35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

GOseek: A Gene Ontology Search Engine using Enhanced Keywords

Kamal Taha, *Member, IEEE*

***Abstract*— We propose in this paper a biological search engine called GOseek, which overcomes the limitation of current gene similarity tools. Given a set of genes, GOseek returns the most significant genes that are semantically related to the given genes. These returned genes are usually annotated to one of the Lowest Common Ancestors (LCA) of the Gene Ontology (GO) terms annotating the given genes. Most genes have several annotation GO terms. Therefore, there may be more than one LCA for the GO terms annotating the given genes. The LCA annotating the genes that are most semantically related to the given gene is the one that receives the most aggregate semantic contribution from the GO terms annotating the given genes. To identify this LCA, GOseek quantifies the contribution of the GO terms annotating the given genes to the semantics of their LCAs. That is, it encodes the semantic contribution into a numeric format. GOseek uses microarray experiment data to rank result genes based on their significance. We evaluated GOseek experimentally and compared it with a comparable gene prediction tool. Results showed marked improvement over the tool.**

I. Introduction

The Gene Ontology (GO) [5] has emerged as one of the most important ontology and the most widely used bio-ontology. Many genomic databases [e.g., [2, 7, 11]) use GO annotations, which assign genes to term nodes to describe these genes. GO ontology is structured as a Directed Acyclic Graphs (DAG). In this graph, GO terms are represented by nodes and the different hierarchical relations between the terms *(mostly "is-a" and "part-of" relations)* are represented by edges. The "is-a" relation represents the fact that a given child term is a subtype of a parent term, and the "part-of" relation represents part-whole relationships. When a gene product is annotated using GO, the DAG displays the term(s) describing this gene product in such a way that reflects how this gene product is related to other gene products.

A number of tools have been developed to utilize the GO annotations stored in the genomic databases. Many of these tools are listed on the GO website [5]. These tools provide great help to biologists. However, most of them do not answer the following question that biologists often have: What is the set of genes that is semantically related to a given a set of genes. Biologists often need to know the set S' of genes that is *semantically related* to a given set S of genes. Determining the set S' helps in understanding gene-disease interactions and advanced disease diagnosis. For instance, biologists in the UAE are trying to determine the set of genes that are related to the genes involved in Type 2 Diabetes (T2D) *(one out of five people in the UAE between the ages of 20 to 79 lives with T2D)*. Few tools have the capability of retrieving the set S', such as DynGO [6], which "retrieves genes and gene products that are *relatives* of input genes based on similar GO annotations, and displays the related genes and gene products in an association tree" [6, 9]. However, most of these tools determine the semantic similarities among genes based solely on the proximity of the GO terms annotating these genes, while overlook the *structural dependencies* among these GO terms. This may lead to low *recall* and *precision* of results.

We propose in this paper a search engine called GOseek (**G**ene **O**ntology **S**earch **E**ngine using **E**nhanced **K**eywords). Given a set of genes, GOseek returns the most significant genes that are semantically related to the given genes. GOseek overcomes the limitation of current gene similarity tools outlined above as follows: (1) it employs the concept of *existence dependency* to determine the *structural dependencies* among the GO terms annotating a given set of genes, and (2) it encodes into a numeric format the contribution of these terms to the semantics of their Lowest Common Ancestor (LCA). The framework of GOseek defines semantic similarity measure as a function that returns a numerical value reflecting the closeness in meaning between the GO terms annotating a given set of genes and their LCA. GOseek accepts keyword-based queries with the form $Q($"g_1", "g_2", .., "g_n"), where g_i denotes a gene. The result of the query $Q($"g_1", "g_2"$)$ is a set of genes, where *each* gene in the set is semantically related to *both* g_1 and g_2. GOseek uses microarray experiment data to *rank* result genes based on their *significance*. We evaluated GOseek experimentally and compared it with a comparable gene prediction tool called DynGO [6]. Results showed marked improvement over the tool.

II. Assigning Semantic Weight to LCAs

Notation 2.1, Keyword Context (KC): A KC is a GO term annotated to a query gene product. For example, consider Fig. 1 and the query Q("Br"). The terms organ development (GO:0048513) and nephron morphogenesis (GO:0072028) are KCs because the gene "Br" is annotated with them.

Most genes have several annotation GO terms. If a query contains n gene keywords, there may be m KCs, where $m > n$. Therefore, a strategy is needed for determining the relationships among *all the occurrences* of genes under consideration (i.e., the relationships among all KCs).

K. Taha is with the Department of Electrical and Computer Engineering, Khalifa University of Science, Technology, and Research, Box 127788, Abu Dhabi, UAE. E-mail: kamal.taha@kustar.ac.ae.

GOseek selects from all KCs of a query subsets, where each subset contains the *smallest* number of KCs that: (1) are *meaningfully* related to each other, and (2) have at least one occurrence of each gene keyword annotated to the subset. The KCs in each subset are called Related Keyword Contexts (RKC). Consider for example Fig. 1 and the query Q("Br", "GCNTI"). Each of the genes "Br" and "GCNTI" is annotated to two GO terms. The RKC candidates are the sets: {GO:0048513, GO:0048729}, {GO:0048513, GO:0060993}, {GO:0072028, GO:0048729}, and {GO:0072028, GO:0060993}.

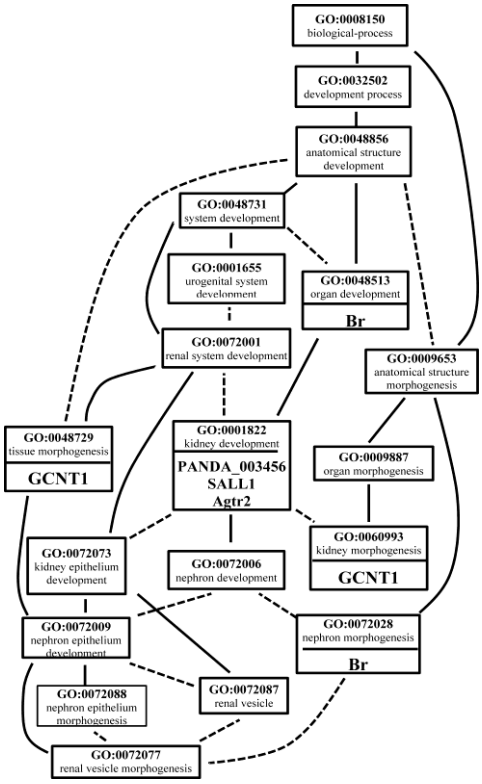

Fig. 1: A fragment of GO Graph showing the ontological relationships of 19 GO terms. Solid edges denote "is-a" relations and dotted edges denote "part-of" relations. Some terms show some of the genes annotated to them.

Since there are more than one RKC, there are more than one LCA of the RKCs. The genes that are most semantically related to input keyword genes are usually annotated to some of these LCAs. These LCAs are the ones that receive the most aggregate semantic contribution from their RKCs and from the terms located in the paths from the RKCs to the LCAs. To identify these LCAs, we quantify the contribution of RKCs and their ancestors to the semantics of the LCAs.

We define the semantic value of a LCA as the aggregate contribution of the terms located in the paths from RKC to the LCA. A KC closer to the LCA contributes more to the semantics of the LCA, while a KC farther from the LCA contributes less. We define the contribution of a GO term t with regard to a KC KC_u to the semantics of a LCA LCA_v as the Semantic Weight (SW) of t related to KC_u (denoted by $SW_{KC_u}(t)$). The SW of LCA_v related to KC_u, is defined as:

$$\begin{cases} SW_{KC_u}(KC_u) = decay^{depth} \\ SW_{KC_u}(t) = \max\{ e_c * SW_{KC_u}(t') \mid t' \in \text{childrenof}\ (t)\} \end{cases} \quad (1)$$

where e_c is the semantic contribution factor for the edge linking term t with its child term t' and $0 < e_c < 1$. We define the contribution of KC_u to LCA_v as $decay^{depth}$, where *decay* is a parameter that can be set in the range 0-1, and *depth* is the depth (hierarchical level) of KC_u, considering the depth of the root term is 0. The ancestors of KC_u contribute less, that is why we have $0 < e_c < 1$. The SW of LCA_v is the aggregate contribution of the semantics of all terms located in the path from each $KC \in RKC$ to LCA_v. Thus, we calculate the SW of LCA_v as:

$$SW(LCA_v) = SW_{KC_x}(LCA_v) + SW_{KC_y}(LCA_v) + \ldots + SW_{KC_z}(LCA_v) \quad (2)$$

where $SW_{KC_i}(LCA_v)$ is the aggregate contribution of the semantics of all terms located in the path from $KC_i \in RKC$ to LCA_v.

We use as a running example throughout the paper the keyword-based query $\boldsymbol{Q}$("Br", "GCNTI"). The query asks for the set of genes that are semantically related to *both* of the gene keywords "Br" and "GCNTI". As shown in Fig. 1: (1) the KCs annotating the gene "Br" are organ development (GO:0048513) and nephron morphogenesis (GO:0072028), and (2) the KCs annotating the gene "GCNTI" are tissue morphogenesis (GO:0048729) and kidney morphogenesis (GO:0060993).

Example 1: Consider the query Q("Br", "GCNTI") of our running example. Let us quantify the contribution of RKCs and their ancestors to the semantics of the LCAs. Consider that parameter *decay* in equation 1 is set to 0.5. Table 3 shows the value of $decay^{depth}$ for each KC. Consider that the semantic contribution factors for 'is-a' and 'part-of' relations are 0.7 and 0.6 respectively. Table 2 shows the SW value of each LCA in the example.

TABLE I. THE VALUES OF $decay^{depth}$ FOR EACH KC IN OUR RUNNING EXAMPLE

KC	GO:0072028	GO:0060993	GO:0048513	GO:0048729
$decay^{depth}$	0.25	0.125	0.125	0.125

TABLE 2. THE LCA OF EACH RKC AND THE SW VALUE OF EACH LCA IN OUR RUNNING EXAMPLE

RKC	GO:0072028 GO:0060993	GO:0048513G O:0060993	GO:0048729 GO:0072028	GO:0048729 GO:0048513
LCA	GO:0001822	GO:0048856	GO:0048856	GO:0048856
SW(LCA)	0.18	0.124	0.18	0.16

III. Determining the LCA Annotating the Most Significant Genes

The degree of association between a LCA and its RKC depends on the value of the SW of the LCA. LCAs with higher SW values have higher association with their RKCs. From the set of LCAs with high SW values, we need to identify the one annotating the most interesting (significant) genes. The genes that are most semantically related to input keyword genes are the ones annotated to the LCA, which annotates the most interesting "significant" genes and has a high SW value. The framework of GOseek selects the LCA with the greatest product of multiplying SW value by the number of significant genes annotated to the LCA. The framework returns the genes annotated to this LCA as the answer for the query. GOseek employs Fisher's exact test and gene expression microarray experiment [1] to find the number of significant genes that would be found by chance to be annotated to LCAs.

Consider that the total number of genes in microarray is "*a*". Consider that the result of the experiment revealed "c" significant genes. Consider that the number of genes annotated to a LCA LCA_v is "*b*". Consider that the number of genes of interest annotated to LCA_v is "*k*". The probability that the number of significant genes annotated to LCA_v is exactly "*k"* out of the "*c*" significant genes is given by the following Fisher's exact test:

$$\Pr(LCA_v) = 1 - \sum_{i=0}^{k-1} \frac{\binom{b}{i}\binom{a-b}{c-i}}{\binom{a}{c}} \qquad (3)$$

- "*a"*: the set of genes in microarray.
- "*b"*: the set of genes annotated to the LCA.
- "*c"*: the number of significant genes.
- "*k*": the number of genes of interest annotated to LCA_v.

The final score of LCA_v is the product of multiplying $SW(LCA_v)$ by $Pr(LCA_v)$ as shown in equation 4.

$$Score(LCA_v) = MAX(SW(LCA_v)) * \Pr(LCA_v) \qquad (4)$$

We multiply $SW(LCA_v)$ by $Pr(LCA_v)$ because each of the two measures evaluates different semantics and characteristics of LCA_v and we look for a LCA that has the greatest product of the two measures. If a LCA has more than one SW value, we take the maximum value.

Example 2: As shown in Table 2, there are two LCAs in our running example: GO:0048856 and GO:0001822. One of the two LCAs annotates the significant genes that are the most semantically related to the input gene keywords "Br" and "GCNTI". Using the microarray information in table 3, let us compute the final scores for the two LCAs. Table 4 shows the number of genes annotated to each of the two LCAs. We use equation 3 to compute the probability of significant genes as follows:

$$\Pr(GO:0048856) = 1 - \sum_{i=0}^{399} \frac{\binom{10153}{i}\binom{13373-10153}{553-i}}{\binom{13373}{553}} = 0.516$$

$$\Pr(GO:0001822) = 1 - \sum_{i=0}^{399} \frac{\binom{7621}{i}\binom{13373-7621}{553-i}}{\binom{13373}{553}} \approx 1$$

Table 5 shows: (1) the probability of significant genes annotated to each LCA, and (2) the final score of each LCA computed using equation 4. Since the final score of GO:0001822 is greater than those of GO:0048856, the genes annotated to GO:0001822 are returned as the answer for the query.

TABLE 3. The Gene Information of Mus Musculus Microarray "Affymetrix GeneChip Mouse Genome 430 2.0 Array" (GPL1261)

Number of genes	*Number of unique genes*	*Number of significant genes*	*Number of unique significant genes*
45851	13373	18307	553

TABLE 4. Number of Genes Annotated to the LCAs GO:0048856 and GO:0001822

GO:0048856		GO:0001822	
Number of genes	*Number of unique genes*	*Number of genes*	*Number of unique genes*
42780	10153	15488	7621

TABLE 5. The SW Value, Probability of Significant Genes, and Score of each LCA

RKC	GO:0072028 GO:0060993	GO:0048513 GO:0060993	GO:0048729 GO:0072028	GO:0048729 GO:0048513
LCA	GO:0001822	GO:0048856	GO:0048856	GO:0048856
SW(LCA)	0.18	0.124	0.18	0.16
Pr(LCA)	1	0.516	0.516	0.516
Score(LCA)	0.18	0.064	0.093	0.083

IV. Experimental results

We implemented GOseek in Java, run on Intel(R) Core(TM)2 Duo CPU processor, with a CPU of 2.6 GHz and 4 GB of RAM, under Windows 7. We experimentally evaluated the quality of GOseek and compared it with DynGO [6]. DynGO "retrieves genes and gene products that are *relatives* of input genes based on similar GO annotations, and displays the related genes and gene products in an association tree" [6].

A. Benchmarking Datasets

Pathways are sets of genes shown to have high functional similarity and can be used to validate similarity measures [4, 8]. A fully described pathway represents the dynamics and dependencies among a set of gene/gene products. Therefore, we used in our experiments pathways as a reference for evaluating and comparing the similarity measures of GOseek and [6]. Given a set S of genes belonging to a same pathway, each of the two methods should return another set S' of genes that is semantically related to set S. In order for sets S and S' to be semantically related, they should be part of the *same pathway*.

We used for the evaluation two different pathway benchmarks: KEGG and Pfam benchmarks. We selected a set of 15 human and 15 yeast diverse KEGG pathways; the genes were retrieved using the DBGET database [3]. We also selected 15 groups of highly related Pfam entries from the Sanger Pfam database [10]. For each group, we retrieved the corresponding human and yeast gene identifiers from the Uniprot database [12]. Assuming that genes belonging to a same KEGG pathway are often related to a similar biological process, the similarity values computed for this dataset should be related to the biological process GO aspect. And, assuming that genes which share common domains in a Pfam clan often have a similar molecular function, the similarity values computed for this second dataset should be related to the molecular function GO aspect.

B. Evaluating Recall and Precision

For each result gene x we constructed a feature vector $\Phi(x)$ relative to all other genes in the result. Each result gene is represented by its best functional distance to all other result genes. The distance between two genes x and y is given by $\| \Phi(x) - \Phi(y) \|$. We clustered results based on their similarity with the KEGG and Pfam pathway benchmarks. The similarity between result vector x and pathway vector y was taken as their normalized dot product:

$$sim\ (x, y) = \frac{\langle x, y \rangle}{\| x \| \| y \|} \qquad (5)$$

Each cluster contains results, whose similarity value with the pathways is 0.3, 0.5, 0.7, 0.8, or 1. For each cluster, we measured the *recall (or true positive rate)* and *precision* of GOseek and of DynGO. Let: (1) G_P be all genes in a pathway and n be the number of these genes, and (2) G_M be the m genes retrieved by a method as semantically related to input gene keywords:

- Recall = $(|G_M \cap G_P| / n)$ (6)
- Precision = $(|G_M \cap G_P| / m)$ (7)

Tables 6-10 show the results. As the tables show: (1) as the similarity between results and pathways increases, recall decreases and precision increases, which is expected since greater similarity means fewer results with higher precision, and (2) GOseek outperforms DynGO [6] in all results.

In summary, the recall and precision values for the two benchmarking datasets show that GOseek outperforms the DynGO method. The results reveal the robustness of the GOseek's method and its ability to reflect the semantic relationships among gene annotations.

TABLE 6. CLUSTER WHOSE SIMILARITY WITH THE PATHWAYS IS 0.3

Recall		Precision	
GOseek	DynGO	GOseek	DynGO
0.92	0.81	0.58	0.34

TABLE 7. CLUSTER WHOSE SIMILARITY WITH THE PATHWAYS IS 0.5

Recall		Precision	
GOseek	DynGO	GOseek	DynGO
0.88	0.81	0.67	0.45

TABLE 8. CLUSTER WHOSE SIMILARITY WITH THE PATHWAYS IS 0.7

Recall		Precision	
GOseek	DynGO	GOseek	DynGO
0.8	0.68	0.72	0.58

TABLE 9. CLUSTER WHOSE SIMILARITY WITH THE PATHWAYS IS 0.8

Recall		Precision	
GOseek	DynGO	GOseek	DynGO
0.75	0.58	0.78	0.64

TABLE 10. CLUSTER WHOSE SIMILARITY WITH THE PATHWAYS IS 1

Recall		Precision	
GOseek	DynGO	GOseek	DynGO
0.68	0.52	0.83	0.72

V. CONCLUSION

We proposed in this paper a biological search engine called GOseek, which overcomes the limitation of current gene similarity tools. Given a set of genes, GOseek returns the most significant genes that are semantically related to the given genes. We experimentally evaluated the quality of GOseek and compared it with DynGO [6]. Results showed that GOseek outperforms DynGO. The results showed also the robustness of GOseek to reflect the semantic relationships among gene annotations.

REFERENCES

[1] Baldi, P. & Hatfield, W. (2002), DNA Microarrays and Gene Expression, Cambridge University Press, Cambridge, UK.

[2] Cherry JM, Adler C, Ball C, Chervitz SA, Dwight SS, Hester ET, Jia Y, Juvik G, Roe T, Schroeder M, Weng S, Botstein D: SGD: Saccharomyces Genome Database. *Nucleic Acids Res* 1998, 26:73-79.

[3] DBGET database. Available at: http://www.genome.jp/dbget/

[4] Guo X, Liu R, Shriver CD, Hu H, Liebman MN: Assessing semantic similarity measures for the characterization of human regulatory pathways. Bioinformatics 2006, 22(8):967-973

[5] GO website: http://www.geneontology.org/

[6] Liu, H., Hu, Z., Wu, C. DynGO: a tool for visualizing and mining of Gene Ontology and its associations. BMC Bioinformatics 6 (201), 2005.

[7] Mouse Genetic Informatics: http://www.informatics.jax.org/

[8] Nagar A, Al-Mubaid H: A New Path Length Measure Based on GO for Gene Similarity with Evaluation using SGD Pathways. IEEE International Symposium on Computer-Based Medical Systems (CBMS 08)

[9] Pesquita C, Faria D, Falcão AO, Lord P, Couto FM (2009) Semantic Similarity in Biomedical Ontologies. PLoS Comput Biol 5(7): e1000443

[10] Sanger Pfam database. Available at: http://pfam.sanger.ac.uk/

[11] Twigger S, Lu J, Shimoyama M, Chen D, Pasko D, Long H, Ginster J, Chen CF, Nigam R, Kwitek A, Eppig J, Maltais L, Maglott D, Schuler G, Jacob H, Tonellato PJ: Rat Genome Database (RGD): mapping disease onto the genome. *Nucleic Acids Res* 2002, 30:125-128.

[12] Uniprot database. Available at: http://www.uniprot.org/