" (e.g., differentiation expression) assigned to g if g is mapped to v
$\Phi(H, v)$	Sum of weighted memberships in classification v over all genes g in set H

is $\Phi_2 \mid \Phi_1$. Conditioning on Φ_1 is perhaps most natural when H is determined experimentally (say by a microarray study), because the size of H often depends on some arbitrary significance cut-off or other practical considerations. Secondly, we consider $\Phi_2 \mid \Phi_2$, which may be more natural when genes of interest are not determined experimentally but based on some other property – for example, sharing a particular transmembrane domain. Finally we consider $\Phi_3 \mid \Phi_3$, which reflects an intuitively compelling situation where one wishes to make use of expression level measurements, which are common to many genomics studies. We describe the implementations of these cases below.

We use all three optimizations described in Section "Improvements to the Naïve Dynamic Program" to compute $\Phi_2 \mid \Phi_1$. In the two-table decomposition, the second table does not actually need to be computed, as $T^2(b)$ is just M choose b, where M is the number of genes that are not members of classification v. The weight counts W_i,j are zero for $j \neq 1$, so we can drop the j subscript and refer simply to W_i. Let I be the maximum value of i for which W_i is positive. The pseudocode for computing table T_1 is as follows:

```
Initialize $T_1(a, b) = 0$ for all $a, b$
For $i = 1$ to $I$
    For $b = 1$ to $|H|$
        For $a = 1$ to $A_{max}$
            For $k = 1$ to $W_i$
                $T(a + k^*i, b + k) \leftarrow T(a, b)^* \binom{W_i}{k} + T(a + k^*i, b + k)$
            End
        End
    End
For $k = 1$ to $W_i$
    $T(k^*i, k) \leftarrow T(k^*i, k) + \binom{W_i}{k}$
    End
End
```

To compute $\Phi_2 \mid \Phi_1$, the summation $\Sigma_b{}' T^1(a, b')*(M$ choose $(b - b'))$ can be substituted for $T(a, b)$ in Eq. 3. However, the denominator is simply ($|G|$ choose $|H|$) in this case, so we can simply compute $\Phi_2 \mid \Phi_1 = \Sigma_a \Sigma_b{}' T^1(a, b')*(M$ choose $(b - b'))$, where the a in the sum goes from $\Phi_2(H, v)$ to A_{max}, and divide by ($|G|$ choose $|H|$).

For $\Phi_2 \mid \Phi_2$ we need to compute tables T^1 and T^2 explicitly. The straightforward computation of T^1 would be wasteful because for any subset of genes H′ we have $\Phi_2(H', r) \geq \Phi_2(H', v)$. Thus, all entries of T^1 below the main diagonal would be zero. Instead, we slightly redefine T^1 so that the columns correspond to $\Phi_2(H', r) - \Phi_2(H', v)$ instead of just $\Phi_2(H', r)$. For simplicity, we omit the weight count optimization, so that the pseudocode for computing T^1 is then:

```
Initialize $T^1(a, b) \leftarrow 0$ for all $a, b$
For each gene $g$ with $\Phi_2(g, v) > 0$
    For $b = 0$ to $\Phi_2(G, r) - \Phi_2(G, v)$
        For $a = 1$ to $\Phi_2(G, v)$
            $T^1(a + \Phi_2(g, v), b + (\Phi_2(g, r) - \Phi_2(g, v))) \leftarrow$
              $T^1(a, b) +$
              $T^1(a + \Phi_2(g, v), b + (\Phi_2(g, r) - \Phi_2(g, v)))$
        End
    End
    $T^1(\Phi_2(g, v), \Phi_2(g, r) - \Phi_2(g, v)) \leftarrow$
      $T^1(\Phi_2(g, v), \Phi_2(g, r) - \Phi_2(g, v)) + 1$
End
```

The computation of T^2 is straightforward. $T^2(b)$ is the number of ways of making $\Phi_2(H', r)$ using only genes g, for which $\Phi_2(g, v) = 0$. We use the weight count trick to speed to computation, defining J to be the maximum j for which $W_{0,j} > 0$. The pseudocode for creating the table is as follows:

```
Initialize $T^2(b) \leftarrow 0$ for all $b$
For $j = 1$ to $J$
    For $b = 1$ to $\Phi_2(G, r)$
        For $k = 1$ to $W_j$
            $T^2(b + j^*k) \leftarrow T^2(b)^* \binom{W_j}{k} + T^2(b + j^*k)$
        End
    End
    For $k = 1$ to $W_j$
        $T^2(j^*k) \leftarrow T^2(j^*k) + \binom{W_j}{k}$
    End
End
```

The p-value $\Phi_2 \mid \Phi_2$ can be computed using Eq. 3 with the substitution $T(a, b) = \Sigma_b{}' T^1(a, b' - a)^* T^2(b - b')$. The p-value $\Phi_3 \mid \Phi_3$ can be computed similarly.

DATA SOURCES

To test our approaches, we used the OXPHOS-CR (oxidative phosphorylation co-regulated) dataset (53 genes) from (Subramanian et al., 2005). The dataset can be characterized as containing OXPHOS genes that are tightly co-regulated across many tissues. Three other datasets were derived from this dataset, each assigned 100, 300, and 500 additional random genes to mimic noisy datasets.

We also applied the four methods of determining significance described above ($\Phi_1 \mid \Phi_1$, $\Phi_2 \mid \Phi_1$, $\Phi_3 \mid \Phi_3$, and $\Phi_2 \mid \Phi_2$) to a publicly available dataset concerning changes in gene expression in the rat uterus at different stages of parturition (Girotti and Zingg, 2003). The Girotti study used rat Genome U34A GeneChips (Affymetrix) to measure five time-points in rat pregnancy: day 0, day 20, at term but not in labor, at term and in labor, and postpartum. The Girotti study presents a literature search for genes showing differential expression in at least one time-point comparison. Filtering criteria applied in the study resulted in 4137 genes (Girotti erroneously reported the number as 4138) forming the population set of genes, of which 431 interesting genes (nonredundant genes that were differentially expressed between day 0 and any other time-point) were annotated using various annotation resources (SwissProt, ExPASy, and other databases) into 17 user-defined classifications. In our study, the GO DRAG was limited to vertices to which the 4137 genes (the set G) are mapped. The 431 genes were used to define the interesting set H. The GO definitions were supplied by GO for the U34A GeneChip (Affymetrix; (The Gene Ontology Consortium, 2000). For Φ_3, $X(g) = 2$ for average fold change over the first two time-points, 1 for less than average fold change (less than $X - 2$ SD), and 3 for higher than average fold change (greater than $X + 2$ SD).

RESULTS AND DISCUSSION

In order to evaluate the quantitative performance of the methods, we ran them on the OXPHOS-CR dataset and took the 35 classifications that were annotated to this dataset to be a "golden standard." In practice, a set of interesting genes will not only be composed of purely interesting genes but also noise (false positives). We therefore ran our methods on the "noisy" datasets that were derived from the golden standard by the addition of random genes (+100 genes, +300 genes, +500 genes).

The results, shown in **Table 2**, show that the $\Phi_2 \mid \Phi_2$ method was able to maintain the ranking of golden standard classifications better than the $\Phi_1 \mid \Phi_1$ (traditional hypergeometric) or the $\Phi_2 \mid \Phi_1$ methods when applied to the noisy datasets. **Figure 3** provides the ROC plot of the rankings of the classifications for the "noisy" datasets resulting from the application of all three methods. Looking at the added 500 random genes dataset (**Figure 3A**), it is clear that the $\Phi_2 \mid \Phi_2$ method provides the best ranking of the classifications. Looking again at **Table 2**, one can see that the methods differ with respect to specificity and sensitivity (with statistical significance threshold set at $p \le 0.05$). Overall, the $\Phi_2 \mid \Phi_1$ and $\Phi_1 \mid \Phi_1$ had good specificity but poor sensitivity. In contrast the $\Phi_2 \mid \Phi_2$ had much better sensitivity than the other two methods but variable specificity. Likewise, $\Phi_2 \mid \Phi_2$ consistently had a better AUC than the other two methods.

In order to further evaluate the performance of the methods, we used the Girotti and Zingg (2003) study, which identified several user-defined classifications without the use of any statistical framework. In **Table 3**, we look at the analogous GO classifications and compare p-values from the four methods described above. Many of these classifications are significantly enriched at $p \le 0.05$ according to at least one of the methods, or are weakly significant. With regard to the biological relevance of the classifications, *defense response*, *cell motility*, *extracellular matrix*, and *cytoskeleton* maintained low p-values across all methods. This is consistent with the parturition process, as it is involves immune suppression and cell remodeling. There are, however, some surprising disagreements between the methods. For example, p-values for *protein modification* range over four orders-of-magnitude, going from 4.2×10^{-5} ($\Phi_2 \mid \Phi_2$) to 0.923 ($\Phi_3 \mid \Phi_3$). (See **Figure A2** in Appendix for scatter plots of different pairs of p-values across all classifications.) It was surprising that p-values could differ so greatly between the methods. **Table 4** shows how many classifications are identified as significantly enriched by each method individually (main diagonal) and by each pair of methods (off diagonal) using the p-value threshold of $p < 0.05$.

Table 2 | Comparison of methods with added noise.

	$\Phi_2 \mid \Phi_1$	$\Phi_2 \mid \Phi_2$	$\Phi_1 \mid \Phi_1$
100 "NOISY" GENES ADDED			
AUC	0.6969	0.7367	0.7107
Sensitivity	0.4857	0.8857	0.3142
Specificity	0.9488	0.3352	0.9261
300 "NOISY" GENES ADDED			
AUC	0.7179	0.8002	0.7867
Sensitivity	0.4571	0.8000	0.4571
Specificity	0.9240	0.5379	0.9438
500 "NOISY" GENES ADDED			
AUC	0.7241	0.8124	0.7754
Sensitivity	0.4000	0.7714	0.4571
Specificity	0.9401	0.7356	0.9476

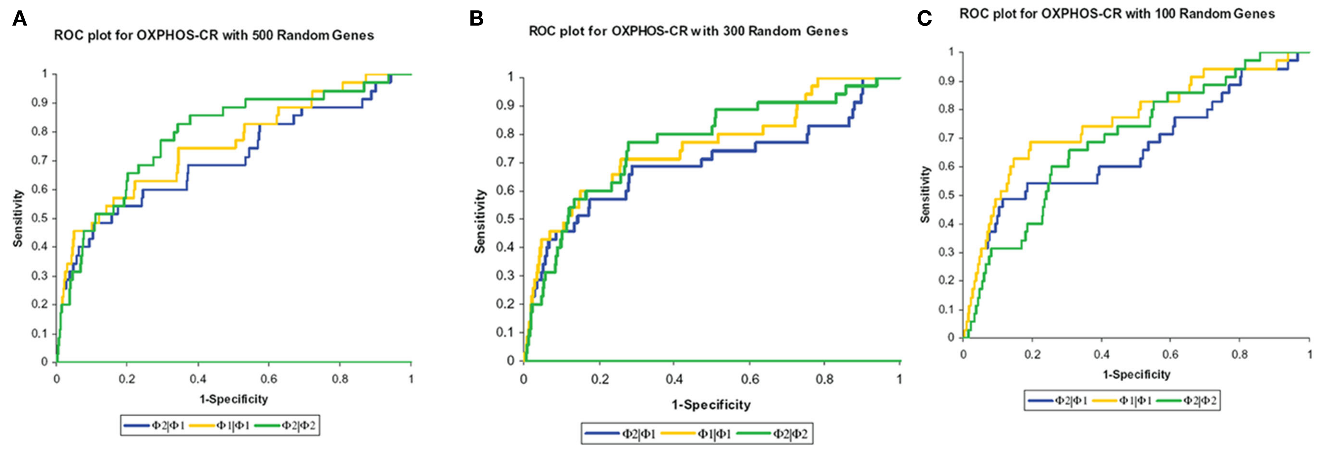


FIGURE 3 | Maintenance of rank. The methods were compared based on their ability to maintain the rank of the "golden standard" classifications. **(A)** ROC plot of the rankings of the classifications from the added 500 random genes dataset. **(B)** ROC plot of the rankings of the classifications from the added 300 random genes dataset. **(C)** ROC plot of the rankings of the classifications from the added 100 random genes dataset.

A conspicuously different value for *cytoskeleton*, between $\Phi_2 \mid \Phi_2$ and $\Phi_1 \mid \Phi_1$, can be observed in **Table 3**. While we have discussed above (see Introduction) the theoretical difference in how significance is computed under $\Phi_2 \mid \Phi_2$ and $\Phi_1 \mid \Phi_1$, it may serve as an informative exercise to look at some of the numbers that give rise to the difference in p-value for *cytoskeleton* under these two settings. Under the setting of Φ_1, 4 significant genes are mapped to *cytoskeleton* out of a total possible 25. Under the setting of Φ_2, 10 significant paths are mapped to *cytoskeleton* out of a possible 49. That is, while there are only 4 of 25 cytoskeleton genes in the interesting set, they are multifunctional – accounting for a total of 10 paths to mappings (an average 2.5 per gene), versus an average of slightly less than 2 paths to mappings for *cytoskeleton* genes over all. Thus, while these paths may describe different precise functions, they collectively implicate the overall *cytoskeleton* classification. Following up on the relevance of the *cytoskeleton* classification with respect to this experimental setting, it becomes clear that interesting genes mapped to it represent a biological reality worthy of distinguishing this category as significant. One of these genes is TPM1, which is an Actin binding protein and provides stability to Actin filaments. Accordingly, Actin is also present. Actin is a component of the cytoskeleton, which provides mechanical support to the cell. Actin also works with Myosin during muscle contraction in a calcium dependent interaction mechanism. Myosin is also one of the interesting genes. Myosin is a contractile protein whose function depends on the hydrolysis of ATP. Together, the presence of these genes reflects the physical demand placed on cytoskeleton with respect to the contraction and remodeling of the uterus, which occur during pregnancy (Slater et al., 2002; Wang and Hirsch, 2003; Salomonis et al., 2005).

Table 3 | Comparison of methods on the Girotti and Zingg (2003) data.

Gene ontology	$\phi_1 \mid \phi_1$	$\phi_2 \mid \phi_1$	$\phi_3 \mid \phi_3$	$\phi_2 \mid \phi_2$
Ribosome	0.093	0.078	0.956	5.91×10^{-5}
Protein modification	0.842	0.653	0.268	4.21×10^{-5}
Defense response	0.209	0.068	0.009	3.02×10^{-4}
Lipid transport	0.423	0.423	0.601	0.285
Lipid metabolism	0.233	0.106	0.616	1.88×10^{-5}
Intracellular protein transport	0.545	0.548	0.504	0.096
Hormone	0.308	0.113	0.248	0.552
Cell differentiation	0.281	0.281	0.423	0.791
Extracellular matrix	0.002	0.003	0.977	0.176
Cytoskeleton	0.259	0.125	0.432	0.017
Cell motility	0.022	0.024	0.128	0.003

Table 4 | Agreement of significance calls.

Method	$\Phi_1 \mid \Phi_1$	$\Phi_2 \mid \Phi_1$	$\Phi_3 \mid \Phi_3$	$\Phi_2 \mid \Phi_2$
$\Phi_1 \mid \Phi_1$	23	21	10	12
$\Phi_2 \mid \Phi_1$		36	14	21
$\Phi_3 \mid \Phi_3$			56	24
$\Phi_2 \mid \Phi_2$				107

Computing p-values for all 397 classifications by $\Phi_2 \mid \Phi_1$, $\Phi_2 \mid \Phi_2$, and $\Phi_3 \mid \Phi_3$ takes, on average, approximately 2 h. This is longer than for $\Phi_1 \mid \Phi_1$ and other statistics in the literature (Khatri and Draghici, 2005). However, it is still much more rapid than annotation by hand (Girotti took weeks to construct her annotation), and is readily justified by the time and expense that goes into collecting the data. Indeed, although microarray data analysis is typically "fast," other modern types of analysis (e.g., analysis of high-throughput sequencing data) typically takes hours or even days, depending on their complexity. Using repeated weights (W_i,j), which allowed us to treat genes with equal scores simultaneously, is an important factor in the speed of the algorithm. For determining the value for $a = \Phi_2(H', r)$, from table T, $|W| = 54$, T had 5581 entries, and required 8,172,874 operations. With 4137 genes and A_{max} as large as 22771, the naïve dynamic program would have required a table with approximately 500 million entries (22771∗22771), and would have required approximately 2 trillion updates (4137∗22771∗22771).

One may wonder whether exact computation of the p-values is necessary as described here. While it is intuitive that approximations will differ from exact p-values, it is not readily apparent that such approximations are inutile. For this reason, we used the hypergeometric equation to approximate $\Phi_2 \mid \Phi_2$ and found that values do indeed differ significantly. When compared, the two were found to be only weakly correlated ($R^2 = 0.2610$). This is a concern not only for ontological analysis, but for broader, integrative analyses as well. Increasingly, the experimental design of high-throughput experimentation has grown to include probabilistic frameworks in which knowing the exact probabilities or p-values is important. An example would be Bayesian modeling (Segal et al., 2001). In large Bayesian networks, multiple classifications may be used as factors for the inference. Small errors in p-values can propagate when being used as conditioning values. Thus, exact computation of p-values is important not just for determining significance, but also for the sake of larger integrative studies.

Our dynamic programming approaches, for $\Phi_2 \mid \Phi_2$ in particular, can easily be generalized to compute p-values for other scoring functions. For example, one might define the score of a gene X with respect to a classification Y by the number of classifications reachable from Y to which X is mapped. Of course, interesting genes could also be selected based on features that have no direct connection to expression or ontology structure, such as presence of an amino-acid motif or of a common upstream motif. Our methods are also applicable to other annotation resources available to the bioinformatics community. For example, databases such as Kyoto Encyclopedia of Genes and Genomes (KEGG; Kanehisa and Goto, 2000) or Munich Information Center for Protein Sequences (MIPS; Mewes et al., 2000), have organized annotations to which statistics can be applied in a similar manner as described here.

While the methods presented here compute individual p-values exactly, one must consider the use of such statistics in context. Naïve application of these methods to a large number of candidate classifications creates a multiple hypothesis testing problem and risks the identification of related classifications. These problems can be addressed by adjusting the p-value threshold and by sampling-based methods (Zhong et al., 2004).

ACKNOWLEDGMENTS

The authors wish to thank Marie-Eve Carrier, Dominic Devost, Karen Hellauer, and Elisa Monaco for their advice and assistance. Funding: This material is based upon work supported by grants from NSERC and CIHR. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the funding agencies.

REFERENCES

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.

Berriz, G. F., King, O. D., Bryant, B., Sander, C., and Roth, F. P. (2003). Characterizing gene sets with FuncAssociate. *Bioinformatics* 19, 2502–2504.

Castillo-Davis, C. I., and Hartl, D. L. (2003). GeneMerge – postgenomic analysis, data mining, and hypothesis testing. *Bioinformatics* 19, 891–892.

Diaz-Uriarte, R., and Dopazo, J. (2003). "Use of GO terms to understand the biological significance of microarray differential gene expression data," in *Methods of Microarray Data Analysis III, Papers from Camda '02*, Chap. 15, eds K. F. Johnson and S. M. Lin (Durham: Kluwer Inc.), 233–247.

Doniger, S. W., Salomonis, N., Dahlquist, K. D., Vranizan, K., Lawlor, S. C., and Conklin, B. R. (2003). MAPPFinder: using geneontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biol.* 4, R7.

Draghici, S., Khatri, P., Bhavsar, P., Shah, A., Krawetz, S. A., and Tainsky, M. A. (2003). Onto-tools, the toolkit of the modern biologist: onto-express, onto-compare, onto-design and onto-translate. *Nucleic Acids Res.* 31, 3775–3781.

Girotti, M., and Zingg, H. H. (2003). Gene expression profiling of rat uterus at different stages of parturition. *Endocrinology* 144, 2254–2265.

Hosack, D. A., Dennis, G. Jr., Sherman, B. T., Lane, H. C., and Lempicki, R. A. (2003). Identifying biological themes within lists of genes with EASE. *Genome Biol.* 4, R70.

Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30.

Khatri, P., and Draghici, S. (2005). Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics* 21, 3587–3595.

King, O. D., Foulger, R. E., Dwight, S. S., White, J. V., and Roth, F. P. (2003). Predicting gene function from patterns of annotation. *Genome Res.* 13, 896–904.

Man, M. Z., Wang, Z., and Wang, Y. (2000). Power sage: comparing statistical tests for sage experiments. *Bioinformatics* 16, 953–959.

Mewes, H. W., Frishman, D., Gruber, C., Geier, B., Haase, D., Kaps, A., Lemcke, K., Mannhaupt, G., Pfeiffer, F., Schüller, C., Stocker, S., and Weil, B. (2000). MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* 28, 37–40.

Salomonis, N., Cotte, N., Zambon, A. C., Pollard, K. S., Vranizan, K., Doniger, S. W., Dolganov, G., and Conklin, B. R. (2005). Identifying genetic networks underlying myometrial transition to labor. *Genome Biol.* 6, R12.

Segal, E., Taskar, B., Gasch, A., Friedman, N., and Koller, D. (2001). Rich probabilistic models for gene expression. *Bioinformatics* 17(Suppl. 1), S243–S252.

Slater, M. D., Zervou, S., and Thornton, S. (2002). Prostaglandins and prostanoid receptors in human pregnancy and parturition. *J. Soc. Gynecol. Investig.* 9, 118–124.

Subramanian, A., Tamayo, P., Mootha, V. K, Mukherjee, S., Elbert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* 102, 15545–15550.

The Gene Ontology Consortium. (2000). Gene ontology: tool for the unification of biology. *Nat. Genet.* 25, 25–29.

Wang, H., and Hirsch, E. (2003). Bacterially-induced preterm labor and regulation of prostaglandin-metabolizing enzyme expression in mice: the role of toll-like receptor 4. *Biol. Reprod.* 69, 1957–1963.

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, R28.

Zhang, B., Schmoyer, D., Kirov, S., and Snoddy, J. (2004). GOTree machine (GOTM): a web-based platform for interpreting sets of interesting genes using gene ontology hierarchies. *BMC Bioinformatics* 5, 16. doi:10.1186/1471-2105-5-16

Zhong, S., Tian, L., Li, C., Storch, K.-F., and Wong, W. H. (2004). Comparative analysis of gene sets in the gene ontology space under the multiple hypothesis testing framework. *Proc. IEEE Comput. Syst. Bioinform. Conf.* 2004, 425–35.

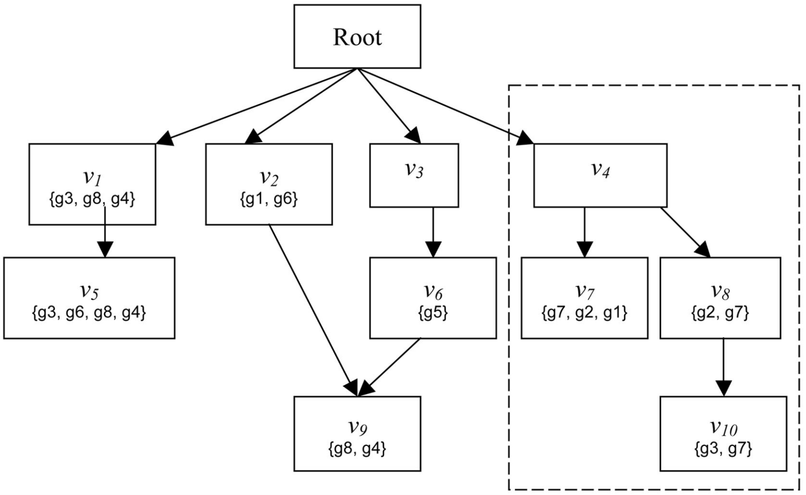
Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 04 November 2011; paper pending published: 19 November 2011; accepted: 06 February 2012; published online: 23 February 2012.

Citation: Iacucci E, Zingg HH and Perkins TJ (2012) Methods for determining the statistical significance of enrichment or depletion of gene ontology classifications under weighted membership. Front. Gene. ***3****:24. doi: 10.3389/fgene.2012.00024*

This article was submitted to Frontiers in Systems Biology, a specialty of Frontiers in Genetics.

APPENDIX



Gene	1	2	3	4	5	6	7	8	Score
Membership score (Φ_1) at root	1	1	1	1	1	1	1	1	8
Membership score (Φ_1) at v_4	1	1	1	0	0	0	1	0	4
Pathway score (Φ_2) at root	2	2	3	4	1	2	3	4	21
Pathway score (Φ_2) at v_4	1	2	1	0	0	0	3	0	7

Ex: Given the example above and a list of interesting genes $\boldsymbol{H=\{g_1, g_4, g_5\}}$
We wish to calculate the probability of observing the score of v_4

Under $(\Phi_1|\Phi_1)$*?*

$\Phi_1(H,v_4)=1 \,/\, \Phi_1(H,r)=3$

$$\Pr(1\,|\,3) = \frac{\binom{4}{1}\binom{4}{2}}{\binom{8}{3}}$$

Under $(\Phi_2|\Phi_1)$*?*

$(\Phi_2(H,v_4)=1,\ \Phi_2(H,\overline{v_4})=5) \,/\, \Phi_1(H,r)=3$

$\forall$ *H' s. t.* $\Phi_2(H',v_4)=1$ & $\Phi_2(H',\overline{v_4})=5$

$\exists$ 2 sets: $\{g_1, g_4, g_5\}$, $\{g_1, g_5, g_6\}$

$$\Pr(1\,|\,3) = \frac{2}{\binom{8}{3}}$$

FIGURE A1 | Explicit examples of weighted membership functions and *p*-value computations.

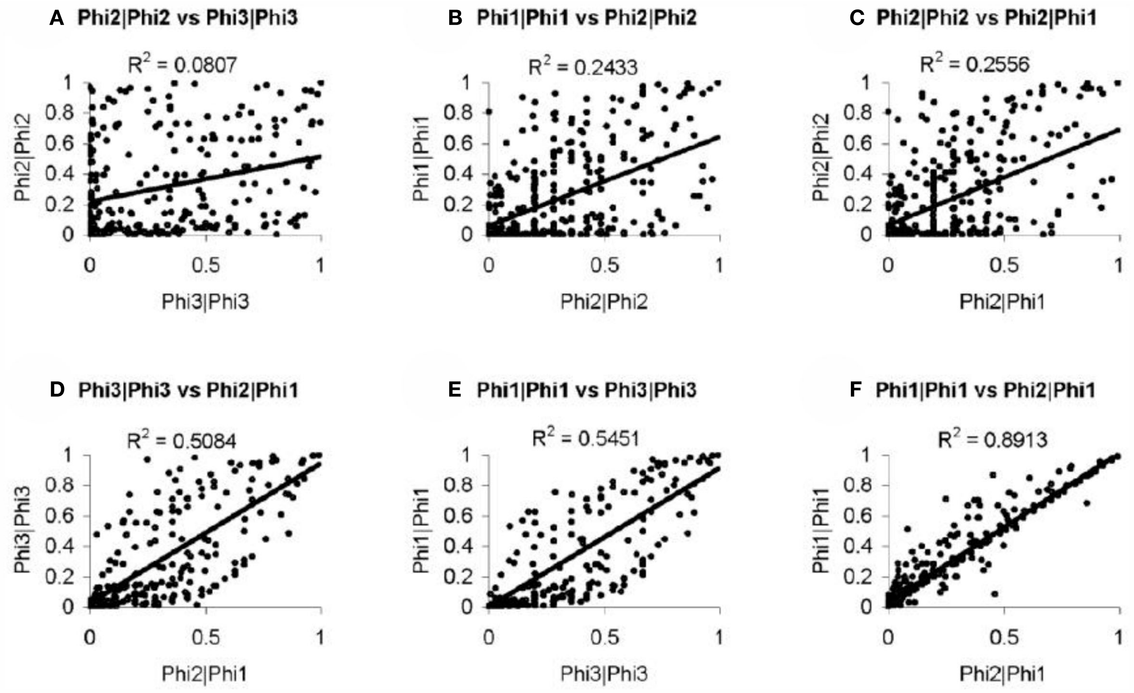


FIGURE A2 | Method comparisons (A–F) pair-wise *p*-value comparisons between methods, these scatter pots show the degree of agreement between methods. R^2 values are provided as a summary of the agreement.