BIOINFORMATICS

Vol. 23 ISMB/ECCB 2007, pages i529–i538
doi:10.1093/bioinformatics/btm195

Information theory applied to the sparse gene ontology annotation network to predict novel gene function

Ying Tao[1,†], Lee Sam[2], Jianrong Li[2], Carol Friedman[1] and Yves A. Lussier[1,2,3,*,‡]

[1]Department of Biomedical Informatics, Columbia University, 622 West 168th Street, VC5, New York, NY 10032, [2]Center for Biomedical Informatics, The University of Chicago, 5841 S Maryland Avenue, Chicago, IL 60637, USA and [3]UCCRC, The University of Chicago, 5841 S Maryland Avenue, Chicago, IL 60637, USA

ABSTRACT

Motivation: Despite advances in the gene annotation process, the functions of a large portion of gene products remain insufficiently characterized. In addition, the *in silico* prediction of novel Gene Ontology (GO) annotations for partially characterized gene functions or processes is highly dependent on reverse genetic or functional genomic approaches. To our knowledge, no prediction method has been demonstrated to be highly accurate for sparsely annotated GO terms (those associated to fewer than 10 genes).

Results: We propose a novel approach, information theory-based semantic similarity (ITSS), to automatically predict molecular functions of genes based on existing GO annotations. Using a 10-fold cross-validation, we demonstrate that the ITSS algorithm obtains prediction accuracies (precision 97%, recall 77%) comparable to other machine learning algorithms when compared in similar conditions over densely annotated portions of the GO datasets. This method is able to generate highly accurate predictions in sparsely annotated portions of GO, where previous algorithms have failed. As a result, our technique generates an order of magnitude more functional predictions than previous methods. A 10-fold cross validation demonstrated a precision of 90% at a recall of 36% for the algorithm over sparsely annotated networks of the recent GO annotations (about 1400 GO terms and 11 000 genes in *Homo sapiens*). To our knowledge, this article presents the first historical rollback validation for the predicted GO annotations, which may represent more realistic conditions than more widely used cross-validation approaches. By manually assessing a random sample of 100 predictions conducted in a historical rollback evaluation, we estimate that a minimum precision of 51% (95% confidence interval: 43–58%) can be achieved for the human GO Annotation file dated 2003.

Availability: The program is available on request. The 97 732 positive predictions of novel gene annotations from the 2005 GO Annotation dataset and other supplementary information is available at http://phenos.bsd.uchicago.edu/ITSS/

Contact: Lussier@uchicago.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

In the postgenomic era, annotating gene functions using standardized vocabularies, such as the Gene Ontology (GO), has become a critical task for biologists due to the massive numbers of genes identified though sequencing. GO is organized as a hierarchical structure containing ontological knowledge of biology, which has been manually developed by human experts (Ashburner *et al*., 2000). Despite advances in the gene annotation process, many gene products are still left poorly characterized. For example, though the number of GO annotations for *Homo sapiens* genes increased 66% from 2003 to 2005, the GO Consortium currently only provides annotations for about 16 000 of the ~25 000 known human genes, indicating that a large number of genes remain to be functionally characterized.

Methods for predicting annotations of gene products fall into the rough categories of experimentally based and knowledge-based approaches. In general, experimentally based approaches depend on direct experimental information about genes, while knowledge-based approaches rely on existing knowledge (e.g. results from previous experiments, biomedical literature, GO Annotation datasets, etc.). Experimentally based methods generally focus on a single scale of biology such as protein conformation (Laskowski *et al*., 2005) or gene sequence (Jones *et al*., 2005; Khan *et al*., 2003). In contrast, knowledge-based approaches, such as those employing the literature or GO, provide opportunities for prediction using knowledge from multiple scales of biology. Literature-based methods, including indexing (Perez *et al*., 2004), natural language processing (Chiang *et al*., 2006), computational reasoning (Bada *et al*., 2004) and statistical analysis (Andrade and Valencia, 1998), have generally achieved below 70% precision at predicting gene function. While hybrid approaches exist, most focus on a specific experimental data type (Chen and Xu, 2004; Kemmeren *et al*., 2005) or are difficult to interpret (Shahbaba and Neal, 2006).

In contrast, GO-based methods have been shown to achieve higher accuracies when used on a small number of GO terms in densely annotated regions of the ontology. The GO Annotations provide standardized and integrated gene function annotations, incorporating relevant literature and experimentally based measurements from multiple scales of biology, many of which have been manually curated. It is therefore a unique data source for inferring such annotations based on multiple physical properties.

*To whom correspondence should be addressed.
†This author completed his PhD at Columbia University in 12/2006 and is no longer affiliated with the institution.
‡This author is currently affiliated to The University of Chicago (2006–present). The work was partially conducted during his tenure at Columbia University (2001–2006).

King *et al.* (2003) proposed the first accurate method for doing so using only existing GO annotation patterns by using machine learning algorithms. Using known GO annotations as a gold standard, King *et al.* obtained a precision of 97.7% using decision trees, and a precision of 93.7% and recall of 50% using Bayesian networks (BN) using data from the Saccharomyces Genome Database (SGD) (Cherry *et al.*, 1998). Using FlyBase (Mitchell *et al.*, 2003) data, they obtained a similar results with a precision of 87.5% using decision trees, and a precision of 78.7% and recall of 50% with BNs. However, these levels of precision and recall were only achieved though strict filtering of the datasets, most significantly requiring each candidate GO term be associated to at least 10 genes. This cut the number of candidate GO terms by over an order of magnitude to 170 for SGD and 218 in FlyBase.

To our knowledge, no prediction method has been demonstrated to be accurate for GO terms associated to fewer than 10 genes, an important consideration as the vast majority of GO terms utilized in the annotations fit in this category (e.g. 82.5% of GO terms in *H.sapiens* annotations are associated with less than 10 genes). In addition, current predictions using GO do not use the ontological similarity between otherwise distinct genes annotations.

The semantic similarity between two concepts, or groups of concepts, has been used extensively in the domain of computer science for information retrieval and natural language processing tasks (Jiang and Conrath 1999; Lee *et al.*, 1993) as well as for *k*-nearest neighbor (KNN) machine learning tasks (Yuseop *et al.*, 2001). Recently, semantic similarity has also been utilized within the biological domain for predicting protein–protein interaction networks (Wu *et al.*, 2006) as well as investigating the relationships between GO annotations and gene sequences [Lord *et al.*, 2003a, b] and microarray expression profiles (Wang *et al.*, 2005). Semantic similarity has also been used in clustering genes functionally, a different task from predicting novel gene functions (Chen *et al.*, 2007; Wang *et al.*, 2005). Previous studies have integrated semantic similarity and KNN methodologies to improve missing value estimations in microarray data (Tuikkala *et al.*, 2006), and analyze gene expression data in coordination with an ontology-driven clustering method (Wang *et al.*, 2005). However, to our knowledge, the proposed method is the first use of an information theory-based semantic similarity (ITSS) approach for assigning novel gene functions to known genes directly from the geometry of the network of the GO annotations and the overarching GO alone. We use knowledge from the GO hierarchies to derive predictions with the hopes that by maximizing the number of utilized GO concepts, these predictions will be based upon as much information as possible. The accuracy of the ITSS method has been established for a significantly broader number of GO annotations than previous methods (King *et al.*, 2003), which were evaluated over a small number of GO terms with more constraints.

In this article, we describe a novel technique, named ITSS, for predicting new gene annotations based exclusively on existing GO annotations, and present the results of an evaluation, which show a higher recall than previously reported methods. Given that methods for predicting gene annotations using homology of physical properties such as sequence and expression profile have been proven successful in the past, it is reasonable to speculate that we may also be able to employ the semantic similarity of gene annotations. In this research, we hypothesize that semantic similarity measurements between groups of concepts based on information theory can be used to predict new annotations associated with a gene. The basis of the ITSS approach we propose is a KNN algorithm using an ITSS measure as the metric for assigning new relationship edges to concept nodes in the network. The predictions described in this article rely on two semantic similarity scores: (1) between two genes' concepts in the GO annotations, and (2) between two groups of GO concepts within the ontology. Through the use of this technique, we are able to more fully exploit the ontological knowledge contained in the structure of GO and its annotations using semantic similarity scores to calculate predictions of novel gene annotations, and provide more interpretable predictions over a broader number of GO terms and genes than previously evaluated prediction methods.

2 SYSTEM AND COMPUTATIONAL METHODS

We have developed the ITSS algorithm to assign unknown annotations to a gene based on the similarity of its known annotations and those of other genes. In this section, we will first introduce the algorithm used for calculating semantic similarity between any two concepts. Next, we will describe the algorithm for calculating the semantic similarity between any two groups of concepts, and last, we will explain how semantic similarity can be used as a metric in the KNN algorithm for predicting new GO annotations for a gene.

2.1 Semantic similarity between any two concepts within an ontology

The first algorithm is for calculating the semantic similarity between 'any two concepts' in an ontology. For example, the simplified ontology seen in Figure 1a consists of nine different concepts. 'Any two concepts' means that the algorithm can be used to calculate the semantic similarity between any two concepts, including identical concepts.

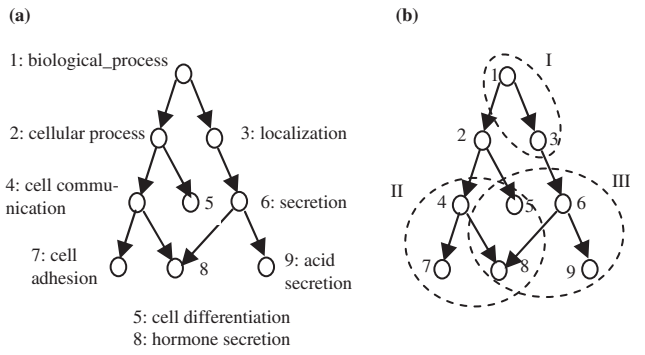

Fig. 1. Semantic similarity between concepts. The semantic similarity between any two concepts, or any two groups of classified concepts is illustrated. (**a**) Semantic similarity can be calculated between any two of the nine concepts. (**b**) Semantic similarity can also be calculated between any two arbitrarily defined groups of concepts. Group I contains concepts 1 and 3, Group II is comprised of concepts 4, 5, 7 and 8, and Group III contains concepts 5, 6, 8 and 9. Concepts can be shared between concepts (e.g. concepts 5 and 8 are members of both Group II and Group III).

The GO is comprised of three subontologies, 'molecular functions', 'cellular components' and 'biological processes'. Because these three subontologies contain orthogonal types of entities, they are considered to be different ontologies in our methods. Therefore, the algorithms described in this section will calculate the semantic similarity between any two concepts from the same subontology in GO. If the two concepts are in different subontologies of GO, then semantic similarity is equal to be zero. For example, the semantic similarity can be calculated between the two concepts 'oxidoreductase activity' and 'peptidase activity', which are both from the same subontology of GO, 'molecular function'.

There are generally three main algorithms, based on information theory, for calculating the semantic similarity between two concepts in an ontology, which were respectively proposed by (Jiang and Conrath, 1997), Lin (1998) and Resnik (1995). In our study, we used Lin's algorithm because it returns a normalized value between 0 and 1, and outperformed other methods in our dataset (Supplementary Fig. S1). Lin's algorithm for calculating the semantic similarity between concepts a and b is defined as:

$$sim(a,b) = 2 \times ic(ms(a,b))/[ic(a) + ic(b)] \quad (1)$$

where

- ic (c), the information content of c, is defined as $-\log(p(c))$, where p (c) is the probability of the occurrence of c. In this study, the occurrence probability of a concept c is defined in Equation (2) (Lord *et al.*, 2003a)

$$p(c) = \frac{(1 + \text{number of all descendants of c})}{\text{total number of concepts in an ontology}} \quad (2)$$

- ms (a,b), the minimum 'subsumer' of concepts a and b, is defined as the common ancestor that has the minimum probability of occurrence.
- ic (ms (a, b)), therefore, is the information content of the minimum 'subsumer' of concepts a and b.
- *Example of the calculation*. To compute the semantic similarity between 'protein binding' and 'single-stranded DNA binding', we note that 'protein binding' has 561 descendants, 'single-stranded DNA binding' has 2 descendants, and the entire 'molecular function' hierarchy contains 7063 concepts. Thus p ('protein binding') $= (1 + 561)/7063 = 0.0796$ and p ('single-stranded DNA binding') $= (1 + 2)/7063 = 0.000425$. Their minimum 'subsumer' is 'binding', which has 961 descendants with p ('binding') $= (1 + 961)/7063 = 0.136$. Therefore, the semantic similarity according to Lin's algorithm is $2 \times (\log 0.136)/[-\log 0.0796 - \log 0.000425] = 0.388$.

2.2 Semantic similarity between two groups of concepts

The second algorithm calculates the semantic similarity between 'any two groups' of concepts within an ontology based on the similarity between a pair of GO concepts calculated as described in the first step. These two groups can be obtained in any way as long as they are all in the same ontology. For example, using the ontology seen in Figure 1b, we can arbitrarily select groups of concepts, such as Groups I, II and III. The semantic similarities can be calculated between any two of these arbitrarily defined groups. These groups can also share identical concepts as shown in Figure 1.

In this particular research, we define a group of concepts as those GO concepts that are associated with a single gene. For example, all of the concepts within the 'molecular function' subontology that are associated with the gene BRCA1 (breast cancer 1, early onset) compose a group, which contains the concepts 'DNA binding', 'protein binding' and 'transcription coactivator activity'. All of the concepts within the 'molecular function' subontology that are associated with the gene BRCA2 (breast cancer 2, early onset) comprise another group, which contains the concepts 'nucleic acid binding', 'protein binding' and 'single-stranded DNA binding'. The semantic similarity between these two groups tells how similar the genes BRCA1 and BRCA2 are in terms of their molecular functions.

Based on these methods for determining the degree of similarity for a pair of concepts, we used the following 'pairwise' method for calculating the semantic similarity between two groups of concepts within an ontology. The pairwise algorithm (Jiang and Conrath, 1999) was compared to the 'cross-join' algorithm (Wang *et al.*, 2004), and was found systematically superior in three preliminary studies (Supplementary Fig. S1). Before performing the semantic similarity calculation, the concepts within one group are paired-up with those of another group. This pairing process is illustrated in Figure 2. First, for each concept in group A, the most similar concept is found in group B. Then, for each concept in group B, the most similar concept is found in group A. If two concepts across the groups are reciprocally found to be most similar to one another, these two concepts are considered to be a pair. All of the reciprocal pairs constitute a set P, which is always non-empty because each concept will always have a 'most similar' partner concept. The 'pairwise' formula for two groups of concepts is:

$$\frac{sim(A,B) = 2 \times \sum_{(a_i,b_i) \in P, sim(ai,bi) \geq t} sim(a_i,b_i)}{(|A| + |B|)} \quad (3)$$

where

- A, B represent the two groups of concepts; (a_i, b_i) is a pair in P, the same indices i means that a_i and b_i are from the same pair.
- If the similarity between a_i and b_i is too low, we usually do not regard them as a pair. Therefore, to reduce noise, we use a **threshold value t** to remove pairs with low similarities.

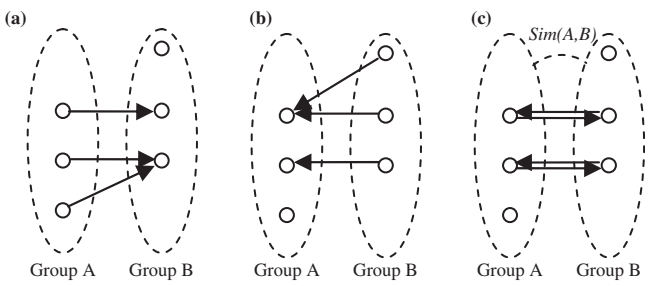

Fig. 2. Determining semantic similarity between groups of concepts using a pair-wise method. The small circles represent concepts, and the dashed ovals indicate the groups of concepts. The geometric distances between the circles illustrate the semantic distances between concepts; a larger semantic distance indicates a lower semantic similarity between concepts. (**a**) For each concept in Group A, the concept in Group B with the maximum semantic similarity (i.e. shortest distance) is determined. The arrows pointing from Group A to Group B indicate these relations. (**b**) For each concept in Group B, the concept in Group A with the maximum semantic similarity (shortest distance) is determined. The arrows pointing from Group B to Group A indicate these relations. (**c**) The bidirectional arrows illustrate the resulting reciprocal relations that are returned as pairs of concepts with the maximum semantic similarity. The similarity score $sim(A,B)$ is calculated using Equation (3).

Table 1. An example of similarity score results from the KNN algorithm for gene annotation predictions

	Genes	Similarity score to target gene according to Equation (3)	DNA repair? (GO:0006281)
Target gene	*BRCA2*	−	?
Training genes	*BRCA1*	0.646	+
	TBPL1	0.614	−
	APEX1	0.613	+
	...	...	...

- $|A|$ and $|B|$ represent the numbers of concepts in set A and B. The items $|A|$ and $|B|$ are used here to reduce the calculated impact of groups with extra concepts beyond paired concepts.

2.3 Prediction of new annotations for a gene using ITSS

Based on the metric of semantic similarity between two concepts or two groups of concepts, the ITSS method employs the simple KNN classification algorithm (Duda and Hart, 1973) to predict new annotations for a gene. The process of KNN is illustrated in the example in Table 1, and detailed below:

- Select a target gene and a target GO term. In the example in Table 1, we selected a gene that is known to be highly related to breast cancer, BRCA2. Our goal is to predict whether BRCA2 participates in the biological process 'DNA repair'. In the GOAh file dated 2003, 'DNA repair' was not associated with BRCA2. However, this annotation was added to the GOAr file dated 2005.
- Calculate the semantic similarities between GO annotations of the target gene and GO annotations of all genes in the training set based on Equation (3). In this example, we set the threshold value $t = 0.7$ based on a previous optimization process.
- Sort, in descending order, the genes in the training set according to the semantic similarities of their annotations to the target gene. Table 1 shows the first 10 training genes, their semantic similarities and categories (i.e. whether they have been annotated as 'DNA repair'). Their categories are '+', indicating the gene has the annotation 'DNA repair', or '−', indicating that the gene does not have this annotation.
- Collect the categories of the first k-training genes based on a predefined k value. In the example in Table 1, if we set $k = 4$, then we will obtain the categories of BRCA1, TBPL1, APEX1 and TRIM24.
- Apply different **cutoff values**, a positive integer less than k, to the number of positive categories required to obtain the prediction category for the target gene. If the number of positive categories is greater than the cutoff value, then a positive category will be returned. Otherwise, a negative category will be returned. In the example in Table 1, if we use cutoff value of either 0 or 1, a positive category will be returned, because the number of positive cases is 2, which is greater than both 0 and 1. However, if we use a cutoff value equal to 2 or 3, then we will get a negative category because the positive number 2 is not greater than either of the cutoff values.

2.4 ITSS parameter optimization

To obtain the best predictions, there are two parameters of the ITSS algorithm that must be optimized: (1) the number of neighbors is KNN (***k*-value**), and (2) a similarity threshold [***t*-value** in Equation (3)]. The optimal k-value was determined by randomly selecting 100 or 500 genes to comprise a testing set and applying different values for k. The optimal **value** t was determined by using the entire datasets of genes. The values of k and t were judged as optimal when the prediction F values are maximal.

2.5 Statistical analysis

The performance of the different prediction algorithms was assessed by comparing the areas under the resulting receiver operating characteristic (ROC) curves, calculated using the 'trapezoidal rule'. The SE of the area under an ROC curve is calculated using the following Equation (4):

$$SE(A) = \sqrt{\frac{[A(1-A) + (n_a - 1)(Q_1 - A^2) + (n_n - 1)(Q_2 - A^2)]}{(n_a + n_n)}} \quad (4)$$

where A is the area under the curve, n_a and n_n are the number of positive and negative results, respectively, taken from the gold standard, and Q_1 and Q_2 are estimated by $Q_1 = A/(2-A)$ and $Q_2 = 2A^2/(1+A)$. Equation (5) defines the SE of the difference between two areas A_1 and A_2:

$$SE(A_1 - A_2) = \sqrt{SE^2(A_1) + SE^2(A_2)} \quad (5)$$

The z-score is equal to $|A_1 - A_2|/SE(A_1 - A_2)$, indicating how far and in which direction the observation deviates from its distribution's mean expressed in units of its distribution's SD. The conservative Bonferroni-type adjustment (Sokal and Rohlf, 1995) accounted for the multiple *a posteriori* comparisons with two types of random controls.

3 RESULTS AND EVALUATION

3.1 Materials

We used a GO file that contains hierarchical relations organized as three exclusive axes of biological concepts. The version of GO used in this study, dated August 2005, was downloaded from http://www.geneontology.org/GO.downloads.shtml, containing 9633 distinct Biological Processes, 1570 distinct Cellular Components and 7063 distinct Molecular Functions, excluding the 1000 terms annotated as obsolete.

Two GOA files for *H.sapiens*, which contain annotations relating human genes to their biological processes, molecular functions and cellular components in GO, were used in this study. The GOAh in this article, is dated March 2003, and was obtained directly from NCBI. GOAh contains 51 830 distinct gene-GO entries, including 11 221 distinct human genes and 3448 unique GO terms. The second GO Annotation file, referred as GOAr, is dated August 2005, and was downloaded from NCBI's Entrez Gene at (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene). It contains 86 348 distinct gene-GO entries, including 15 442 distinct human genes and 4610 distinct GO terms. A detailed comparison is summarized in Table 2.

Table 2. Summary of the content of two GO Annotation (GOA) tables: historical GOA file (GOAh) dated March 2003, and more recent GOA file (GOAr) dated August 2005

	Distinct GO terms				Distinct genes
	Total	Terms having 10 or more gene annotations (%)	Terms having 3–9 gene annotations (%)	Terms having two or less gene annotations (%)	Total
GOAh	3511	648 (18%)	954 (27%)	1909 (55%)	11 221
GOAr	4610	832 (18%)	1327 (29%)	2451 (53%)	15 442

3.2 Experiments

In order to determine the accuracy of the predictions, we conducted two experiments and an in-depth manual evaluation:

(i) A 10-fold cross-validation was performed to compare ITSS to published predictive algorithms on the GO annotations databases of SGD and FlyBase.
(ii) As no such previous studies exist for *H.sapiens*, a 10-fold cross-validation in conditions comparable to that of the first experiment, and a 'historical rollback' validation were conducted on the *H.sapiens* database. We manually assessed 100 randomly selected positive predictions from the *H.sapiens* data resulting from the use of the optimal algorithm parameter values derived from the validation studies.

3.3 Experiment 1. Comparison of ITSS to published predictive algorithms for the SGD and FlyBase datasets

To evaluate the ITSS approach in comparison to other machine learning algorithms that do use semantic distance-based methods, we compared the prediction results of the ITSS algorithm to those of the Decision Tree and Bayesian Network studies performed by King *et al*. (2003). To obtain fair comparisons, we repeated the experimental methods of King *et al*. (2003) as precisely as possible. As the original SGD and FlyBase GOA files were not available, we used SGD and FlyBase GOA files from 2005, and removed entries later than 22 January 2002 according to their PubMed IDs, to produce datasets of relatively similar size to those utilized by King *et al*. (2003) who used a SGD file containing 6403 genes and a FlyBase file containing 13 500 genes; we calculated datasets containing 6099 and 11 142 genes, respectively.

We replicated the 10-fold cross-validation methods utilized by King *et al*. (2003) and also followed their procedures to get similar sets of GO terms. For this study, we used 165 GO terms for testing in both the SGD and FlyBase datasets, drastically reduced from the entire SGD (2261) or FlyBase (3859) datasets due to the 10 gene association constraint. The genes in each GOA file were randomly partitioned into 10 sets of approximately equal size. Each of these 10 sets of genes will be used as testing set, in turn, and the aggregate of the remaining nine sets as the training set. The task was to predict if a gene in the testing set is associated with a certain GO term, using the known annotations in the corresponding GOA files as a gold standard. Knowledge of any association between the gene and the target term is hidden from the ITSS algorithm in order to provide an unbiased binary association prediction for the term.

If a prediction was positive, i.e. the gene was indeed assigned to the term in the GOA file, the prediction was considered a true positive (TP). If a prediction was positive but the gene was not assigned to the term in the GOA file, then this prediction was considered a false positive (FP). If a prediction was negative but the gene was not assigned to the term in the GOA file, then this prediction was counted as a true negative (TN). If a prediction was negative and the gene was assigned to the term in the GOA file, then this prediction was considered to be a false negative (FN). The True-Positive Rate (TPR), equal to TP/(TP+FN) and the False-Positive rate (FPR), equal to FP/(FP+TN), were then calculated. The prediction results were represented as ROC curves, in which the FPR is plotted on the x-axis and the TPR on the y-axis (Metz, 1978). The cutoff values of the ITSS algorithm were varied to generate the different data points on the curves.

In order to demonstrate that the predictions are effective, we employed two types of random controls. The first control, random algorithm (RA), assigned a random decimal number to the value of calculated semantic similarity used in the ITSS. The second control, random data (RD), follows a permutation resampling design where the ITSS algorithm was applied to a fictitious GOAr annotation file constructed by randomly shuffling the relationships between genes and their GO annotations to purposely randomize annotation patterns while preserving the total number of occurrences of each annotated GO terms. RA and RD provide an estimate of the maximum number of FP predictions which can be useful to understand the meaning of the observed uncorroborated predictions in the full study.

We conducted these evaluations with optimized parameters for the ITSS algorithm ($k=4$, $t=1$). The comparisons of ROC curves are shown in Figure 3. Because in biological predictions a low FPR is usually more desirable than a high TPR, we used the ROC area comparison method (Hanley and McNeil, 1983) in only the areas of those ROC curves where the FPR was below 0.001. In the SGD dataset, as shown in Figure 3a, DT performed slightly better than ITSS algorithm, but the difference was not statistically significant ($z=1.538$, $P=0.124$), and the results of the proposed ITSS algorithm were a little better than those associated with the BN, but again,

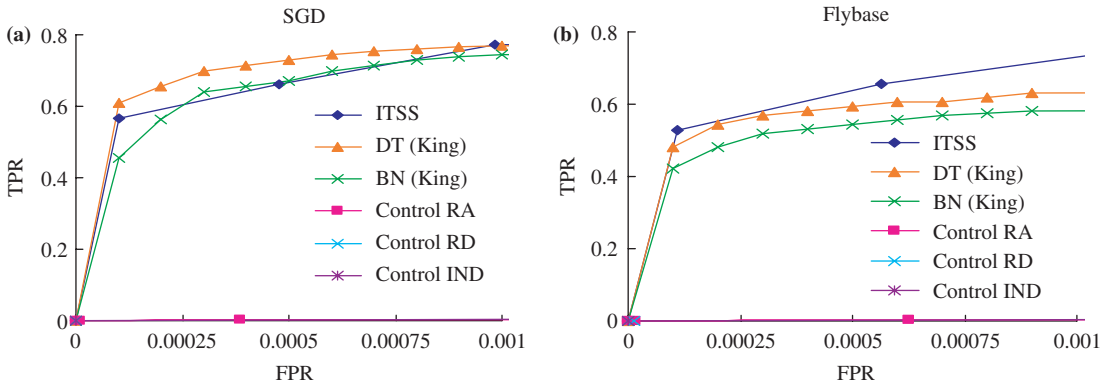

Fig. 3. ROC Curves for comparisons of ITSS to previous machine learning approaches using 10-fold cross-validation. **(a)** Comparison of methods in SGD dataset using the 10 gene association constraints to obtain comparable datasets to previously published results. **(b)** Comparison of methods in Flybase dataset similarly constrained as the SDG dataset. ITSS: information theory-based semantic similarity algorithm, Control RA: random algorithm control of ITSS, Control RD: random data control of ITSS, DT (King): decision trees by King *et al*., BN (King): Bayesian's networks by King *et al*., Control IND: independent control by King *et al*. It should be noted that the curves of controls are so close to the horizontal axis that they can hardly be seen.

Table 3. Comparisons of ITSS algorithm to other machine learning algorithms used in previously published work

GOA dataset	Prediction method	Total predictions	# of Genes	# of GO concepts	Precision (%)	Recall (%)
Panel a[a]						
SGD	DT	1 088 510	6 403	*170*	98	50
	BN	1 088 510	6 403	*170*	94	50
	ITSS	1 006 335	6 088	*165*	95	57
FlyBase	DT	2 943 000	13 500	*218*	87	50
	BN	2 943 000	13 500	*218*	80	50
	ITSS	1 838 430	11 142	*165*	94	53
Panel b[b]						
SGD	ITSS	7 172 424	6 099	*1 176*	48	52
FlyBase	ITSS	20 946 960	11 142	*1 180*	52	54

[a]Results of the 10-fold cross-validation using decision trees (DT) and BN conducted by King *et al*. (2003) using GO terms associated with at least 10 genes when recall is close to 50% and performance of ITSS in comparable conditions.
[b]Results of 10-fold cross-validation using ITSS algorithm with GO terms associated with at least three genes when recall is close to 50% (Conditions in which previous algorithms were not demonstrated to operate). *There is a significant 6-fold increase of GO concepts* upon which the ITSS can operate in condition (panel a) as compared to (panel b) (in italic font) in Table 3.

the difference was not statistically significant ($z = 0.439$, $P = 0.67$). In the FlyBase dataset, as shown in Figure 3b, the ITSS algorithm produced significantly better predictions than either the DT ($z = 2.34$, $P = 0.019$) or BN ($z = 4.9$, $P = 8.4 \times 10^{-7}$) methods. As illustrated in Table 3, the ITSS method performs as well or better than the DT and BN methods.

To explore the performance of the method in real-world conditions where most genes are poorly or not annotated (less than 10 gene annotations per GO) and previous methods based on annotation have not been demonstrated to operate, we applied the ITSS algorithm to the entire SGD dataset comprised of 2261 distinct GO terms and 6099 genes. We stratified the accuracy of the calculated predictions according to the number of genes associated with each GO term, and found that the ITSS algorithm performed well (above 0.6 precision and 0.5 recall) for those GO terms that were associated with three or more genes. Therefore, we performed 10-fold cross-validation evaluations incorporating all GO terms with at least three associated genes in both the SGD and FlyBase datasets, summarized in Table 3. The total number of TP predictions resulting from these experiments was over three times larger than those presented in previous studies.

3.4 Experiment 2. Predictions in the *H.sapiens* dataset

3.4.1 10-fold cross-validation For the *H.sapiens* dataset, we initially conducted a 10-fold cross-validation using both the GOAr and GOAh files. After removing those GO terms marked as 'obsolete' and the three ambiguous terms 'biological process unknown' 'molecular function unknown' and 'cellular component unknown', we further limited our dataset to include only those GO terms that had at least three associated genes. As a result, we obtained 2072 and 1390 distinct GO terms from the

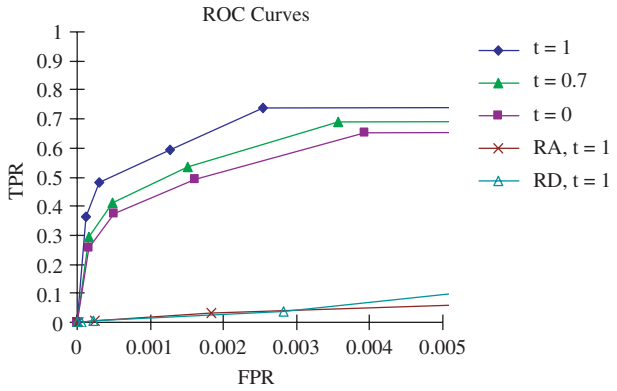

Fig. 4. ROC curves of GOAr dataset in 10-fold cross-validation. The precision–recall curve of GOAh is available in Supplement S2.

Table 4. Summary of manual validation results for 100 randomly selected predictions obtained from GOAh

Expert curator's opinion	Number of predictions	Examples	
		Gene	GO term
Correct (found in GOAr and confirmed by the expert)	17	GJA4	Cell communication
Correct (judged by the expert and confirmed with journal article)	34	ZNF638	DNA binding
Uncorroborated	49	WNT2	Cell–cell signaling
Total	100		

GOAr and GOAh files, respectively. Because at least one annotation is necessary as the clue for making the prediction and one as the target GO term, we also limited the datasets used in these validation experiments to include only those genes with at least two associated GO terms. We obtained 13 509 and 11 076 such genes from the GOAr and GOAh files, respectively. The GOA dataset was not filtered with respect to evidence code (all annotations were kept). Thus, we generated 27 990 648 (2072 GO terms × 13 509 genes) predictions applying the 10-fold cross-validation methods over the GOAr dataset, among which 77 602 positively corresponded to the gold standard. We also derived 15 395 640 (1390 GO terms × 11 076 genes) predictions based on the GOAh dataset, of which 44 020 predictions were positive according to the gold standard.

Figure 4 shows the predictions resulting from the application of the ITSS algorithm to the GOAr file during the 10-fold cross-validation experiment as TP-FP curves with a variable parameter t (threshold). The precision-recall curve of GOAr as well as the predictions resulting from the 10-fold cross-validation utilizing the GOAh file can be found in Supplementary Figure S2. The ROC and precision-recall curves results from GOAr and GOAh (Fig. 4 and Supplementary Fig. S2) are very similar. In this evaluation, when applying the optimization methods to the ITSS algorithm, as described in Methods section, we obtained the best predictions when $k=4$ and $t=1$. The impact of parameter t is illustrated in Supplementary Figure S2. As shown in Figure 4, the ITSS algorithm provides significantly better predictions over the GOAr dataset than either of the two controls ($z>217$, $P<2.2\times10^{-16}$ when compared to RA, and $z>216$, $P<2.2\times10^{-16}$ when compared to RD, (see Methods, subsection 'Statistical Analysis'). The maximum precision was 90% at a recall of 36%, and the maximum recall was 74% with a precision of 45%.

3.4.2 Historical rollback validation To further evaluate the ITSS algorithm in situations that mirror real life, we predicted new annotations using the older GO association file GOAh (2003) in the *H.sapiens* dataset and then validated the newly predicted annotations using the newer association file GOAr (2005) as a second evaluation. Using similar procedures as our cross-validation, we excluded from the GOAh file those GO terms marked as 'obsolete' and the three ambiguous terms 'biological process unknown', 'molecular function unknown' and 'cellular component unknown'. We further limited our testing dataset to include only those GO terms that had at least three associated genes, resulting in 9589 genes and 1377 GO terms from the GOAh file.

To further validate the effectiveness of the ITSS method, a blinded expert manually examined a sample of 100 random positive predictions from GOAh that were randomly selected from a corpus of the 2704 most plausible positive predictions obtained by using the best parameters for the ITSS algorithm (as determined by the optimization method described in Section 2): $k=4$, $t=0.7$, and a cutoff equal to 3. This set of 2704 positive predictions can be found in Supplementary Figure S3. The expert was a senior postdoctoral molecular biology research scientist with more than 10 years of laboratory experience.

A summary of the manual assessment results is provided in Table 4. Of the 100 assessed predictions, 51 were considered correct and validated in the scientific literature according to the expert, leaving 49 uncorroborated, but not necessarily wrong. Of the corroborated predictions, 17 were found directly in the GOAr file, and 19 others were found to be a parent of the GO concept associated to the gene in the GOAr file. For example, the gene MTERF was predicted to be associated with DNA binding (GO:0003677), which is the direct parent of double-stranded DNA binding (GO:0003690) in the GOAr file, and was judged to be correct by the expert. Thus, accepting direct parents to be correct predictions, as they are related on a high level, improved the measurement of the precision significantly. An additional six predictions could have been determined to be correct by extending the gold standard to include all ancestors of the concepts in GOAr. None of the uncorroborated predictions would have been erroneously assigned a TP value, as judged by the expert, if all ancestors of the concepts in GOAr file were to be accepted as the gold standard. However, by extending the gold standard to include all descendents of GO terms found in the GOAr file, 20 uncorroborated and 8 additional corroborated results are generated. There was one prediction (gene TNFSF15 associated with

GO:0007267: cell-cell signaling) in which the GO term cell-cell signaling has no hierarchical relations with any terms in the GOAr file. This prediction was validated by the expert based on published literature (Haridas *et al*., 1999). Therefore, the final number of correct predictions, as validated by the expert, was 51, yielding a precision of 51%. (95% confidence interval: 43–58%, $n = 100$). The confidence interval was determined using the hypergeometric distribution. Additional details and bibliographic references can be found in Supplementary Figure S4. We also applied the ITSS prediction algorithm to the GOAr file, and generated 97 732 new positive predictions (Supplementary Fig. S5).

4 DISCUSSION

Comparison with previous studies shows that the ITSS prediction approach is able to produce comparable or better predictions than the best implementations of DT or BN when applied to similar datasets. Most importantly, the ITSS algorithm was able to make predictions in the sparsely annotated GO terms, although precision of the resulting predictions dropped from 90% to approximately 50% for a constant recall of about 50%. This functionality is particularly important because GO terms with fewer than 10 gene annotations, which were excluded from previous prediction studies, occupy over 80% of total number of annotated GO terms that represent biological processes, cellular components and molecular functions. We demonstrated that the ITSS method is capable of generating predictions for these previously untapped GO terms of sparsely annotated GO terms, ultimately providing a 3-fold increase in the number of TP predictions. Any additional valid predictions in this space are likely to yield a higher impact than for those GO terms that are already well annotated. Even with this reduction in precision, the ITSS algorithm provides significantly more predictions over a broader number of GO terms than previously evaluated methods.

4.1 Predictions for the *H.sapiens* dataset

When compared to the two controls, the results of both the 10-fold cross-validation and historical validation in the *H.sapiens* datasets confirm that the integration of KNN and information theoretic semantic similarity methodologies is a valuable technique for predicting new gene annotations. To our knowledge, this study provides the first example of the application of a prediction algorithm to GO annotations in *H.sapiens*. As expected from the validation experiments over yeast (SGD) and fly (FlyBase) data, the ITSS algorithm performs significantly better than either the RA or RD controls. In a historical rollback, which assumes that techniques similar to ITSS were not applied to the dataset over the period evaluated (Supplementary Fig. S2), the precision of the ITSS algorithm can be estimated between 43% and 58%, lower than the 90% estimate of the 10-fold cross validation. Moreover, this rollback experiment over the *H.sapiens* dataset illustrated that the task of predicting future gene annotations is significantly more difficult than calculating contemporary ones. These results suggest that the 10-fold cross-validation overestimates the accuracy of the ITSS algorithm and that future studies should also include evaluations with historical rollback. The manual assessment we conducted led to a conservative estimate of 51% precision on predictions over a large and sparsely annotated network spanning 9589 human genes and 1377 GO terms, many of which were annotated with less than 10 genes. This compares favorably to previous manual assessment of predictions conducted in more favorable conditions. For example, King *et al*. (2003) observed 38% and 44% precision for predictions conducted on small, densely annotated subsets of SGD and FlyBase, respectively. It is notable that this subset contained only GO terms with 10 and more gene annotations, perhaps indicating poorer or equal performance of their predictive system in comparison to the ITSS method under optimal, densely annotated network conditions (King *et al*., 2003). A comparison of the 10-fold cross-validation results conducted on the more recent GOAr and the older GOAh files found no obvious differences in prediction results, indicating that the discrepancy observed in the historical validation was not due to intrinsic structural problems with the gold standard GOAr dataset. A reasonable explanation for the higher accuracy observed in the 10-fold cross-over designs is related to the high likelihood of functional codiscovery of related genes in genomic research (Rzhetsky *et al*., 2006) clustering them both functionally and temporally. Therefore, the GO Annotations are more likely to be updated in terms of functionally related gene groups during the same time period. With this in mind, the discrepancy is likely due to the similarity of functionally related genes in terms of annotation, making them good predictors for each other, and the propensity of the 10-fold cross-validation method to randomly split sets in such a way that it is likely to choose a gene within a specific functional group as a candidate for prediction. Conversely, predicting new annotations using historical data is likely to be more difficult because those annotations that can be easily inferred may have already been added to the GOA files during the same time period, and fewer patterns for predicting new annotations exist once these times periods are removed in the rollback. In addition, the benchmarks associated with the historical validation are often minimal or incomplete estimations. Considering the GOAh and GOAr files only differ by only 2 years of data, some of the FP found in historical validation may be borne out by future studies, increasing the observed precision and recall rates over time. As such, current precision and recall results for the historical validation can be interpreted as conservative minimum estimates. Thus a combination of cross-validation and historical rollback methods will provide a more comprehensive evaluation protocol for prediction algorithms in the future. As this is the first validation of its kind over the GO Annotations, it is still unknown if other machine learning approaches will also experience similar variance between the historical validation and the 10-fold cross-validation.

It is worth noting that the impact of applying semantic similarity metrics to these two types of validations is dichotomous. The extension of the threshold t to include non-identical concepts (e.g. $t = 0.7$) improves the historical validation results by up to 12.7% (see Supplementary Fig. S2). This is not the case for cross-validation methods, where the

optimal value in 10-fold cross-validation is $t = 1$, meaning that implicit hierarchical knowledge contained in the ontologies are not utilized to infer concept relations. This indicates that the GOA files contain patterns that are sufficient for conducting validation studies based on known annotations. However, for validation studies based on historical data more closely reflecting realistic conditions, the threshold t must be lowered to include more ontological knowledge in order to relate two different concepts. This demonstrates that superficial patterns based only on identical concepts are insufficient for predicting new gene annotations in a realistic setting, and the semantic relations between ontologically structured concepts must be used. As evaluations of other algorithms that use superficial patterns, which are subsequently validated using historical data have not been reported, we cannot perform an explicit comparison between the performance of ITSS and other algorithms in a historical validation.

The manual assessment results show that the precision of the ITSS algorithm could be increased further since many predicted annotations are semantically compatible to true knowledge, and will be judged as correct. The expert judged some predictions to be correct based on the semantic knowledge about the predicted GO concepts. For instance, the ITSS method predicted that the MTERF gene has the function DNA binding (GO:0003677) but, in the more recent GOA file, the gene was annotated with the term double-stranded DNA binding (GO:0003690). Therefore, the expert was able to determine the prediction to be correct based on semantic knowledge contained within the GO, because DNA binding is an ancestor concept of double-stranded DNA binding.

Because the ITSS method is entirely reliant on the known curated annotations of a gene in GOA, it is dependent on the timeliness of those annotations. However, in many cases corroborating evidence for a particular annotation exists in the literature for a significant amount of time before actually being added to the corresponding GOA file. This annotation lag is illustrated in our manual evaluation, where most of the evidence utilized to corroborate predictions are dated prior to the GOAh release date (2003). For example, the evidence that the gene GP6 (glycoprotein VI) has a 'receptor activity' was published in 2000 (Ezumi *et al*., 2000). However, the annotation 'receptor activity' for GP6 was not yet added to the GOA files as of 2003 (GOAh), but appeared in the GOAr file dated 2005. Therefore, by applying the ITSS algorithm to the GOAh file, the association between 'receptor activity' and GP6 was predicted as novel because in 2002 the fact was not annotated in the GOAh file, though the publication was otherwise available since 2000. While the method is not able to make predictions for completely un-annotated genes, the results of the manual validation indicate that the ITSS method may help experts find annotation omissions, and keep much of the associated 'computer executable knowledge' up-to-date.

4.2 Future work

The ITSS method and results were comparable to other machine learning algorithms in 10-fold cross-over designs and provided better future predictions than these techniques over a broader number of genes and GO terms when comparing the manual curations. Therefore, the possibility that in situations closely resembling real life this approach could outperform those based on superficial annotation patterns merits future study.

5 CONCLUSIONS AND FUTURE WORK

In this study we demonstrate the efficacy of ITSS, a high throughput computational approach capable of automatically predicting GOA with equal or higher overall accuracy than previous methods for a significantly broader range of GO terms. The ITSS prediction approach is able to accurately provide predictions for sparsely annotated gene functions and processes where previous methods were not demonstrated to work, generating an order of magnitude more predictions in GOA as a result. In contrast to other machine learning methods that provide a prediction giving no justification or line of reasoning behind the predictive process, the proposed similarity-based algorithms are readily interpretable: GOA contributing to the 'similarity scores' and gene deemed similar contributing to the 'KNN vote' can be straightforwardly verified. As a result, prediction reliability can be easily judged by investigating similar genes. To our knowledge, this is the first study demonstrating the feasibility of using the semantic similarity-based algorithm for the prediction of GO annotations. The novel prediction method has been shown to faithfully recapitulate known 'future' biological knowledge artificially removed from the dataset through a conservative historical rollback validation. In addition, we conducted an in-depth evaluation demonstrating the higher level of difficulty involved in predicting future GO annotations using a rollback method as compared to a conventional 10-fold cross-over validation with contemporary annotations removed. This method holds promise in facilitating a high throughput approach to generating hypotheses in genomic and biomedical research and it is likely to be applicable to other networks of annotations as well.

ACKNOWLEDGEMENTS

The authors thank D. Maglott for providing GOAh, Dr X. Li for her expert evaluation, and T. Borlawsky for editorial assistance. This study is partially supported by grants K22LM008308, R01LM007659, R01LM008635 and 1U54CA121852.

Conflict of Interest: none declared.

REFERENCES

Andrade,M.A. and Valencia,A. (1998) Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families. *Bioinformatics*, **14**, 600–607.

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

Bada,M. *et al*. (2004) *Using Reasoning to Guide Annotation with Gene Ontology Terms in GOAT*. ACM Press, New York, NY, USA, pp. 27–32.

Chen,J.L. *et al*. (2007) Evaluation of high-throughput functional categorization of human disease genes. *BMC Bioinformat*., **8**, S7.

Chen,Y. and Xu,D. (2004) Global protein function annotation through mining genome-scale data in yeast Saccharomyces cerevisiae. *Nucleic Acids Res*., **32**, 6414–6424.

Cherry,J.M. *et al*. (1998) SGD: saccharomyces genome database. *Nucleic Acids Res*., **26**, 73–79.

Chiang,J.H. *et al*. (2006) GeneLibrarian: an effective gene-information summarization and visualization system. *BMC Bioinformat*., **7**, 392.

Duda,R.O. and Hart,P.E. (1973) *Pattern Classification and Scene Analysis*. Wiley, New York.

Ezumi,Y. *et al*. (2000) Molecular cloning, genomic structure, chromosomal localization, and alternative splice forms of the platelet collagen receptor glycoprotein VI. *Biochem. Biophys. Res. Commun*., **277**, 27–36.

Hanley,J.A. and McNeil,B.J. (1983) A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, **148**, 839–843.

Haridas,V. *et al*. (1999) VEGI, a new member of the TNF family activates nuclear factor-kappa B and c-Jun N-terminal kinase and modulates cell growth. *Oncogene*, **18**, 6496–6504.

Jiang,J. and Conrath,D. (1999) Multi-word complex concept retrieval via lexical semantic similarity. In *Proceedings 1999 International Conference on Information Intelligence and Systems*. IEEE, pp. 407–414.

Jiang,J.J. and Conrath,D. (1997) semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of International Conference on Research in Computational Linguistics Research on Computational Linguistics (ROCLING X)*. Academica Sinica, Taipei, Taiwan, pp. 19–33.

Jones,C.E. *et al*. (2005) Automated methods of predicting the function of biological sequences using GO and BLAST. *BMC Bioinformat*, **6**, 272.

Kemmeren,P. *et al*. (2005) Predicting gene function through systematic analysis and quality assessment of high-throughput data. *Bioinformatics*, **21**, 1644–1652.

Khan,S. *et al*. (2003) GoFigure: automated gene ontologyTM annotation. *Bioinformatics*, **19**, 2484–2485.

King,O.D. *et al*. (2003) Predicting gene function from patterns of annotation. *Genome Res*., **13**, 896–904.

Laskowski,R.A. *et al*. (2005) Protein function prediction using local 3D templates. *J. Mol. Biol*., **351**, 614–626.

Lee,J.H. *et al*. (1993) Using term dependencies of a thesaurus in the fuzzy set model. *Microproc. Microprog*., **39**, 105–108.

Lin,D. (1998) An information-theoretic definition of similarity. Machine learning. In *Proceedings of the Fifteenth International Conference (ICML'98)*. Morgan Kaufmann Publishers, Madison, WI, USA, pp. 296–304.

Lord,P.W. *et al*. (2003a) Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation. *Bioinformatics*, **19**, 1275–1283.

Lord,P.W. *et al*. (2003b) Semantic similarity measures as tools for exploring the gene ontology. *Pac. Symp. Biocomput*., 601–612.

Metz,C.E. (1978) Basic principles of ROC analysis. *Semin. Nucl. Med*., **8**, 283–298.

Mitchell,J.A. *et al*. (2003) From phenotype to genotype: issues in navigating the available information resources. *Methods Inf. Med*., **42**, 557–563.

Perez,A.J. *et al*. (2004) Gene annotation from scientific literature using mappings between keyword systems. *Bioinformatics*, **20**, 2084–2091.

Resnik,P. (1995) Using information content to evaluate semantic similarity in a taxonomy. In Mellish,C.S. (ed.) *IJCAT-95 Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*. Morgan Kaufmann Publishers, Elsevier, Amsterdam, The Netherlands, pp. 448–453.

Rzhetsky,A. *et al*. (2006) Microparadigms: chains of collective reasoning in publications about molecular interactions. *Proc. Natl Acad. Sci. USA*, **103**, 4940–4945.

Shahbaba,B. and Neal,R.M. (2006) Gene function classification using Bayesian models with hierarchy-based priors. *BMC Bioinformat*., **7**, 448.

Sokal,R.R. and Rohlf,F.J. (1995) *Biometry: The Principles and Practice of Statistics in Biological Research*. W.H. Freeman and company, Elsevier, Amsterdam, The Netherlands.

Tuikkala,J. *et al*. (2006) Improving missing value estimation in microarray data with gene ontology. *Bioinformatics*, **22**, 566–572.

Wang,H. *et al*. (2005) An ontology-driven clustering method for supporting gene expression analysis. In *Proceedings of the 18th IEEE International Symposium on Computer-Based Medical Systems*. Dublin, Ireland, pp. 389–394.

Wang,H. *et al*. (2004) Gene expression correlation and gene ontology-based similarity: an assessment of quantitative relationships. In *Proceedings of IEEE*. La Jolla, CA, USA, pp. 25–31.

Wu,X. *et al*. (2006) Prediction of yeast protein-protein interaction network: insights from the gene ontology and annotations. *Nucleic Acids Res*., **34**, 2137–2150.

Yuseop,K. *et al*. (2001) Collocation dictionary optimization using WordNet and k-nearest neighbor learning. *Mach. Trans*., **16**, 89–108.