Peng et al. BMC Bioinformatics 2014, 15(Suppl 2):S5
http://www.biomedcentral.com/1471-2105/15/S2/S5

BMC Bioinformatics

PROCEEDINGS Open Access

Towards integrative gene functional similarity measurement

Jiajie Peng[1,2], Yadong Wang[1*], Jin Chen[2,3*]

From The Twelfth Asia Pacific Bioinformatics Conference (APBC 2014)
Shanghai, China. 17-19 January 2014

Abstract

Background: In Gene Ontology, the "Molecular Function" (MF) categorization is a widely used knowledge framework for gene function comparison and prediction. Its structure and annotation provide a convenient way to compare gene functional similarities at the molecular level. The existing gene similarity measures, however, solely rely on one or few aspects of MF without utilizing all the rich information available including structure, annotation, common terms, lowest common parents.

Results: We introduce a rank-based gene semantic similarity measure called InteGO by synergistically integrating the state-of-the-art gene-to-gene similarity measures. By integrating three GO based seed measures, InteGO significantly improves the performance by about two-fold in all the three species studied (yeast, *Arabidopsis* and human).

Conclusions: InteGO is a systematic and novel method to study gene functional associations. The software and description are available at http://www.msu.edu/~jinchen/InteGO.

Background

The Gene Ontology (GO) provides a structured, controlled vocabulary of terms, which are interrelated forming a directed acyclic graph (DAG) for describing and categorizing (into three categories) the attributes for genes, gene products and sequences [1]. The "molecular function" (MF) category describes fundamental biochemical activities (including specific binding to ligands or structures of a gene product) at the molecular level [2]. As a popular resource used for functional annotation, MF provides rich information and a convenient way to study gene functional similarity by comparing terms with which the genes are annotated [3-7], which subsequently supports a wide variety of applications, such as assessing target gene functions [8], predicting gene functional associations [9], inferring protein nomenclature [10], predicting sub-cellular localization [11], discovering new pathways [12], *etc.*

In order to compute gene-to-gene functional similarities using GO, various computational approaches have been developed. These approaches can be classified into two distinct categories: 1) group-wise, meaning calculating gene-to-gene similarity directly based on a statistical framework considering all the terms annotated to the target genes [13-15], and 2) pair-wise, *i.e.*, indirectly computing gene-to-gene similarity using term-to-term similarities computed with GO semantic measures [12,16-21]. Each of the aforementioned measurements adopts one or a few kinds of knowledge in the GO efficiently. However, they do not rely on all of the rich information available in the GO databases. In this paper, we propose a new rank-based gene semantic similarity measure called InteGO (Integrated Gene Ontology measure), which can integrate the state-of-the-art gene-to-gene measures [12,13,17] (therefore considering more information than these measures) to bring the performance of the GO-based functional similarity studies to a higher level.

In the first GO-based measure category (group-wise), by combining elements of the topology and annotation information, the Yu measure calculates a probabilistic

* Correspondence: ydwang@hit.edu.cn; jinchen@msu.edu
[1]School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China
[2]MSU-DOE Plant Research Laboratory, Michigan State University, East Lansing, MI 48824, USA
Full list of author information is available at the end of the article

level of similarity from GO, in order to directly compute gene similarity [13]. The main idea of the Yu measure is that a pair of genes should be very similar if they are included in a functional group with a few proteins, whereas the similarity is lower if the gene pair belongs to a large gene group. Mathematically, given two gene $g1$ and $g2$, the gene-to-gene similarity can be calculated with:

$$GeneSim_{Yu}(g_1, g_2) = -\ln \frac{n_{g1,g2}}{N} \quad (1)$$

where $n_{g1,g2}$ is the total number of gene pairs that have the same set of lowest common ancestors (LCAs) as g_1 and g_2; N is the total number of gene pairs in the selected GO category. A LCA is the common ancestor with the highest information content (IC). In the illustrative example in Figure 1, there are in total 45 gene pairs possible among the ten genes; the LCA of gene pair g_1 and g_2 is t_1, and the number of gene pairs (which LCA is also t_1) is 9. Therefore, the similarity of g_1 and g_2 based on the Yu measure is $-ln(9/45) = 1.61$. The Yu measure considers both the elements of topological distance and the LCA distance. However, it simplifies the computation of shared information of both genes without using all of the common parents of the GO terms annotated to g_1 or g_2, which neglects the locations of LCAs and the aggregate semantic contributions from the parents of the target terms (due to the high complexity of graph matching). Alternatively, the SORA [15] measure computes the IC of a term set by means of combining inherited and extended information content of the terms based on the structure of GO. Gene functional similarity is estimated using the IC overlap ratio of term sets. However, like the Yu measure, it ignores valuable information implicit in the semantics, *i.e.*, the common parents of the GO terms, when calculating the shared IC and relationships among involved terms.

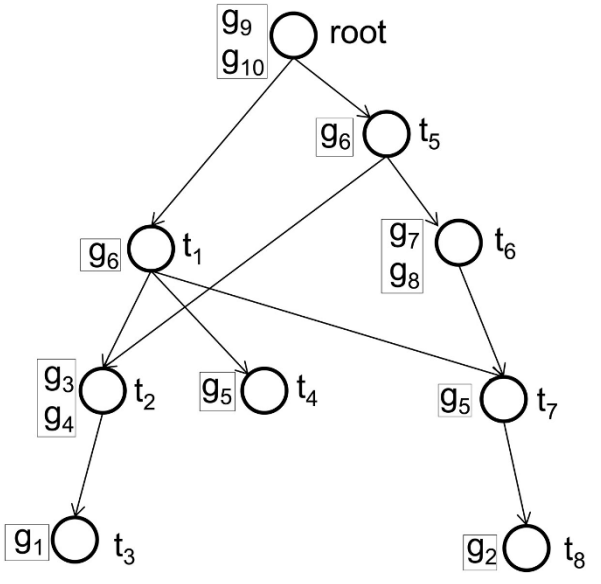

Figure 1 An illustrative example of Gene Ontology (GO). An illustrative example of GO forming a directed acyclic graph (DAG), in which nodes and edges represent GO terms and "is-a" or "part-of" relationships between terms. $\{t_1, ..., t_7, root\}$ is the set of GO terms, and $\{g_1, ..., g_{10}\}$ is the set of genes annotated to these terms.

In the measures in the second category (pair-wise), the pair-wise term comparisons originally developed for natural language processing [16,18-21] are utilized, and are strongly dependent on the specific taxonomy. Among the earlier developed methods, an IC based measure called the Resnik measure has showed strong correlations between its results and gene expression similarities on yeast [16,22]. Mathematically, given a GO term t, its IC is defined as a negative log likelihood $IC(t) = -\log(|G_t|/|G_{root}|)$, where G_t and G_{root} are the sets of genes annotated to term t and the root term (including all of its descendants) respectively. In the Resnik measure, the similarity between term t_1 and t_2 is defined as the IC of LCA: $TermSim_{Resnik}(t_1, t_2) = IC(LCA_{12})$. Although the Resnik measure strongly correlated with the gene expression data [22], terms sharing the same LCA have the same semantic similarity, even if they are at very different levels of GO. Consequently, it cannot differentiate term pairs that are far from LCA with term pairs close to the same LCA. In the illustrative example in Figure 1, the common parent of t_2 and t_7 is t_1, which is the same as the LCA of t_3 and t_8. According to the Resnik measure, $Sim_{Resnik}(t_2, t_7) = Sim_{Resnik}(t_3, t_8) = 0.51$, but clearly the distance from t_2 and t_7 to the LCA is shorter. To take both the distance from LCA to the target terms and the distance from LCA to root into account [17], a later-developed measure called the Schlicker measure was proposed:

$$TermSim_{Schliker}(t_1, t_2) = \frac{2 \times IC(LCA_{12})}{IC(t_1) + IC(t_2)} \times \left(1 - \frac{|G_{LCA_{12}}|}{|G_{root}|}\right) \quad (2)$$

where $G_{LCA_{12}}$ is the set of genes annotated to the LCA of t_1 and t_2.

In Eq 2, the first part on the right side of the equation quantifies the distance from terms t_1 and t_2 to their LCA, and the second part measures the distance from the LCA to the root, where a short former distance and a long later distance indicate a higher similarity. Experimental results revealed that the Schlicker measure agrees with sequence similarity [17]. In the same example in Figure 1, the Schlicker measure is able to differentiate term pair (t_2, t_7) and (t_3, t_8) with $TermSim_{Schlicker}(t_2, t_7) = 0.15$ and $TermSim_{Schlicker}(t_3, t_8) = 0.09$. However, the common problem of the Schlicker measure

and the Resnik measure is that they only consider a single common ancestor, neglecting the fact that two GO terms may have multiple common ancestors in the GO structure [23].

Recently, the Wang measure was proposed to consider all of the parent terms of the target terms [12]. Given a term t_1 and its parent term p, the semantic contribution of p to t_1, denoted as $S_{t1,p}$, is defined as the maximal semantic contribution of the paths from t_1 to p. The GO term similarity in the Wang measure is defined in Eq 3, where P_1 (or P_2) are the sets of all of the parents of t_1 (or t_2).

$$TermSim_{Wang}(t_1, t_2) = \frac{\sum_{p \in P_1 \cap P_2} (S_{t_1,p} + S_{t_2,p})}{\sum_{t \in P_1} S_{t_1,p} + \sum_{t \in P_2} S_{t_2,p}} \quad (3)$$

The experiment result shows that this measure performs significantly better than Resnik measure on yeast protein functional similarities [12]. However, the Wang measure ignores both the topological distances among the LCAs and the statistics of gene annotations that the Yu measure has taken into consideration. For the same example in Figure 1, to compare the similarity of term t_3 and t_8, all of the common parents of the two terms, $P_3 = \{t_1, t_2, t_3, t_4, t_5, root\}$ and $P_8 = \{t_1, t_5, t_6, t_7, t_8, root\}$, are considered by the Wang measure.

For the Resnik, Schlicker and Wang measures, gene-to-gene similarity is computed based on the GO term similarities that annotate to the target genes. In Wang *et al* [12], let g_1 and g_2 be two genes and T_1 and T_2 be the sets of GO terms annotated to g_1 and g_2, the gene-to-gene similarity is calculated by Eq 4:

$$GeneSim(g_1, g_2) = \frac{\sum_{t \in T_1} TermSim(t, T_2) + \sum_{t \in T_2} TermSim(t, T_1)}{|T_1| + |T_2|} \quad (4)$$

where t is a GO term, $TermSim(t, T_x) = \max_{ti \in Tx} \mathrm{Sim}(t, t_i)$, which represents the highest similarity between t and term set T_x. Note that, for both $|T_1|$ and $|T_2|$, only the terms with $TermSim(t, T_x) \neq 0$ are counted.

To the best of our knowledge, the existing measures emphasize on only one or few types of relationships between genes but ignores the others. One of these measures may be better than the others on one specific set of terms and genes, but may perform worse than the other measures on another gene set. Since none of the existing measures takes into account all of the aspects of GO (structure, annotation, LCA, all of the common parent, etc), which is of course a challenging task, it is hypothesized that the integration of multiple measures can improve the performance, since integration of multiple methods has been widely applied for performance boosting [24-26]. In this paper, we proposed a rank-based gene semantic similarity measure called InteGO by synergistically integrating the state-of-the-art gene-to-gene similarity measures. The integrated measures are called seed measures in the rest of paper. The major contributions of our work are:

- While the existing measures only consider one or few aspects of the problem, InteGO is an integrative approach, which conceptually considers all of the information in GO to reduce incorrect score assignments. In addition, InteGO employs an adaptive approach for the optimization of the seed measure integration.
- A rank-based approach is used to integrate multiple seed measures. Since the values from different seed measures have different scales and distributions, a direct integration of the values may lead to biased results. With our rank-based approach, InteGO unifies the scale and distribution among different seed measures, ensuring fair comparison.
- InteGO is an open framework, which adds the flexibility to integrate more GO similarity measures, more advanced evaluation and integration methods in the future.

InteGO was systematically tested on three species with different levels of complexity of GO annotations, i.e., yeast, *Arabidopsis* and human. The experimental results on all of the three species show that InteGO performs consistently better than the other measures in all of the tests.

Method

In order to integrate multiple seed measures in InteGO, two key problems need to be solved: first, how to select the most appropriate seed measures for integration; second, how to integrate all of the scores from the different seed measures. To solve these problems, InteGO is divided into two steps: 1) to compute similarity scores with every seed measure individually and rank the scores, and 2) to evaluate and integrate the ranks of multiple seed measures.

Rank-based similarity

The outputs of the different gene-to-gene similarity measures have different scales and distributions. Therefore, a direct integration of the values may lead to biased results. In InteGO, we unify the scale and distribution among different seed measures with a rank-based approach. One common problem of rankbased approaches though is the data size dependence, *i.e.*, while a rank-based approach can work well on a relative large dataset, it is often inadequate on a small set of data. For example, if only two

genes are provided by a user, the similarity rank of the two genes is always one, regardless how high or how low the actual similarity score is. Therefore, instead of requiring users to always provide a large set of genes to compare (which is not reasonable all of the time), InteGO maintains a background set of genes (*BG*) for every species of interest to unify the similarity scores from the multiple seed measures. *BG* must satisfy two requirements: 1) it is large enough; 2) it unbiasedly includes the full spectrum of gene similarity scores, ranging from the lowest to the highest.

The framework of InteGO is shown in Figure 2. In the steps with grey background, the similarity scores in *BG* are pre-calculated with all of the seed measures and saved in a database called *GeneSimDB*. When a user inputs a gene set *G*, the similarity scores of all of the gene pairs in *G* and all of the gene pairs between *G* and *BG* will be calculated with all of the seed measures, and be merged into *GeneSimDB*. If *G* is a subset of *BG*, InteGO will output the results directly. Finally, all of the gene pairs in *GeneSimDB* are sorted incrementally based on their gene similarity scores and are ranked. The ranked gene similarity score $RankSim(g_1, g_2, m)$ for genes g_1 and g_2 in *G* is calculated as:

$$RankSim(g_1, g_2, m) = \frac{2 \times r^m_{g_1,g_2}}{(|BG \cup G|)^2} \quad (5)$$

where $r^m_{g_1,g_2}$ is the rank of gene pair g_1 and g_2 using seed measure *m*, and *BG* is the predefined background gene set, and *G* is the user provided gene set. The ranked similarity indicates how similar a given gene pair is in the background of all of the gene pairs.

One advantage to use the rank-based measure is to unify different scales and distributions among the seed measures. Therefore, the agreement among the ranks could indicate the functional similarities appropriately. An illustrative example is shown in Table 1. Given ten gene pairs, three measures (M_A, M_B and M_C) are used to calculate the gene-to-gene semantic similarities based on the GO. The first column of the values show that the similarity scores of measure M_A, M_B and M_C have different scales and different distributions. For example, the semantic similarity of gene pair 3 is 3.0 for measure M_A and 0.9 for measure M_B, although they both mean the highest functional similarity in their own datasets. The second column of the values show the ranks of the gene pairs under each seed measure in assenting order.

Adaptive integration approach

The rank-based semantic similarities of gene pairs from every seed measure provide an unique opportunity to compute the gene-to-gene similarities with all the information of GO utilized by the seed measures. A key problem here is how to select the most appropriate integration approach. Although there are many integration approaches all working well on certain domains, there does not exist one method that is always better than the others. In fact, to choose an appropriate integration method is largely dependent on the content of the study. Therefore, we propose an adaptive approach to automatically select the most appropriate integration method from a set of candidates. The main idea of the adaptive approach is to score all of the methods in the pool of the

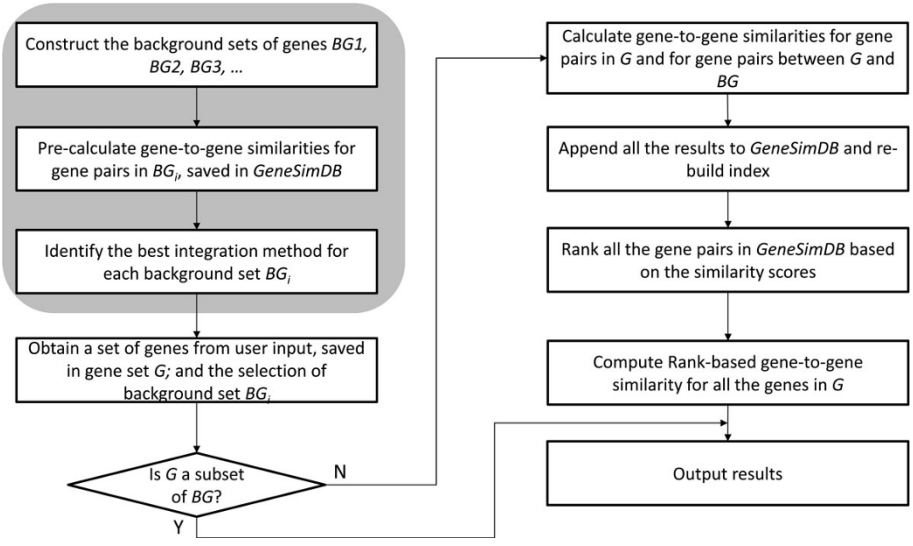

Figure 2 Framework of InteGO for calculating the rank-based gene-to-gene similarities in MF. Framework of InteGO for calculating the rank-based gene-to-gene similarities in MF. The boxes in the grey block are the pre-processing modules for the preparation of the background gene set.

Table 1 Illustrative example for integration similarity.

Gene Pairs	semantic Similarity			Rank of Similarity			Integration of Ranks			
	M_A	M_B	M_C	M_A	M_B	M_C	MAX	MIN	MEAN	MEDIAN
Gene pair 1	2.4	0.2	0.04	9	2	4	0.9	0.2	0.5	0.4
Gene pair 2	1.8	0.6	0.12	6	7	8	0.8	0.6	0.7	0.7
Gene pair 3	3.0	0.9	0.03	10	10	3	1.0	0.3	0.8	1.0
Gene pair 4	1.2	0.3	0.05	5	3	5	0.5	0.3	0.4	0.5
Gene pair 5	0.9	0.1	0.06	3	1	6	0.6	0.1	0.3	0.3
Gene pair 6	0.5	0.5	0.02	2	6	2	0.6	0.2	0.3	0.2
Gene pair 7	1.0	0.4	0.09	4	4	7	0.7	0.4	0.5	0.4
Gene pair 8	1.8	0.4	0.13	6	4	9	0.9	0.4	0.6	0.6
Gene pair 9	0.2	0.7	0.01	1	8	1	0.8	0.1	0.3	0.1
Gene pair 10	2.1	0.8	0.16	8	9	10	1.0	0.8	0.9	0.9

Illustrative example for integration similarity, where M_A, M_B and M_C are three seed gene-to-gene functional similarity measures.

candidate integration approaches with the background set *BG*, and then select the best one.

InteGO provides four integration methods: max, min, mean and median. As an open system, InteGO also allows users to use their own integration methods. Mathematically, let $RankSim(g1, g2, m)$ be rank-based similarity of gene $g1$ and $g2$ using seed measure m, InteGO is defined as:

$$InteGO(g_1, g_2, I) = \begin{cases} max_{m\in M} RankSim(g_1, g_2, m) & if\ I = max \\ min_{m\in M} RankSim(g_1, g_2, m) & if\ I = min \\ mean_{m\in M} RankSim(g_1, g_2, m) & if\ I = mean \\ median_{m\in M} RankSim(g_1, g_2, m) & if\ I = median \\ integration_{m\in M} RankSim(g_1, g_2, m) & if\ I = other\_integration \end{cases} \quad (6)$$

where M is a set of seed measures and I is an integration method which is max, min, mean, median of all of the ranks, or any other integration method that is defined by the user. For an illustrative example in Table 1, the results based on the four different integration methods are shown in the third column.

To automatically determine which integration method is the best, all of the gene pair similarities in *BG* are calculated based on each candidate integration method and are evaluated systematically with biological data. Recent studies used the correlation coefficient of gene expression correlations or gene sequence similarities to evaluate the MF based gene similarities [22]. However, it is not always correlated between gene functional similarities and gene expression correlation or sequence similarities [12]. Furthermore, previous studies show that enzymes are usually categorized biochemically with EC (Enzyme Commission) numbers but not their nucleotide or amino acid sequences [27,28], which indicates that it could be a better way for using EC numbers to explain molecular function with the criteria that the molecular functions of a group of genes are similar if they have the same EC numbers [12,29,30].

To systematically use EC to choose an integration method, all of the genes in *BG* are grouped based on their EC numbers (four digits), and then the differences between the inter- and intra-EC gene-to-gene similarities are tested. With an integration method, the higher the ratio between intra-EC gene similarities and inter-EC gene similarities, the better the integration method is. Quantitatively, we utilize the logged fold change (LogFC) measure which has been widely used in the gene expression studies [31]. The LogFC score of EC *ei* is defined in Eq 7:

$$LogFC(e_i) = \frac{1}{|E|} \times \sum_{e_j\in E; G(e_j)\cap G(e_i)=\theta} \frac{\sum_{g\in G(e_i)} diff_g(e_i, e_j)}{|G(e_i)|} \quad (7)$$

where $G(e_i)$ is set of all of genes which EC number is e_i; E is a set of ECs which do not have overlapped genes with e_i ($G(e_j) \cap G(e_i) = \varnothing$); $diff_g(e_i, e_j)$ is computed as:

$$diff_g(e_i, e_j) = \ln \frac{|G(e_i)| \times \sum_{g'\in G(e_j)} (1 - GeneSim(g, g') + c)}{|G(e_j)| \times \sum_{g*\in G(e_i)} (1 - GeneSim(g, g*) + c)} \quad (8)$$

where c is a Laplacian smoothing parameter which is a constant small positive number; $G(e_i)$ is the set of all of the genes assigned to EC e_i except gene g; $G(e_j)$ is the set of all of the genes assigned to EC e_j; g is a gene assigned to e_i. In Eq 8, the numerator represents the inter-EC distance and the denominator represents the intra-EC distance. The higher the $diff_g(e_i, e_j)$ is, the more obvious the positive difference between inter-EC difference and intra-EC difference is.

For example, given nine genes in *BG*, four of which have the same EC number, labeled as e_1, and the other five genes belong to another EC number, labeled as e_2. To calculate the LogFC score for e_1, we first compute $diff_g(e_1, e_2)$ with Eq 8, meaning that every gene in e_1 is compared with every other gene in e_1 for the average intra-EC difference, and then every gene in $e1$ is compared with every gene in e_2 to get the inter-EC differences. $logFC(e_1)$ is the average of all of the $diff_g(e_1, e_2)$ scores for the genes assigned to e_1.

The method that has the highest LogFC scores for all of the ECs are considered as the most appropriate integration method for *BG*. If a user input set *G* is much smaller than *BG* (which often happens), we assume the selected method is also the most suitable for $G \cup BG$. If the size of *G* is comparable to *BG*, it is not necessary to use *BG*, then the integration method shall be selected based on the evaluation on *G*.

Results

To systematically evaluate the performance of InteGO, we tested it on three model organisms with different levels of GO annotation scale and complexity. For each of them, we adopted EC numbers and protein sequences as independent biological evidences.

Data preparation

The GO annotation and structure data were downloaded from the GO website (http://www.geneontology.org/GO.downloads.shtml). To systematically evaluate different GO-based gene-to-gene similarity measures on MF, the pathway and EC number information of Yeast, *Arabidopsis* were downloaded from the Saccharomyces genome database (http://biocyc.org/YEAST/organism-summary?object=YEAST), PlantCyc (http://ftp.plantcyc.org/Pathways) and HumanCyc (http://humancyc.org) respectively. Note that our EC based evaluation method requires that an EC has at least two genes. In yeast, *Arabidopsis* and human, 95, 325 and 312 ECs satisfy the criteria. The protein sequences were downloaded from the Saccharomyces genome database (http://www.yeastgenome.org/download-data/sequence), TAIR (http://www.arabidopsis.org/tools/bulk/sequences/index.jsp) and UniProt (http://www.uniprot.org) respectively.

Let *E* be the set of all of the ECs that have at least one gene assignment, we define *BG* as the set of all of the genes that has at least one EC assignments in *E*. This definition of *BG* ensures that for any gene in *BG* the intra-EC similarity is valid. The sizes of *BG* are 218, 1,348 and 1,504 for yeast, *Arabidopsis* and human respectively. An experiment on the variation of the background set (see Additional file 1) reveals that the use of a relatively smaller background set may affect performance significantly. Additional file 2, 3 and 4 show that the distribution of the gene-to-gene similarities with Yu, Schlicker and Wang measures, where the similarity scores are spread in the full spectrum of the range. In summary, the background gene sets are well prepared.

InteGO was implemented with Java JDK 1.6 and JUNG library [32]. The experiments were run on a windows 7 computer with Intel i7 CPU and 10 GB RAM.

Selecting seed measures

In order to select the most appropriate seed measures for InteGO, we screened four existing measures (Yu, Resnik, Schlicker and Wang) using the EC based evaluation method. Figure 3 shows that for the Yu, Schlicker and Wang measures, it is not distinguishable that one measure is significant better than another. The Yu, Schlicker and Wang measures all performed the best on yeast with the highest median value. The Schlicker measure performs best on *Arabidopsis*, while the Yu measure is best on human. Therefore, we chose all of the three as the seed measures in InteGO. We did not choose the Resnik measure, because it is clearly not as good as the other measures in all of the three species. Note that the upper-bound and the lower-bound of the LogFC scores in Figure 3 were set to 5 and -0.05 respectively to eliminate outliers.

In addition, Figure 4 shows that although all of the three seed measures perform equally well in some ECs, each measure has its own favorable EC groups. For example, the Schlicker and Wang measures perform the best in 51 and 52 out of the total 325 *Arabidopsis* ECs respectively (see Figure 4(b)), which is greater than the Yu measure (20 ECs). However, the Yu measure performs the best in 159 out of the total 315 human ECs, which dominant the EC group distribution in human (see Figure 4(c)). Therefore, an appropriate integration of these measures may combine the advantages of different measures and improve the overall performance. Note that although only four measures were screened in the experiment, more measures can be evaluated and added later since InteGO has an open framework.

Selecting integration method

In order to select the most appropriate integration method, four different approaches (MAX, MIN, MEAN and MEDIAN) were tested and compared. Figure 5 shows that MAX performs the best among the four integration methods. In yeast, although almost all of the measures have the same median value, the 25th percentile of MAX is 5, significantly higher than the Yu, Schlicker and Wang measure (1.68, 3.00 and 2.04 respectively) and the other integration methods. In *Arabidopsis* and human, the median of MAX are both 5, which is also significantly higher than that of all of the other integration methods. It indicates that the performance of MAX, a simple integration approach, has been increased to around 2-fold. This is because the integration considers all of the aspects of GO, while an individual seed measure, although nicely designed, is compromised in that it focuses on only one of few kinds of knowledge in GO. The other integration measures, especially MIN, however, cannot distinguishably improve the gene similarity performance. As shown in Figure 5(c), the result of MIN is even worse than the seed measures. It indicates that the performance of gene-to-gene similarity could be significantly improved only by the appropriate integration.

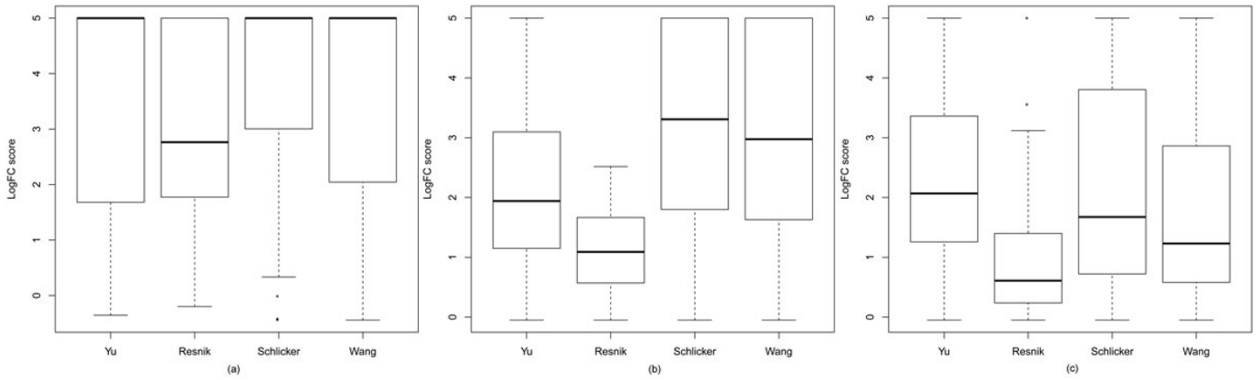

Figure 3 Logged fold change (LogFC) score comparison. Logged fold change (LogFC) score comparison for four different similarity measures in Molecular Function (MF) category on yeast (a), *Arabidopsis*(b) and human (c).

As mentioned in the previous section, the seed measures have their own favorable EC groups. To test whether MAX take advantage of all of the strength of the seed measures, we compared MAX with the Yu, Schlicker and Wang measure on all of the ECs. Figure 6 (a), (b) and (c) show that MAX dominant the EC groups, clearly different to the results in Figure 4. In detail, MAX performs the best in 140 and 172 out of 325 and 315 ECs in *Arabidopsis* and human respectively, while the numbers are only 2, 9, 6 in *Arabidopsis* and 2, 2, 1 in human for the Yu, Wang and Schlicker measures respectively. In summary, the experiment indicates that integrating multiple measures could improve the performance of gene similarity measurement and MAX is the best integration method.

Statistics analysis was carried out to test whether the results of the best integration measure (MAX) of InteGO is statistically the best. We compared InteGO with the three seed measures using TukeyHSD test [33]. The p-values shown in Table 2 and the 95% family-wise confidence level (Additional file 5, 6 and 7) indicate that the results of MAX are significant better than the results of all of the seed measures in yeast, *Arabidopsis* and human, with the only exception that the Schlicker measure's results are comparable in yeast, in that the Schlicker measure performs very well in yeast, so there is little room for InteGO to improve.

Protein sequence based performance evaluation

In addition to use EC as the evaluation criteria, protein sequence similarities were employed as independent evidence for further performance study. Although the correlation coefficient between semantic similarity and sequence similarity is not as strong as EC, it is generally accepted that as sequence similarity increases, so does the chance that these proteins are homologues, in which

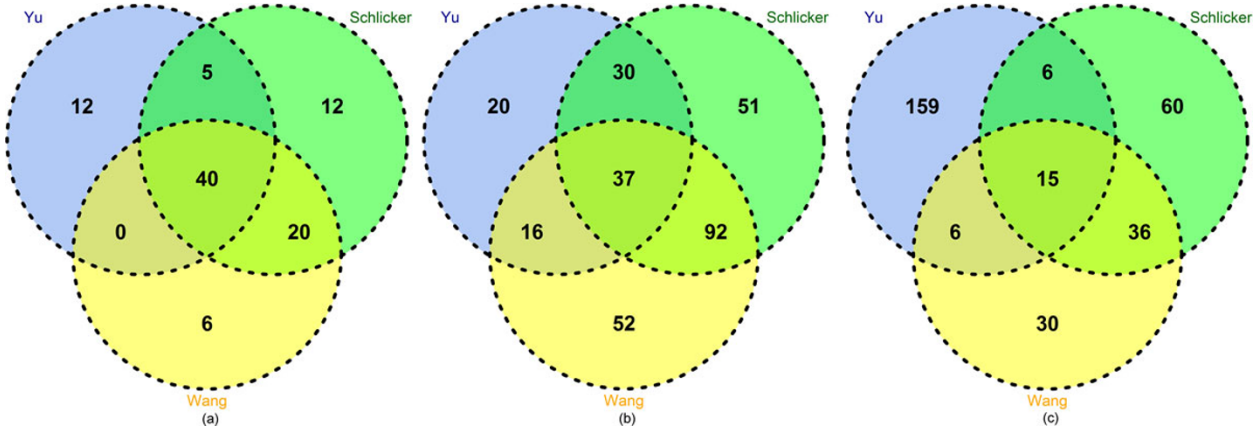

Figure 4 Venn Diagram for Yu measure, Schlicker measure and Wang measure with number of ECs on which perform best on yeast (a), Arabidopsis (b) and human (c). Venn Diagram for Yu measure, Schlicker measure and Wang measure with number of ECs on which perform best on yeast (a), *Arabidopsis* (b) and human (c). Blue, green and yellow represent Yu measure, Schlicker measure and Wang measure respectively.

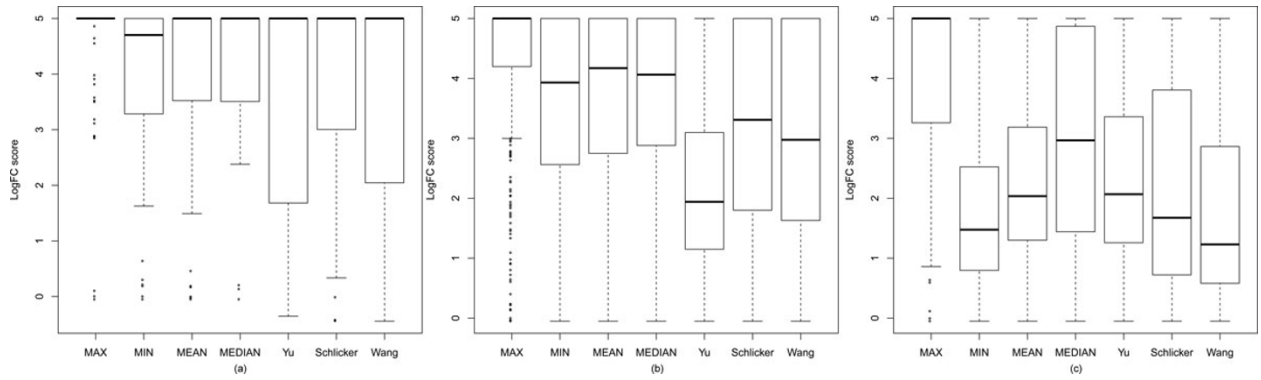

Figure 5 Logged fold change (LogFC) score comparison for four integration measures and three integrated measures. Logged fold change (LogFC) score comparison for four integration measures (MAX, MIN, MEAN and MEDIAN) and three integrated measures (Yu measure, Schlicker measure and Wang measure) in Molecular Function (MF) category on yeast (a), *Arabidopsis* (b) and human (c).

case they are likely to have identically annotated molecular functions [34]. In our test, sequence similarity scores ($ln(BitScore)$) of all genes in the *BG* of the three species were calculated with BLAST, resulting in 20,652 yeast, 772,609 *Arabidopsis* and 942,609 human gene pairs. As shown in Figure 7, the semantic similarity measurements show a correlation with sequence similarity. The covariance scores (see Additional file 8) on all of the three species reveal that InteGO is overall the best measure.

Conclusions

Comparing gene at the functional level is vital for various of applications [3-7]. The existing GO semantic based measures either calculate gene-to-gene similarities directly [13], or indirectly compute gene-to-gene similarities with term-to-term similarities [12,17]. Unfortunately, none of them takes into account all of the respects of rich information in GO (structure, annotation, LCA and all of the parents term, *etc*). In this paper, we proposed a new measure called InteGO to appropriately integrate the seed measures with the following advantages: 1) InteGO employs an adaptive approach which enables the optimization of seed measure integration; 2) it applies a rank-based integration approach, which unifies the scale and distribution differences among different seed measures; 3) InteGO is an open-platform measure that allows users to add/delete seed measures, redefine the background gene set and change the rank-based integration method.

To demonstrate the advantages of InteGO, we compared its EC-assigned gene similarities and sequence similarities with three existing measures (the Yu, Schlicker and Wang measure) in yeast, *Arabidopsis* and human. Comparing with these state-of-the-art measures, the experimental results show that InteGO increases the LogFC scores to about two-fold. It indicates that integrating multiple measures appropriately can improve the performance of the functional similarity measure. Especially, we found that taking the maximal ranks from all of the seed measures performs the best. The covariances between semantic

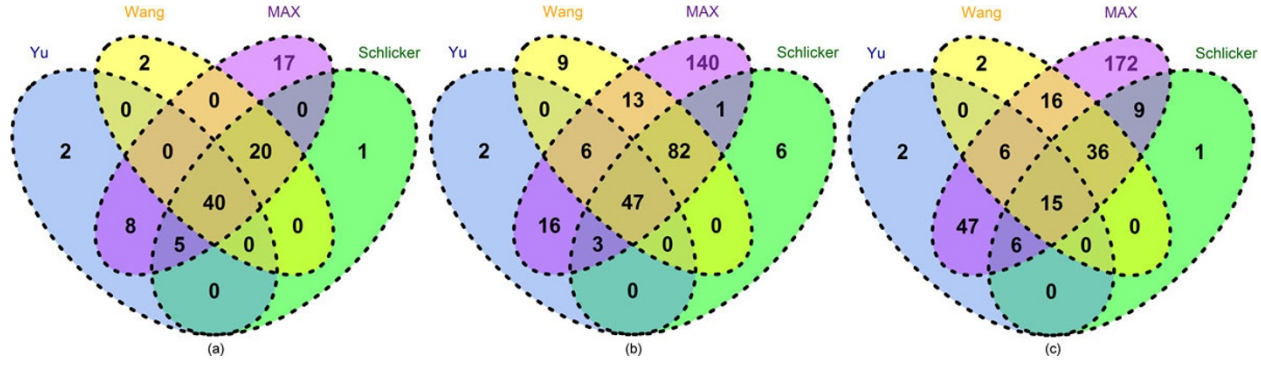

Figure 6 Venn Diagram for the best integration measure MAX and Yu measure, Schlicker measure and Wang measure. Venn Diagram for the best integration measure MAX and Yu measure, Schlicker measure and Wang measure with number of ECs on which perform best on yeast (a), *Arabidopsis* (b) and human (c). Purple, blue, green and yellow represent MAX measure, Yu measure, Schlicker measure and Wang measure respectively.

Table 2 Adjusted P-values for comparing MAX with Yu, Schlicker and Wang measure using TukeyHSD.

Measures	Adjusted p-value		
	yeast	*Arabidopsis*	human
MAX vs. Schlicker	2.8E-1	<**1.0E-7**	<**1.0E-7**
MAX vs. Wang	**1.0E-2**	<**1.0E-7**	<**1.0E-7**
MAX vs. Yu	**1.1E-4**	<**1.0E-7**	<**1.0E-7**
Wang vs. Schlicker	5.9E-1	9.6E-1	3.2E-1
Yu vs. Schlicker	6.0E-2	<**1.0E-7**	1.9E-1
Yu vs. Wang	5.8E-1	<**1.0E-7**	**1.2E-3**

Adjusted P-values for comparing MAX with Yu, Schlicker and Wang measure using TukeyHSD. Significant p-values are in bold fonts.

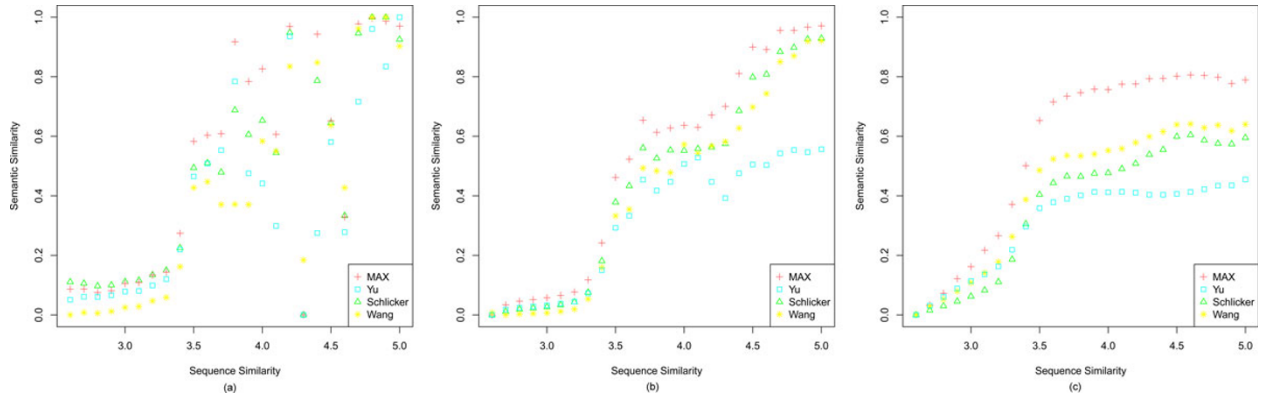

Figure 7 Comparing InteGO with the Yu, Schlicker and Wang measures with protein sequence similarity. Comparing InteGO with the Yu, Schlicker and Wang measures with protein sequence similarity on yeast (a), *Arabidopsis* (b) and human (c), where the x-axis is BLAST sequence similarity (*ln*(*BitScore*)) and y-axis is the normalized semantic similarity based on GO.

similarities and protein sequence similarities shows InteGO is clear the best out of all the tested measures.

In InteGO, to maintain a large background gene set is expensive. Therefore, extending InteGO from MF to BP or even other biological or medical ontologies is not a trivial problem. In the future, we will continue to improve InteGO to be more efficient and to be applicable on more ontologies. As an open framework, the performance of InteGO may be further improved by synergistically integrating more seed measurements. We will continue to integrate and compare InteGO with more recent gene-to-gene measurements in the future. We will continue to explore better integration methods, such as using EM algorithm to optimize the weight for each seed measure, to achieve better performance.

Additional material

Additional file 1: Average LogFC scores for different sizes of background set. To test whether the selection of *BG* will affect the integration performance, we compared the results for different background set on yeast. First, given the full set of *BG*, a subset of gene pairs were randomly selected with the percentage varying from 10% to 100%. This process was repeated for 100 times. Second, as shown in Additional file 1, the logFC scores for each subset size were calculated based on the randomly selected gene pairs. Since we do not use the full set, the computable ECs are also a subset of all of the computable ECs. In Additional file 1, the LogFC score increases linearly from 0 to 10 when the coverage increases from 10% to 90%, then suddenly jumps to a high score (13.8) when all of the background genes were used, indicating that first, the size of the background set affects the integration measure significantly, second, to use the full background set is the best, although it slightly increases the computational time.

Additional file 2: Distribution of the gene-to-gene similarities with Yu measure. Distribution of the gene-to-gene similarities with Yu measure for all of the genes in the Background Gene Set (BG) on yeast.

Additional file 3: Distribution of the gene-to-gene similarities with Schlicker measure. Distribution of the gene-to-gene similarities with Schlicker measure for all of the genes in the Background Gene Set (BG) on yeast.

Additional file 4: Distribution of the gene-to-gene similarities with Wang measure. Distribution of the gene-to-gene similarities with Wang measure for all of the genes in the Background Gene Set (BG) on yeast.

Additional file 5: The 95% family-wise confidence level of TukeyHSD test on yeast. The 95% family-wise confidence level of TukeyHSD test on yeast, which compared MAX with all the three seed measures (Schlicker, Wang and Yu measure).

Additional file 6: The 95% family-wise confidence level of TukeyHSD test on Arabidopsis. The 95% family-wise confidence level of TukeyHSD test on *Arabidopsis*, which compared MAX with all the three seed measures (Schlicker, Wang and Yu measure).

Additional file 7: The 95% family-wise confidence level of TukeyHSD test on human. The 95% family-wise confidence level of

TukeyHSD test on human, which compared MAX with all the three seed measures (Schlicker, Wang and Yu measure).

Additional file 8: The covariance sores comparing with sequence similarity. The covariance sores comparing with sequence similarity on yeast, *Arabidopsis* and human for Max, Yu, Schlicker and Wang measure.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
JC conceived the project. **JP**, **JC** and **YW** designed the algorithm and experiments. **JP** implemented the algorithm and finished the experiments.

Acknowledgements
This project has been funded by the U.S. Department of Energy (Chemical Sciences, Geosciences and Biosciences Division, grant no. DE-FG02-91ER20021 to J.C; the National High Technology Research and Development Program of China grant (no. 2012AA020404, 2012AA02A602 and 2012AA02A604) and the National Natural Science Foundation of China grant (no. 61173085) to Y. W.

Declarations
The publication costs for this article were funded by the corresponding author's institution.
This article has been published as part of *BMC Bioinformatics* Volume 15 Supplement 2, 2014: Selected articles from the Twelfth Asia Pacific Bioinformatics Conference (APBC 2014): Bioinformatics. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcbioinformatics/supplements/15/S2.

Authors' details
[1]School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China. [2]MSU-DOE Plant Research Laboratory, Michigan State University, East Lansing, MI 48824, USA. [3]Department of Computer Science and Engineering, Michigan State University, East Lansing, MI 48824, USA.

Published: 24 January 2014

References
1. Ashburner M, Ball CA, Blake JA, *et al*: **Gene Ontology: tool for the unification of biology.** *Nature Genetics* 2000, **25**:25-29.
2. Harris M, Clark J, Ireland A, *et al*: **The Gene Ontology (GO) database and informatics resource.** *Nucleic Acids Research* 2004, **32(Database)**:D258-D261.
3. Rhee S, Wood V, Dolinski K, Draghici S: **Use and misuse of the gene ontology annotations.** *Nature Review Genetics* 2008, **9(7)**:509-515.
4. Peng J, Chen J, Wang Y: **Identifying cross-category relations in Gene Ontology and constructing genome-specific term association networks.** *BMC Bioinformatics* 2013, **14(Suppl 2)**:S15.
5. Chen JL, Liu Y, Sam LT, Li J, Lussier Y: **Evaluation of high-throughput functional categorization of human disease genes.** *BMC Bioinformatics* 2007, **8(Suppl 3)**:S7.
6. Kemmeren P, Kockelkorn T, Bijma T, Donders R, Holstege F: **Predicting gene function through systematic analysis and quality assessment of high-throughput data.** *Bioinformatics* 2005, **21(8)**:1644-1652.
7. Zhu M, Gao L, Guo Z, Li Y, Wang JD, Wang , Wang C: **Globally predicting protein functions based on co-expressed protein-protein interaction networks and ontology taxonomy similarities.** *Gene* 2007, **391(1-2)**:113-119.
8. Lewis B, Shih I, Jones-Rhoades M, Bartel D, Burge C: **Prediction of mammalian microRNA targets.** *Cell* 2003, **115(7)**:787-798.
9. Vafaee F, Rosu D, Broackes-Carter F, Jurisica I, *et al*: **Novel semantic similarity measure improves an integrative approach to predicting gene functional associations.** *BMC systems biology* 2013, **7**:22.
10. Papadopoulos V, Baraldi M, Guilarte T, *et al*: **Translocator protein (18kDa): new nomenclature for the peripheral-type benzodiazepine receptor based on its structure and molecular function.** *Trends in Pharmacological Sciences* 2006, **27(8)**:402-409.
11. Lu Z, Hunter L: **GO molecular function terms are predictive of subcellular localization.** *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* NIH Public Access; 2005, 151.
12. Wang J, Du Z, Payattakool R, Yu P, Chen C: **A new method to measure the semantic similarity of GO terms.** *Bioinformatics* 2007, **23(10)**:1274-1281.
13. Yu H, Jansen R, Stolovitzky G, Gerstein M: **Total ancestry measure: quantifying the similarity in tree-like classification, with genomic applications.** *Bioinformatics* 2007, **23(16)**:2163-2173.
14. Batet M, Sanchez D, Valls A: **An ontology-based measure to compute semantic similarity in biomedicine.** *Journal of Biomedical Informatics* 2011, **44**:118-125.
15. Teng Z, Guo M, Liu X, Dai Q, Wang C, Xuan P: **Measuring gene functional similarity based on groupwise com-parison of GO terms.** *Bioinformatics* 2013, **29(11)**:1424-1432.
16. Resnik P: **Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language.** *Journal of Artificial Intelligence Research* 1999, **11**:95-130.
17. Schlicker A, Domingues F, Rahnenfuhrer J, Lengauer T: **A new measure for functional similarity of gene products based on Gene Ontology.** *BMC Bioinformatics* 2006, **7**:302.
18. Lin D: **An information-theoretic definition of similarity.** In *Proceedings of the 15th international conference on Machine Learning. Volume 1*. San Francisco; 1998:296-304.
19. Jiang J, Conrath D: **Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy.** *Proceedings of International Conference Research on Computational Linguistics* Taiwan; 1997.
20. Wu X, Pang E, Lin K, Pei Z: **Improving the Measurement of Semantic Similarity between Gene Ontology Terms and Gene Products: Insights from an Edge-and IC-Based Hybrid Method.** *PLoS One* 2013, **8(5)**:e66745.
21. Pesquita C, Faria D, Bastos H, Ferreira A, Falcao A, Couto F: **Metrics for GO based protein semantic similarity: a systematic evaluation.** *BMC Bioinformatics* 2008, **9(Suppl 5)**:S4.
22. Sevilla JL, Segura V, Podhorski A, *et al*: **Correlation between gene expression and GO semantic similarity.** *IEEE ACM Transactions on Computational Biology and Bioinformatics* 2005, **2(4)**:330-338.
23. Pesquita C, Faria D, Falcao A, Lord P, Couto F: **Semantic similarity in biomedical ontologies.** *PLoS computational biology* 2009, **5(7)**:e1000443.
24. Yang D, Tang J, Yang X, *et al*: **An integration strategy to measure enzyme activities for detecting irreversible inhibitors with dimethoate on butyrylcholinesterase as a model.** *International Journal of Environmental and Analytical Chemistry* 2011, **91(5)**:431-439.
25. Ward MO: **Xmdvtool: Integrating multiple methods for visualizing multivariate data.** *Proceedings of the Conference on Visualization* IEEE Computer Society Press; 1994, 326-333.
26. Goldkuhl G, Lind M, Seigerroth U: **Method integration: the need for a learning perspective.** *IEE Proceedings-Software* 1998, **145(4)**:113-118.
27. Claudel-Renard C, Chevalet C, Faraut T, Kahn D: **Enzyme-specific profiles for genome annotation: PRIAM.** *Nucleic Acids Research* 2003, **31(22)**:6633-6639.
28. Karp P: **Call for an enzyme genomics initiative.** *Genome Biology* 2004, **5(8)**:401.
29. Diaz-Mejia J, Perez-Rueda E, Segovia L: **A network perspective on the evolution of metabolism by gene duplication.** *Genome Biol* 2007, **8(2)**:R26.
30. Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes.** *Nucleic Acids Research* 2000, **28**:27-30.
31. Allison D, Cui X, Page G, Sabripour M: **Microarray data analysis: from disarray to consolidation and consensus.** *Nature Review Genetics* 2006, **7**:55-65.
32. O'Madadhain J, Fisher D, Smyth P, White S, Boey Y: **Analysis and visualization of network data using JUNG.** *Journal of Statistical Software* 2005, **10(2)**:1-25.
33. Miller R: *Simultaneous statistical inference* McGraw-Hill New York; 1966.
34. Lord P, Stevens R, Brass A, Goble C: **Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation.** *Bioinformatics* 2003, , **10:** 1275-1283.

doi:10.1186/1471-2105-15-S2-S5
Cite this article as: Peng *et al.*: Towards integrative gene functional similarity measurement. *BMC Bioinformatics* 2014 **15**(Suppl 2):S5.