

Improving Feature Representation of Natural Language Gene Functional Annotations Using Automatic Term Expansion

Ji He

Abstract—Despite increasing work for describing gene functions using controlled vocabulary, natural language style gene functional annotations are most easily available and are most widely used by biologists. And intelligent analysis of these data in large scale is of great importance in the post-genome era. While the vector space model (VSM) based TF*IDF feature representation is widely adopted for text document analysis, it has significant limitations when applied to these data, primarily due to the high conciseness and high noisiness of the functional annotations. To improve TF*IDF feature representation, this paper proposes two automatic term expansion (ATE) methods based on query expansion (QE) in information retrieval (IR) theory. The effectiveness of ATE was examined through its application to the measurement of pattern proximity of gene functional annotations. Our comparative results show that ATE is effective in retrieving functionally correlated genes corresponding to a random query gene on this particular data type, and has the capability to produce more accurate measurement of the pattern similarity, with reference to genes' biological functions.

I. INTRODUCTION

GENE function is one of the central topics in biology and bioinformatics [1]. In the literature, a large number of computational methods exist for predicting and validating gene functions based on evidences from various aspects such as sequence similarity, phylogenetic profiles, gene co-regulation patterns, protein-protein interactions, and protein complexes [2]. Regardless of the approach used to determine the functions of genes, informatics researchers eventually face the question of how to describe gene functions in a way so that they can be easily accessible and understandable by both human and machine.

Presently, biologists commonly annotate gene functions using concise terms in natural language. Particularly, almost every protein in the NCBI non-redundant protein database (NR) is associated with this type of functional description. Natural language however is well known to be challenging to machine understanding. With an attempt to improve this situation, some ongoing work such as the Gene Ontology (GO) [3] and the KEGG Orthology (KO) [4] promote the use of controlled vocabulary to facilitate computational analysis, and are gaining increasing popularity. However, it is laborious and time consuming to establish a comprehensive ontology dictionary. So far, no ontology system claims to have a full coverage of biological concepts. In addition, mapping confirmed biological functions to the ontology space

is shown to be tedious [5], despite some existing methods for automatic assignment of ontology terms based on either biological [6][7] or text literature data [5]. These two factors primarily limit the coverage of ontology-based functional annotations. As a comparison, presently, the well-known ontology-annotated GO sequence database contains about 190,000 reference sequences, which is considerably smaller than the natural language-annotated NCBI NR database that has more than 6,000,000 reference sequences. In numerous laboratories, it is still a *de facto* standard for biologists to carry out BLAST searches [8] of their first-hand sequence data against the NCBI NR database, and consequently annotate them in natural language. Not surprisingly, one may foresee that for a considerable period of time, natural language functional annotations will remain most easily available to biologists.

This situation however poses a significant challenge to the analysis of gene functions. For example, given a new sequence library, very often it is desired to identify all genes that are associated with some particular biological functions, or to quickly survey its distribution according to predefined categories. Manual keyword search and organization of these data have shown to be tedious and inefficient, especially in view of the data explosion in the post-genome era. As such, it is desired to have an automatic system for more intelligent analysis of these data.

It well known that the measurement of pattern proximity plays a crucial role in computational intelligence research, especially in fields like information retrieval, pattern recognition, clustering analysis, and reasoning [9][10]. Particularly, in our application, the evaluation of the correlation between two biological terms and further two functional annotations primarily affects the performance of automatic gene selection and categorization processes. As such, this paper focuses on the measurement of pattern proximity for natural language functional annotations as the start point.

The rest of this paper is organized as follows: Section II briefly reviews a typical paradigm for measuring text proximity based on term frequency - inverse document frequency (TF*IDF) feature representation in the vector space model (VSM), analyzes its limitation on gene functional annotation data, and proposes our improvements using automatic term expansion (ATE). Section III evaluates the efficacy of the proposed methods through comparative experiments. And lastly, Section IV summarizes our findings and proposes future work.

This work was supported by the Samuel Roberts Noble Foundation base fund.

Ji He is with the Bioinformatics Core Facility, Samuel Roberts Noble Foundation, Ardmore, OK 73401, USA (phone: 580-224-6727; fax: 580-224-6692; email: jhe@noble.org).

II. METHODS

We started our work on the foundation of the well-established information retrieval (IR) theory. Intuitively, one may treat the functional description of each gene as a short text document and hence consider our study as a content-based text mining application. As a typical approach, by extracting the contents of the text documents and embed them as data points into a vector space, we may further apply an appropriate numeric function to evaluate the proximity between two arbitrary vectors. This approach is briefly reviewed below.

A. Vector Space-based Text Feature Representation and Pattern Proximity: A Brief Review

The construction of a vector space starts with the selection of a set of *feature terms* (text tokens or words) according to statistical and/or heuristical criteria (a particular approach used in our study is summarized in Section III-A). This selected feature set, denoted as $T = \{t_1, t_2, \dots, t_M\}$, essentially serves as the bases of the M-dimensional vector space so that the content of each document may be represented as a weighted histogram in reference to the feature set. Given a text document X , one of the most widely adapted practice to abstract its content into a vector format (referred to as the document's *feature vector*) is the TF*IDF representation [11], formulated as

$$\begin{aligned} X &= (x_1, x_2, \dots, x_M) \\ &= (tf_1 \cdot idf_1, tf_2 \cdot idf_2, \dots, tf_M \cdot idf_M), \end{aligned} \quad (1)$$

where tf_i is the term frequency (TF) (i.e. the occurrence) of feature term t_i in document X , and idf_i is the inverse document frequency (IDF) of t_i in reference to the whole dataset, defined as

$$idf_i = -\log \frac{df_i}{N}, \quad (2)$$

in which df_i is the document frequency (DF) of t_i , i.e. the number of documents that contain t_i , and N is the total number of documents in the dataset. In addition, in many studies that involve pattern proximity, it is desired that the feature vectors be appropriately normalized in order to prevent category proliferation [12]. Among the most widely used normalization methods are the first-level ratio-scaling normalization given by

$$x'_i = \frac{x_i}{\max\{x_i, i = 1, \dots, M\}} \quad (3)$$

and the second-level Euclidean normalization given by

$$x'_i = \frac{x_i}{\sqrt{\sum_{i=1}^M x_i^2}}. \quad (4)$$

Once the contents of text documents are represented as data points in the vector space, one may evaluate the proximity among two arbitrary vectors X and Y using a meaningful

similarity or dissimilarity/distance measure. Among the most widely used measures are the Euclidean distance defined as

$$D(X, Y) = \sqrt{\sum_{i=1}^M (x_i - y_i)^2}, \quad (5)$$

and the cosine similarity (also know as the correlation around zero) given by

$$\cos(X, Y) = \frac{\sum_{i=1}^M (x_i \cdot y_i)}{\sqrt{\sum_{i=1}^M x_i^2} \cdot \sqrt{\sum_{i=1}^M y_i^2}}. \quad (6)$$

TF*IDF feature representation and VSM-based pattern proximity measures have solid theoretical foundation and have shown good performance in numerous real-life applications, such as large-scale text document clustering and classification [13][14]. However, they did not produce satisfactory results in our earlier studies on gene functional annotations. Our further investigation identified two characteristics of gene functional annotations that mainly caused this performance degradation, namely: 1) *Annotations are considerably short and concise*. In many cases, the actual function of a gene is inferred with just a few number of terms with low term frequencies. When converted into the vector space, they do not stand out to actually represent the feature of the text. In other words, there is a high level of noisiness in the TF*IDF feature vectors. 2) *Many gene functions appear rather subjectively annotated*. The usage of synonyms and polysemes in these annotations is very common. In addition, there are a large number of terms referring slight variations of same concept (e.g. "beta-glucosidase" and "glycoside") and terms referring different aspects of same or closely related biological processes (e.g. "transmembrane" and "transporter"). These terms cannot be collated together with a typical stemming procedure (the stemming algorithm we used is reported in Section III-A), and are consequently treated independently in the vector space. As a result, in our studies, the TF*IDF feature vectors were often high-dimensional, sparse and noisy. Very often, the annotation of a gene showed zero similarity to those of a large number of genes according to Equation 6, because they did not have terms exactly in common, although they appeared highly correlated through human inspection. This in turn affected the performance of down-stream analysis.

Our studies aim to solve these problems respectively through two aspects: Firstly, we investigate techniques as improvements over the conventional TF*IDF feature representation and particularly the IDF-based term weighting method, so that terms which better reflect genes' biological functions would receive relatively higher weights, and hence the noisiness caused by irrelevant terms would be correspondingly suppressed. And secondly, instead of assuming a complete independence among the feature terms, we study their inherit correlations, and take these information into account for a more accurate estimation of pattern proximity. The following section reports our proposed approaches in detail.

B. Automatic Term Expansion for Text Feature Representation

Our idea of automatic term expansion (ATE) originates from the well studied query expansion (QE) theory in IR. Given a user query and a number of initial retrieval results, QE refers to the process of interactively and/or automatically reformulating user’s search query to improve retrieval performance in terms of either precision, recall, or both [15]. To apply the concept of QE to our application, we consider every gene functional description as an individual query and its retrieval results in reference to a full dataset are ranked according to a pattern proximity measure (e.g. given by either Equation 5 or Equation 6). Then by applying an appropriate automatic QE algorithm, one may polish the query through modifying its feature vector and identify other genes with more relevant biological functions – in other words, improve the pattern proximity measurement. Following this concept, we proposed two automatic term expansion methods based on literal and conceptual term co-occurrence, respectively introduced below.

1) *ATE using literal term co-occurrence statistics:* This method is derived from the idea of Mitra et. al. [15] which assumes that two closely correlated terms will occur together in many documents. On the other hand, if two terms refer to different concepts, their co-occurrence will not be strongly correlated. In Mitra et. al.’s work, terms with high correlations to those in existing retrieval results were down-weighted in order to maximize the variety of future results. As a different approach, our study aim to expand the correlated term set in order to reduce the sparseness of the vector space, and to up-weight them in order to improve the representation of the original query. Specifically, the *term-to-term correlation* between t_i and t_j based on their *literal co-occurrence* in the whole dataset S , notated as $\text{Cor}(t_i, t_j)$ is calculated by

$$\text{Cor}(t_i, t_j) = \min\{P(t_i|t_j), P(t_j|t_i)\}, \quad (7)$$

where the posterior probability $P(t_i|t_j)$ is estimated as

$$P(t_i|t_j) = \frac{\# \text{ of documents in } S \text{ containing both } t_i \text{ and } t_j}{\# \text{ of documents in } S \text{ containing } t_j}. \quad (8)$$

Then given a gene’s feature vector in Equation 1 in respect to the feature term set $T = \{t_1, t_2, \dots, t_M\}$, each attribute value x_i is reformulated as

$$\begin{aligned} x'_i &= \sum_{j=1}^M U(t_i, t_j) \cdot \text{Cor}(t_i, t_j) \cdot x_j \\ &= \sum_{j=1}^M U(t_i, t_j) \cdot \text{Cor}(t_i, t_j) \cdot t_{f_j} \cdot idf_j, \end{aligned} \quad (9)$$

where $U(t_i, t_j)$ is a *utilization switch* in response to the term-to-term correlation between t_i and t_j , defined as

$$U(t_i, t_j) = \begin{cases} 1 & \text{if } \text{Cor}(t_i, t_j) \geq \alpha \\ 0 & \text{otherwise,} \end{cases} \quad (10)$$

where $\alpha \in [0, 1]$ is a positive threshold value.

Since $\text{Cor}(t_i, t_i) = 1.0$, Equation 9 essentially up-weighs all terms. The degree of an up-weight depends on the number of terms that are highly correlated to it and the levels

of their term-to-term correlations. Understandably, general terms have low correlations to all other terms according to Equation 8, due to their high TF. On the other hand, specialized terms (which usually infer particular biological functions in our application) that frequently co-occur in same functional annotations will be highly up-weighted. Hence, this process practically boosts “featured” terms and reduces the noisiness caused by general terms during feature representation.

Readers may note that the concept of ATE is particularly reflected on those terms with zero attribute values in the original TF vector, i.e. that do not appear in the original document. Given a term t_i with its original attribute value $x_i = 0$, if there exist at least one other term t_j that satisfies $x_j \neq 0$ and $\text{Cor}(t_i, t_j) \geq \alpha$, then in the reformulated feature vector, $x'_i \neq 0$ according to Equation 9. In other words, the “bag of words” originally collected from the document is expanded. Understandably, the degree of ATE is primarily controlled by the parameter α , as only terms whose pair-wise correlations are above the threshold are used for ATE.

Lastly, it is also understandable that the reformulated feature vector also requires necessary normalization (e.g. using Equation 3 or Equation 4) to reduce category proliferation.

2) *ATE using conceptual term co-occurrence statistics based on GO annotations:* As a natural extension to our proposed literal co-occurrence-based ATE approach, we also consider the possibility of utilizing other sources of gene functional annotations to assist our analysis of the natural language annotations. Understandably, if there is a reliable grouping of genes according to their actual biological functions, one may infer the correlations of the terms in their functional annotations also based on their co-occurrence in reference to this grouping, and further apply them to ATE. We refer to this type of statistics as *conceptual co-occurrence* in comparison to the literal co-occurrence mentioned above, as two terms are considered co-occurred as long as they appear in the same conceptually functional group.

The particular paradigm we reported here utilizes the data from gene ontology (GO) [3], being one of the major approaches for gene functional annotation in the conceptual space. Presently, GO represents over 24,000 controlled vocabulary (*GO terms*) in a directed acyclic graph (DAG) covering three orthogonal taxonomies, namely molecular function, biological process, and cellular component. Our studies primarily focus on the biological process taxonomy as it is most comprehensive among the three for the time being and is widely considered to be most related to genes’ biological functions. The hierarchical organization of the taxonomy allows us to estimate the correlations between two GO terms using well-established graph- and/or IR-based models.

Given a set of genes S , each member X being associated (annotated) with a set of GO terms $G(X)$ and a set of natural language terms $T(X)$, whose unions over the full gene set are T and G correspondingly, we say a GO term g is associated with a natural language term t if $\exists X, g \in G(X)$ and $t \in$

$T(X)$ – in plain words, there exist at least one genes which are annotated with both g and t . An intuitive way to estimate the conceptual co-occurrence of two natural language terms t_i and t_j could be to apply the same method in Equation 7, with a straight-forward modification on the estimation of posterior probability $P(t_i|t_j)$ so that,

$$P(t_i|t_j) = \frac{\# \text{ of GO terms associated with both } t_i \text{ and } t_j}{\# \text{ of GO terms associated with } t_j}. \quad (11)$$

This function however contains a significant drawback by treating all GO terms equally. Given the hierarchical organization of GO, one would naturally understand that genes grouped under a relatively specialized GO term would show higher correlation than those grouped under a more general GO term. With this consideration, we introduce a custom weighing to all GO terms in Equation 11 corresponding to their levels of specialization, more specifically, based on the widely adopted *information content* (IC) measurement [16].

Briefly saying, the information content of a lexical concept/class c is quantified as the negated log of its likelihood $p(c)$ in the corpus, formalized as

$$I(c) \equiv -\log(p(c)) = -\log\left(\frac{f(c)}{N}\right), \quad (12)$$

where $f(c)$ is the frequency of the instances of concept c and N is the corpus size¹. To apply IC measurement to GO, we treat each GO term as a conceptual class that subsumes the term itself as well as all its descendent (children) terms. Hence the likelihood on a GO term g is calculated according to

$$p(g) = \frac{\text{size_of}\{C(g)\}}{\text{size_of}\{C(\text{root})\}}, \quad (13)$$

where $C(g)$ is the set of terms being subsumed by g , and root is the most top level (root) term. With this definition, the more specialized a GO term g is, the lower the likelihood $p(g)$ is, and hence the higher information content $I(g)$ it has. Particularly, the information content of the root term has the lowest IC value 0.0. With this weighing, we further reformulate Equation 11 into

$$P(t_i|t_j) = \frac{\sum_{g_k \in G(i,j)} I(g_k)}{\sum_{g_k \in G(j)} I(g_k)}, \quad (14)$$

where $G(i, j)$ refers to the set of GO terms that are associated with both t_i and t_j , whereas $G(j)$ associated with t_j . By replacing Equation 8 with Equation 14 and applying it back into Equation 7 and Equation 9, we complete the formalization of the ATE process based on conceptual co-occurrence.

III. COMPARATIVE EXPERIMENTS AND RESULTS

Our experiments examined the effectiveness of the proposed ATE methods on a real-life dataset, the Affymetrix 22k ATH1 Arabidopsis genechip, by comparing them against the conventional TF*IDF feature representation method without

¹Readers may also note that the definition of IDF in Equation 2 is a particular application of IC.

ATE, with reference to pattern proximity estimation on the GO conceptual space. Our comparative paradigm and results are reported below.

A. Dataset and Preprocessing

The Affymetrix 22k ATH1 Arabidopsis genechip is a widely-used microarray genechip for genome-wide transcriptome study of the model plant *Arabidopsis thaliana*. It contains 22,500 probe-sets representing approximately 24,000 genes. The full annotations of all probe-sets were downloaded through the NASCArrays web server². Since our interests were in the gene-level annotations in both natural language style and in GO, we firstly removed all control probe-sets, cross-hybridization probe-sets (each of which present multiple genes) and probe-sets with no confirmed gene-mapping from the dataset. This process generated a dataset with one-to-one mapping between probe-sets and genes. Further, we manually inspected all gene annotations and removed those probe-sets annotated with unknown gene functions as they did not provide valuable information in our validation. In addition, in our experiments, we used only the biological process taxonomy of GO. And thus, genes with no confirmed biological process GO annotation were also removed. Lastly, the natural language style functional annotations of all remaining genes were abstracted into vector format following the preprocess described following:

The preprocess started with the selection of text features in order to construct the vector space. In the literature, N-grams [17] and bag of words (BOW) [18] are two typical approaches for this task. N-gram extracts all possible N-length substrings of a text string, whereas BOW collects naturally delimited words disregarding grammar and order information. Our study adopted the BOW approach as it retains the readability of the words. A large number of statistical and heuristic methods have been studied in the history for keyword selection (examples including document frequency, CHI-square statistics, information gain [19] and evolutionary search [20]). Since in our study there is no pre-defined categories for reference and considering the fact that most gene functional annotations are notably short, we adopted a relatively straight-forward document frequency-based filtering approach for feature selection. This filtering process is summarized below:

- 1) **Removing custom patterns:** We maintain a small list of custom “stop-patterns” such as species names in “[...]” format (e.g. “[Arabidopsis thaliana]”), gene and protein family IDs (e.g. “[InterPro:xxxxx]”) that we do not feel necessary to include into the feature set. Text that match these custom patterns based on regular expressional searches are firstly removed.
- 2) **Tokenizing text:** Words are extracted from text based on natural delimiters (white space, comma, full stop, etc.).

²NASCArrays is available at <http://affymetrix.arabidopsis.info/narrays/experimentbrowse.pl>.

- 3) **Stemming:** The root forms of all words are identified based on the Porter stemming algorithm [21]. Thus words with the same morphological origins (such as "transport", "transporting" and "transporter") are equally treated.
- 4) **Removing stopwords:** Words that appear in a maintained stoplist are then removed. A stoplist is dictionary of highly common words (stopwords) in reference to generic corpora (e.g. "I", "is", "very") that are widely considered insignificant in typical content-based IR studies.
- 5) **Removing too general and too rare words:** This filters words based on their document frequency (DF) with two thresholds F_{min} and F_{max} . Too general words (with $df > F_{max}$) and too rare words (with $df < F_{min}$) are thought insignificant in our study and are removed. Practically, we decide the DF thresholds based on human inspection and in reference to the total number of texts in the dataset.

In practice, our experiments used 32 stop-patterns, 617 stopwords and thresholds of $F_{min} = 12$ and $F_{max} = 1,200$, which in the end generated a list of 1,492 keywords.

Based on the selected keyword feature set, the natural language style gene annotations were converted into vector format according to Equation 1 and null vectors with all-zero attribute values were further removed, leaving a final dataset of 14,197 genes for our experiments ³.

B. Experimental Design and Evaluation Measures

Our experiments were designed to compare our proposed two ATE techniques, based on literal term co-occurrence as proposed in Section II-B1 (ATE-L for ease of notation) and conceptual term co-occurrence in Section II-B2 (ATE-C) respectively, over the conventional BOW-based TF*IDF feature representation. Our comparisons were two-fold. Firstly, we investigate if these ATE techniques may help to identify larger number of genes that are functionally correlated to a random query gene from the sparse dataset. Secondly, we evaluate if the measurement of the pattern proximity between two genes could be improved to better infer the correlation of their biological functions through ATE.

A challenge to our study is the lack of a "golden standard" for evaluating the degree of similarity (correlation) between two genes' biological functions. Presently there is no discriminative training and/or testing dataset that quantitatively notates how two genes' functions are correlated. It is also not practicable for us to manually verify the pair-wise correlations of all genes' functions in a genome-wide scale. As our work-around, we use similarity scores based on GO as the reference measurement, as the GO annotations for most Arabidopsis genes had been well curated and validated by human. Our evaluation of GO similarity was based on Resnik's information content (IC) approach [22] which has been reported to outperform some other measures in the literature [23]. The measurement is summarized as following:

³The dataset and all processing programs are available upon email request.

Given two gene products G_X and G_Y annotated with P and Q GO terms respectively, notated as $G_X = \{g_{x1}, g_{x2}, \dots, g_{xP}\}$ and $G_Y = \{g_{y1}, g_{y2}, \dots, g_{yQ}\}$, we first calculate the all pair-wise similarities between two GO terms g_{xi} and g_{yj} based on the IC of their *minimal subsumer*. A so-called minimal subsumer of two terms g_{xi} and g_{yj} , denoted as $\nabla(g_{xi}, g_{yj})$, is their subsumer that has the minimal likelihood p (and hence maximal IC). To formalize:

$$\begin{aligned} \text{Sim}(g_{xi}, g_{yj}) &\equiv \frac{I(\nabla(g_{xi}, g_{yj}))}{\max\{I(g_{xi}), I(g_{yj})\}} \\ &= \frac{-\log(\min\{p(g)|g \in S(g_{xi}, g_{yj})\})}{-\log(\max\{p(g_{xi}), p(g_{yj})\})}, \end{aligned} \quad (15)$$

where $I(\cdot)$ is given by Equation 12, $p(\cdot)$ is given by Equation 13, and $S(g_{xi}, g_{yj})$ is the subsumer set of term g_{xi} and g_{yj} , essentially being all their common ancestor terms. Understandably, there are $P \cdot Q$ pair-wise term-to-term similarities based on all GO annotation terms for gene products G_X and G_Y . And in order to evaluate the overall similarity between these two gene products $\text{Sim}(G_X, G_Y)$, we followed a straight-forward yet most common ordered weighted average (OWA) approach [24] by taking their arithmetic average:

$$\text{Sim}(G_X, G_Y) = \frac{\sum_{i=1}^P \sum_{j=1}^Q \text{Sim}(g_{xi}, g_{yj})}{P \cdot Q}. \quad (16)$$

Understandably, $\text{Sim}(G_X, G_Y) \in [0, 1]$, with a higher value indicating a higher similarity level between the biological functions of two gene products. Readers may note our Equation 16 consists of a normalization of Resnik's definition for the ease of comparison in the same scale.

Reaching this point, our experimental design is distinctly outlined: given a set of N genes $S = \{X_1, X_2, \dots, X_N\}$, each with both natural language annotation and GO annotation, for any gene X_i , we first calculate its pattern similarity to all other genes $X_j, j \neq i$ based on their GO annotations, this will produce a vector of $N - 1$ similarity values, denoted as $\vec{\text{Sim}}_0 = (\text{Sim}_0(X_i, X_1), \dots, \text{Sim}_0(X_i, X_j), j \neq i, \dots, \text{Sim}_0(X_i, X_N))$ to serve as the "golden standard". Then we calculate the same set of pair-wise similarities using the feature vectors generated by each feature representation method based on natural language annotations, denoted as $\vec{\text{Sim}} = (\text{Sim}(X_i, X_1), \dots, \text{Sim}(X_i, X_j), j \neq i, \dots, \text{Sim}(X_i, X_N))$. Understandably, the number of non-zero values Sim indicate the number of genes being identified to be with functional correlation to the query gene, which is recorded in our experiment and used for comparison. To evaluate the goodness of the correlation calculation, naturally, one may understand that a higher correlation between Sim and Sim_0 suggests a better performance. Specifically, we used the Pearson correlation coefficient to obtain a quantitative evaluation of this correlation. Without losing generalization, the Pearson correlation coefficient r between two K -dimensional vectors X and Y is defined as

$$r = \frac{\sum_{i=1}^K ((x_i - \bar{X}) \cdot (y_i - \bar{Y}))}{\sqrt{\sum_{i=1}^K (x_i - \bar{X})^2} \cdot \sqrt{\sum_{i=1}^K (y_i - \bar{Y})^2}}, \quad (17)$$

where \bar{X} is the mean of X given by $\bar{X} = (\sum_{i=1}^k x_i)/K$. Clearly, $r \in [-1, 1]$, with a higher value suggesting a higher correlation, and hence a better performance in our experiment.

C. Results and Discussions

We applied the cosine similarities of genes' functional annotations in natural language, based on earlier-mentioned three different feature representation methods, namely naive TF*IDF, ATE-L and ATE-C, and evaluated their performance based on twenty-fold test⁴. With twenty-fold test, the ATH1 dataset was randomly and evenly divided into 20 subsets (in practice, 17 subsets each with 710 data and 3 subsets each with 709 data). Each time, our program used nineteen subsets as the training set for inferring both literal and conceptual term-to-term correlations, then applied these statistics for ATE on a unique testing set. The separation of training set and testing set ensures that our evaluation based on GO annotation does not lead to biased results due to the use of GO for inferring conceptual co-occurrence statistics.

In each unique twenty-fold test, different term-to-term correlation threshold α in Equation 10 that ranged from 0.1 to 0.9 were tested. With each ATE method using each threshold on each gene as the query, the number of genes that had non-zero pattern similarity to the query gene and the correlation score defined in Equation `refeq:pearson-correlation` were recorded. The same evaluation was performed over TF*IDF without ATE on each testing set. Results from all twenty subsets were finally aggregated to form an overall comparison.

As shown in Figure 1 and Figure 2, for a random query gene, TF*IDF without ATE returned an average of 36.34 genes with non-zero pattern similarity, out of the testing subset in average size of 709.85. The average correlation between two random genes' pattern similarities calculated by TF*IDF and their biological functional correlations inferred with GO was 0.245. When the term-to-term correlation threshold α was relatively low, especially when $\alpha \leq 0.3$, the performance of both ATE-L and ATE-C are greatly affected by α . However, with relatively high α thresholds ($\alpha \geq 0.5$), both two methods become less sensitive to this parameter. In generally, a lower α value leads to a larger number of returned genes with non-zero pattern similarities to a random query gene, whereas the similarity calculation by both ATE-L and ATE-C tend to be less accurate (as indicated by smaller correlation scores compared to that of TF*IDF). This suggests that, in practice, using a too low α setting may lead to query over-expansion and in turn degrade the performance of ATE. On the other hand, and satisfactorily, with a reasonably medium-to-high α value range (particularly $\alpha \geq 0.4$ in our experiments), the pattern similarities generated by ATE-L are stably in the same level of those by TF*IDF, while the number of correlated genes discovered

⁴Our experiment used twenty-fold test instead of more traditional ten-fold test due to the high computational cost on pair-wise pattern proximity calculation with varying threshold values.

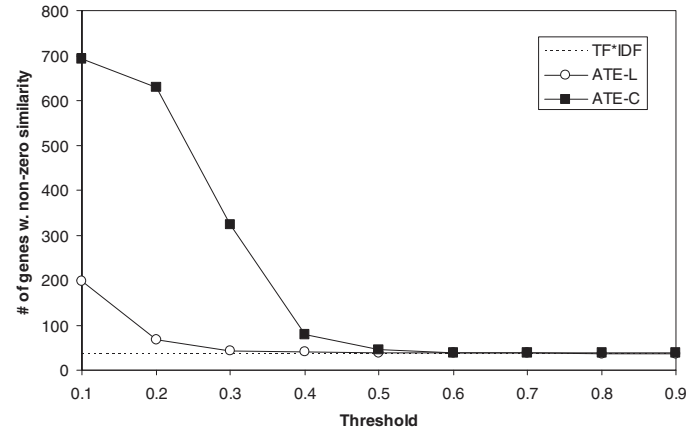


Fig. 1. The average numbers of genes with non-zero pattern similarity to a random query gene returned by ATE-L and ATE-C in response to different term-to-term correlation thresholds, compared to those returned by TF*IDF without ATE, with reference to the testing subsets with an average size of 709.85.

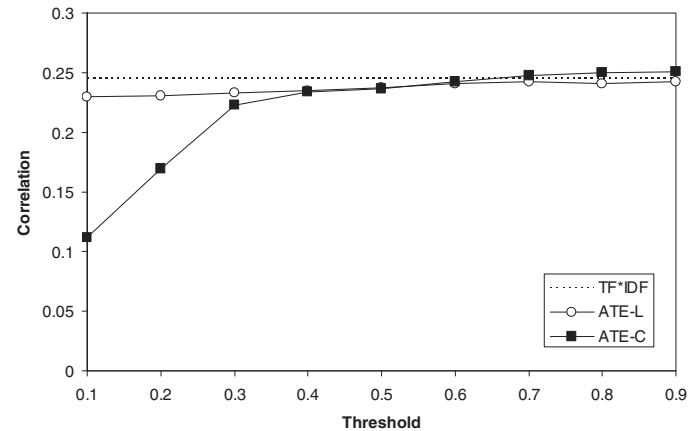


Fig. 2. The average correlations between the correlation of two random genes' biological functions inferred with GO and the pattern similarity calculated by ATE-L and ATE-C in response to different term-to-term correlation thresholds, compared to those of TF*IDF without ATE.

by ATE-L is constantly higher than that by TF*IDF. In other words, with a proper control of the degree of expansion, ATE based on literal co-occurrence could generally help to identify more genes of user interests based on a random search query. Furthermore, using the same α threshold in this range, ATE-C's performance appeared to be marginally better than ATE-L. More significantly, with a relatively high threshold $\alpha \geq 0.6$, ATE-C outperformed both ATE-L and TF*IDF in both evaluation aspects. This indicates that with an appropriate mapping between text terms and biological concepts, it is possible to better infer the correlations among text terms and further apply this knowledge to improve the analytical results over natural language text data.

IV. CONCLUDING REMARKS AND FUTURE WORK

In this paper, we analyzed the challenges of representing the feature of a common biological data type, gene functional annotations in natural language. We showed that the conventional content-based feature representation method in the IR

field had limited power on this particular problem, mainly due to the typically short and concise descriptions and high diversity and ambiguity of biological terms in the primary data. Based on the well-established query expansion (QE) theory, we proposed two automatic term expansion (ATE) methods based on literal and conceptual term co-occurrence statistics respectively. In our controlled comparative experiments on the Affymetrix 22k ATH1 Arabidopsis genechip dataset, both ATE methods have yield performance gains in varying degrees over the conventional feature representation method TF*IDF, through their applications to measuring the pattern similarities among genes, in reference to genes' biological functions. Our studies hence conclude that ATE is a powerful technique for improving the feature representation of natural language style gene functional annotations.

While we obtained satisfactory results in our controlled experiments, naturally, there are a number of questions to be answered in our future work. Firstly, it deserves to further study the possibility of incorporating both literature and conceptual co-occurrence statistics into a single framework that is more effective than ATE using either aspect of data. Secondly, it would be interesting to further evaluate if the performance of ATE could be further improved using knowledge from other data sources such as biological literature and WordNet lexical database [25]. Lastly but most importantly, not limited to the natural language gene functional annotation data, it would be most valuable to carry out research on knowledge integration from heterogenous domains to better understand gene functions in a more intelligent manner.

ACKNOWLEDGEMENTS

The author appreciates colleague Dr. Yongzhen Pang's assistance in preparing and organizing the Affymetrix Arabidopsis ATH1 genechip dataset. The author would like to thank all reviewers for their critic reading and valuable comments.

REFERENCES

- [1] T. Murali, C.-J. Wu, and S. Kasif, "The art of gene function prediction," *Nature Biotechnology*, vol. 24, no. 12, pp. 1474–1476, 2006.
- [2] A. Vazquez, A. Flammini, A. Maritan, and A. Vespignani, "Global protein function prediction from protein-protein interaction networks," *Nature Biotechnology*, vol. 21, pp. 697–700, 2003.
- [3] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene ontology: tool for the unification of biology. the gene ontology consortium." *Nature Genetics*, vol. 25, no. 1, pp. 25–29, 2000.
- [4] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa, "KEGG: Kyoto encyclopedia of genes and genomes," *Nucleic Acids Research*, vol. 27, no. 1, pp. 29–34, 1999.
- [5] E. Stoica and M. Hearst, "Predicting gene functions from text using a cross-species approach," in *Proceedings of the Pacific Symposium on Biocomputing (PSB)*, vol. 11, 2006, pp. 88–99.

- [6] Y. Tao, L. Sam, J. Li, C. Friedman, and Y. A. Lussier, "Information theory applied to the sparse gene ontology annotation network to predict novel gene function," *Bioinformatics*, vol. 23, no. 13, pp. i529–538, 2007.
- [7] X. Mao, T. Cai, J. G. Olyarchuk, and L. Wei, "Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary," *Bioinformatics*, vol. 21, no. 19, pp. 3787–3793, 2005.
- [8] S. F. Altschul, W. Gish, W. Miller, E. W. Meyers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, 1990.
- [9] H. Eidenberger, "Evaluation and analysis of similarity measures for content-based visual information retrieval," *Multimedia Systems*, vol. 12, no. 2, pp. 71–87, 2006.
- [10] S.-F. Ding, S.-X. Xia, F.-X. Jin, and Z.-Z. Shi, "Novel fuzzy information proximity measures," *Journal of Information Science*, vol. 33, no. 6, pp. 678–685, 2007.
- [11] S. Robertson, "Understanding inverse document frequency: On theoretical arguments for IDF," *Journal of Documentation*, vol. 60, no. 5, pp. 503–520, 2004.
- [12] G. Carpenter, S. Grossberg, and D. Rosen, "Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system," *Neural Networks*, vol. 4, pp. 759–771, 1991.
- [13] A. Rauber, E. Schweighofer, and D. Merkl, "Text classification and labelling of document clusters with self-organising maps," *Journal of the Austrian Society for Artificial Intelligence*, vol. 19, no. 3, pp. 17–23, October 2000.
- [14] Y. Yang and X. Liu, "A re-examination of text categorization methods," in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1999, pp. 42–49.
- [15] M. Mitra, A. Singhal, and C. Buckley, "Improving automatic query expansion," in *Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, 1998, pp. 206–214.
- [16] S. Ross, *A First Course in Probability*. Macmillan, 1976.
- [17] W. B. Cavnar and J. M. Trenkle, "N-gram-based text categorization," in *Proceedings of the Annual Symposium on Document Analysis and Information Retrieval (SDAIR)*, Las Vegas, US, 1994, pp. 161–175.
- [18] D. D. Lewis, "Naive (Bayes) at forty: The independence assumption in information retrieval." in *Proceedings of the European Conference on Machine Learning (ECML)*, no. 1398, 1998, pp. 4–15.
- [19] Y. Yang and J. Pedersen, "A comparative study on feature selection in text categorization," in *Proceedings of the Fourteenth International Conference on Machine Learning (ICML)*, 1997, pp. 412–420.
- [20] Y. Kim, W. Street, and F. Menczer, "Feature selection in unsupervised learning via evolutionary search," in *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 2000.
- [21] M. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130–137, 1980.
- [22] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 1995, pp. 448–453.
- [23] J. L. Sevilla, V. Segura, A. Podhorski, E. Guruceaga, J. M. Mato, L. A. Martinez-Cruz, F. J. Corrales, and A. Rubio, "Correlation between gene expression and GO semantic similarity," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 2, no. 4, pp. 330–338, 2005.
- [24] R. R. Yager, "On ordered weighted averaging aggregation operators in multi-criteria decision making," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 18, no. 183–190, 1988.
- [25] C. Fellbaum, *WordNet: An Electronic Lexical Database*. The MIT Press, 1998, ISBN 978-0262061971.