Fuzzy Measures on the Gene Ontology for Gene Product Similarity

Mihail Popescu, James M. Keller, and Joyce A. Mitchell

Abstract—One of the most important objects in bioinformatics is a gene product (protein or RNA). For many gene products, functional information is summarized in a set of Gene Ontology (GO) annotations. For these genes, it is reasonable to include similarity measures based on the terms found in the GO or other taxonomy. In this paper, we introduce several novel measures for computing the similarity of two gene products annotated with GO terms. The fuzzy measure similarity (FMS) has the advantage that it takes into consideration the context of both complete sets of annotation terms when computing the similarity between two gene products. When the two gene products are not annotated by common taxonomy terms, we propose a method that avoids a zero similarity result. To account for the variations in the annotation reliability, we propose a similarity measure based on the Choquet integral. These similarity measures provide extra tools for the biologist in search of functional information for gene products. The initial testing on a group of 194 sequences representing three proteins families shows a higher correlation of the FMS and Choquet similarities to the BLAST sequence similarities than the traditional similarity measures such as pairwise average or pairwise maximum.

Index Terms—Similarity measure, fuzzy measure, Choquet integral, Gene Ontology.

1 Introduction

The pace of gene discovery has increased tremendously over the past 10 years with the emphasis on sequencing the human genome and various other genomes. With this increased pace has come the need for tools to assist with the analysis of similarities between genes and among gene families. Genes are grouped in various ways, including being part of gene families, being part of a metabolic pathway, and being coregulated under various environmental conditions. The rapid expansion of knowledge about various protein isoforms that are produced from the same mRNA transcript but with alternate splicing is also creating needs for new measures of similarity among genes. In analyzing the similarity (or dissimilarity) between gene products, the obvious features to consider are the DNA sequence and the expression values. However, for many gene products, additional information is available. One form of information is symbolic, taking the form of associated Gene Ontology (GO) terms [1] and terms from a thesaurus used to index the publications about the gene or gene product [2]. Our goal is to incorporate these symbolic features into gene similarity functions that utilize as much common supportive evidence as possible (especially that information contained in the ontologic or taxonomic structure) while minimizing the effect of ambiguity and/or incomplete annotations. This paper describes several measures for gene product similarity. These novel similarity measures for gene product comparison, the FMS and Choquet, are based on fuzzy measures [3], [4] and fuzzy set theory that have been shown to be very effective in other domains [4], [5].

There are two categories of approaches to compute the similarity of two objects described by sets of terms that belong to a taxonomy. In the first category, the terms in the sets are considered individually; this category can be further divided into two approaches, pair-based and set-based. In the second category, the similarity measures use graph similarity techniques.

In the first category, the first approach is to aggregate the similarities between all pairs of terms from the two sets. The pairwise similarities are aggregated using a function such as maximum or average. Lord et al. [6] used the average of the pairwise GO term similarities to compute the similarity between two gene products, while Speer et al. [7] used the maximum for the same task. A good review of the pairwise similarities between individual objects that belong to a taxonomy is given in [8]. In [9], the pairwise similarity between Gene Ontology (GO) terms was used to search multiple biological databases. The similarity was computed using the information content [10] of a GO term. Ganesan et al. [11] develop several similarity measures for information retrieval, using various techniques. One measure involves a combination between average and maximum, called Optimistic Genealogy Measure (OGM), based on the depth in the hierarchy, to compare different customers based on their buying behavior. In the second approach, called the "bag of words" approach [12], the similarity is computed using set similarity measures such as Dice, Jaccard, or cosine [13]. A generalization on the cosine measure based on the depth in the hierarchy was one of the measures assessed by Ganesan et al. [11]. A widely acknowledged problem with the depth-based similarity is

- *M. Popescu is with the Health Management and Informatics Department, University of Missouri, Columbia, MO 65211. E-mail: popescum@missouri.edu.*
- *J.M. Keller is with the Electrical and Computer Engineering Department, University of Missouri, Columbia, MO 65211. E-mail: kellerj@missouri.edu.*
- *J.A. Mitchell is with the Department of Medical Informatics, School of Medicine, University of Utah, Salt Lake City, UT 84132. E-mail: Joyce.Mitchell@hsc.utah.edu.*

Manuscript received 20 Oct. 2004; revised 9 Mar. 2005; accepted 1 July 2005; published online 31 July 2006.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-0176-1004.

that the distance in a taxonomy is not uniform due to the variation in density of the various subtaxonomies [11]. Many approaches to Web content data mining have been developed with "bag of words" approaches [14].

In the second category, the objects in each set are considered as a tree (or graph) that is a part of the original taxonomy. The similarity between two sets is cast as a tree (graph) similarity problem. This problem is encountered in many domains where the information can be represented as a tree, such as 3D structure matching [15], MESH-based document retrieval [12], 2D shape recognition [16], multiagent systems [17], natural language processing [18], database search [9], etc. In the general case, this problem is NP-complete [16]. However, various techniques are described for computing the similarity in polynomial time [12], [15], [16], [19].

Fuzzy measures have not been used extensively in bioinformatics. Use of fuzzy techniques such as fuzzy clustering [20], fuzzy neural networks [21], fuzzy rule systems [22], and fuzzy relations [23] has been reported for microarray analyses. Similar techniques were used in bioinformatics in applications related to document content analysis [14].

In this paper, we are extending the work of Lord et al. [1], [6] who investigated semantic similarity measure to explore the Gene Ontology. We compare our new fuzzy measures to traditional set similarity measures such as Jaccard, Dice, and vector cosine and to pairwise similarities such as average and maximum as applied to the GO. We show that, by utilizing more information than the traditional measures, the fuzzy measures correlate better with the sequence-based similarity measures. These measures can also be applied to other semantic knowledge sources, especially those with knowledge structured into taxonomies such as MeSH.

The proposed fuzzy measure similarities address inconsistencies and inabilities of existent numeric comparisons used for gene products. BLAST scores do not account for the functions of the proteins, as do annotation-based similarities. Second, cardinality-based measures (such as Jaccard and Dice) ignore the information content of the annotation terms in their construction. A frequently used annotation term (given, say, by promiscuous domains such as SH3 and ATP-binding cassettes [29]) could artificially make two gene products look more similar than they actually are. Finally, the average of pairwise term information content [6] is inconsistent in the sense that self-similarity is not 1 when a product is annotated with more than one term. Similarly, the maximum of pairwise term information content is inconsistent since the similarity between two gene products that share just one term is 1, regardless of the rest of their annotation terms. It follows under this calculation that two gene products that share a "promiscuous domain" are very similar. Our proposed measures do not have these inconsistencies. In Section 6, we give numeric evidence to support this statement with respect to the collagen family of proteins.

2 Background

Given two gene products, G_1 and G_2, we can consider them as being represented by collections of terms $G_1 = \{T_{11}, \ldots, T_{1i}, \ldots, T_{1n}\}$ and $G_2 = \{T_{21}, \ldots, T_{2j}, \ldots, T_{2m}\}$. Based on the two sets, the goal is to define a similarity between G_1 and G_2, denoted as $s(G_1, G_2)$. The Jaccard and Dice similarity measures are computed as

$$\text{Jaccard similarity}: s_J(G_1, G_2) = \frac{|G_1 \cap G_2|}{|G_1 \cup G_2|}. \quad (1)$$

$$\text{Dice similarity}: s_D(G_1, G_2) = \frac{2|G_1 \cap G_2|}{|G_1| + |G_2|}. \quad (2)$$

Note that, in both the Jaccard and Dice measures, if $G_1 \cap G_2 = \emptyset$, the similarity is zero. This seems reasonable at first glance, but it is possible for two gene products to have terms that are siblings "deep within" the GO. These gene products should have nonzero similarity even though their annotation terms are not identical.

The annotations for the two gene products can be arranged into binary valued vectors $\mathbf{v}_i \in \Re^{NT}$, where NT is the total number of terms in the complete annotation set (a component of 1 if the annotation is present and 0 else). Then, various vector space-based similarity measures are calculated, such as the cosine similarity:

$$s_V(G_1, G_2) = \frac{\mathbf{v}_1 \bullet \mathbf{v}_2}{|\mathbf{v}_1||\mathbf{v}_2|}, \quad (3)$$

where $\mathbf{v}_1 \bullet \mathbf{v}_2$ is the dot product and $|.|$ represents the length of the vector (square root of the total number of annotations for the gene product). One advantage of this approach is that each gene product is described by an NT-dimensional feature vector, allowing the use of well-known vector space clustering algorithms such as c-means and fuzzy c-means [24]. However, if $NT >> 0$ (number of GO terms is large), the vectors v_i become long and sparse, making the clustering more problematic.

In the pairwise approach, similarity is computed considering the terms pairwise, say $s_{ij}(T_{1i}, T_{2j})$, and then the values for the pairs are aggregated using, for example, the average as:

$$s_{AVG}(G_1, G_2) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} s_{ij}}{mn}. \quad (4)$$

The problem with the average pairwise similarity is that it underestimates the similarity. The best illustration of this fact is that the self-similarity is less than one ($s(G_1, G_1) < 1$) if $m, n > 1$. Without normalization of some sort, the average is not a true similarity. If the maximum is used instead, the similarity is overestimated since it is enough that the two gene products share one term for the similarity to be 1. This is especially bad for the multidomain protein. Since they share functions (hence, GO terms), their similarity will be 1, making impossible any discrimination among them. The above problems are illustrated in Example 2.

All the above similarity measures can be easily generalized if we consider that each term, T_k, has a weight g^k associated with it. For example, the Jaccard similarity becomes

TABLE 1
Characteristics of the $\mathrm{GPD194_{12.10.03}}$ Data Set

Characteristics of the $\mathrm{GPD194_{12.10.03}}$ data set

ENSEMBL Family (ENSF)	F_i = Protein Family	No. of Genes	Gene Symbols (in order)	N_i = No. of Sequences
339	myotubularin	7	MTMR1÷4, MTMR6÷8	21
73	receptor precursor	7	FGFR1÷4, RET, TEK, TIE1	87
42	collagen alpha chain	13	COL1A2, COL21A1, COL24A1, COL27A1, COL2A1, COL3A1, COL4A1, COL4A2, COL4A3, COL4A6, COL5A3, COL9A1, COL9A2	86

$$\mathrm{s_{WJ}(G_1, G_2)} = \frac{\sum_{\{\mathrm{i}|\mathrm{T_i}\in \mathrm{G_1}\cap \mathrm{G_2}\}} \mathrm{g^i}}{\sum_{\{\mathrm{j}|\mathrm{T_j}\in \mathrm{G_1}\cup \mathrm{G_2}\}} \mathrm{g^j}}. \quad (5)$$

This will be referred to as the weighted Jaccard similarity.

The measures proposed in this paper try to overcome the limitations mentioned above, i.e., the zero similarity and the under/overestimation. In addition, the new Choquet measure tries to better incorporate into the similarity measure the effect of the reliability of the data elements (GO terms in our case).

We present a pilot study that demonstrates the promise of this new approach. The basis of our illustrative computations is a set of 194 human gene products that were clustered into three protein families using Markov clustering (MCL) [25]. The gene products (and their) families were retrieved on 10 December 2003 using the ENSEMBL browser (http://www.ensembl.org). Table 1 itemizes several characteristics of these clusters, which we call the $\mathrm{GPD194_{12.10.03}}$ data. Since this data is dynamic, we include the date on which it was extracted in the name.

These three gene families were chosen for several reasons. First, each family had multiple well-characterized genes, many of which are involved in human disorders when mutated and all of which could be considered very similar in both structure and function. Second, several of the genes, especially the receptor precursor genes, were characterized by multiple isoforms represented by the multiple sequences and, thus, representing extremely similar gene products. Third, the MCL clustering available through ENSEMBL had pulled together these gene families, allowing us a cluster method by which to benchmark our results. Fourth, the gene families were distinct from one another, but, in some cases, could be considered as having similar functions at a higher level that might be represented in the Gene Ontology (such as having catalytic activity). Thus, our sample had a range of similarities between genes and gene products. We were fortunate because there was quite a range of similarities between members of a gene family, with the myotubularins being quite similar, the receptor precursors having many isoforms, and the collagen alpha chains being quite diverse. The 194 human sequences are mapped to 27 genes (see Table 1).

To validate our similarity measures, we began with the same approach as in Lord et al. [6] by computing the correlation between the new measures and a sequence-based similarity. Unlike Lord et al. [6], we normalized the BLAST [26] bit scores, $\{\mathrm{s_{ij}} : \mathrm{s_{ij}} \in [0,1]; 1 \le \mathrm{i,j} \le 194\}$, using:

$$\mathrm{s(seq_i, seq_j)} = \frac{\mathrm{s_{raw}(seq_i, seq_j)}}{\min\{\mathrm{s_{raw}(seq_i, seq_i)}, \mathrm{s_{raw}(seq_j, seq_j)}\}}, \quad (6)$$

where $\mathrm{s_{raw}}$ is the natural logarithm of the BLAST bit score between $\mathrm{seq_i}$ and $\mathrm{seq_j}$.

The sequence-based similarity matrix obtained is shown in Fig. 1. The range of the numbers in Fig. 1 is [0, 1], 1 (dark) indicating high similarity (low distance).

3 Fuzzy Measure-Based Similarity Measure

The fuzzy measure similarity (FMS) is based on the concept of fuzzy measure, a generalization of probability measure. In this context, the terms in a combined set describing two gene products will be considered as "information sources" that support the similarity of the two genes. Let $\mathrm{G} = \{\mathrm{T_1}, \ldots, \mathrm{T_n}\}$ be a finite set of terms describing a gene

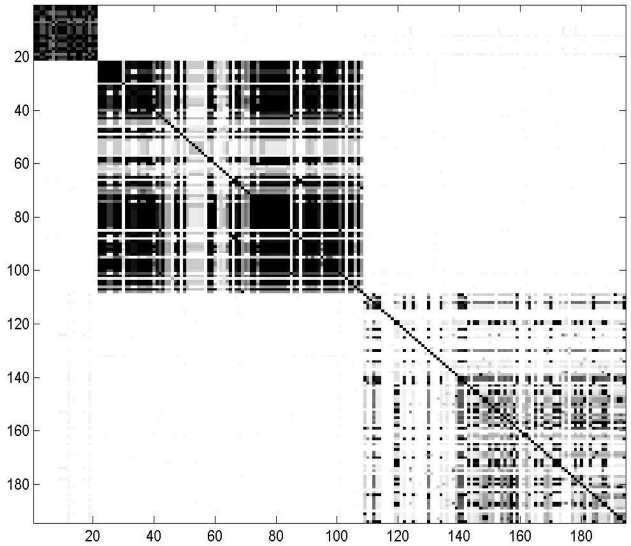

Fig. 1. The BLAST Sequence similarity matrices for the three gene families (the 194 sequences are presorted by family and by gene as shown in Table 1).

product. A fuzzy measure, g, is a real valued function $g : 2^G \rightarrow [0, 1]$, satisfying the following properties:

1. $g(\emptyset) = 0$ and $g(G) = 1$.
2. $g(A) \leq g(B)$ if $A \subseteq B$.

Note that the normal additivity condition of probability theory is replaced by the weaker condition of monotonicity (property 2). For a fuzzy measure g, let $g^i = g(\{T_i\})$. The mapping $T_i \rightarrow g^i$ is called a fuzzy density function. The fuzzy density value, g^i, is interpreted as the (possibly subjective) importance of the single information source T_i in determining the similarity of two genes. Fuzzy measures are quite general since they only require two simple properties to be satisfied. However, it is often the case that the densities can be extracted from the problem domain or supplied by experts. The key to using fuzzy measures involves finding ones that can be built out of the densities. One of the most useful classes of fuzzy measures is due to Sugeno [4]. A fuzzy measure g is called a Sugeno measure (g_λ-fuzzy measure) if it additionally satisfies the following property [3]:

3. For all $A, B \subseteq G$ with $A \cap B = \emptyset$.

$$g_\lambda(A \cup B) = g_\lambda(A) + g_\lambda(B) + \lambda g_\lambda(A) g_\lambda(B) \quad \text{for some } \lambda > -1. \tag{7}$$

The subscript λ will be omitted unless needed for simplicity. If the densities are known, the value of λ for any Sugeno fuzzy measure can be uniquely determined for a finite set G using (7) and the facts $G = \bigcup_{i=1}^{n} \{T_i\}$ and $g_\lambda(G) = 1$, which leads to solving the following equation for:

$$(1 + \lambda) = \prod_{i=1}^{n} (1 + \lambda g^i). \tag{8}$$

This equation has a unique solution for $\lambda > -1$ [3]. We mention that if $n = 1$, we use $g(T_1) = g^1$ instead of $g(T_1) = 1$, as follows from property 1. This insures that two genes that share two GO terms are more similar than two genes that share just one GO term.

For our application, the set of fuzzy density values is constructed from the information sources in the set G in a simple fashion, adapting the approach in [10]. The densities are computed as the information content for the particular term determined from a given corpus, SWISS-PROT in our case. In particular, for each term, T_k, in the GO, we counted the number of occurrences in the corpus of the term or any of its children and converted it to a probability, i.e.,

$$p(T_k) = \left(\frac{\text{count}(T_k + \text{children of } T_k \text{ in corpus})}{\text{count}(\text{all GO terms in corpus})} \right) \quad 1 \leq k \leq |GO|.$$

Then, we define the density $g^k = g(\{T_k\})$ by

$$g^k = ic(T_k) = -\ln(p(T_k) / \max_{T_j \in GO} \{-\ln(p(T_j)\}, \tag{9}$$

where $ic(T_k)$ is the information content of T_k computed as in [10]. The denominator is used to scale the values into the interval [0, 1].

Example 1. Calculation of a Sugeno measure for a gene product term set. The set of supporting GO terms for the sequence with GenBank ID AAN03650 (COL24A1 gene) is:

$$G = \{T_1 = 5201(\text{extracellular matrix structural component}), T_2 = 7155(\text{cell adhesion}), T_3 = 5581(\text{collagen})\}.$$

The associated densities are $\{g^k\} = \{0.58, 0.44, 0.65\}$, calculated as described above. Then, we solve (8) as

$$1 + \lambda = (1 + 0.58\lambda)(1 + 0.44\lambda)(1 + 0.65\lambda) \Rightarrow \lambda = -0.86.$$

The Sugeno measure becomes:

$$g(\{T_1\}) = g^1 = 0.58, g(\{T_2\}) = g^2 = 0.44,$$
$$g(\{T_3\}) = g^3 = 0.65, g(\{T_1, T_2\}) = g^1 + g^2 + \lambda g^1 g^2 = 0.8.$$

Similarly,

$$g(\{T_1, T_3\}) = 0.9, g(\{T_3, T_2\}) = 0.84, g(G) = 1.$$

Given the above development, we define our similarity measure [27].

Definition 1. *Fuzzy measure-based similarity (FMS). The similarity* $s_{FMS}(G_1, G_2)$ *between two sets* G_1 *and* G_2 *of terms is defined as:*

$$s_{FMS}(G_1, G_2) = \frac{g_1(G_1 \cap G_2) + g_2(G_1 \cap G_2)}{2}, \tag{10}$$

where g_1 *is the Sugeno measure defined on* G_1 *from the densities* $\{g^{1i}\}$ *and* g_2 *is the Sugeno measure defined on* G_2 *from the densities* $\{g^{2j}\}$.

Example 2. Case 1: Similarity calculations for two gene products from the same family: Consider the sequence G_1 with GenBank ID AAH35609 (MTMR4 gene) and the sequence G_2 with GenBank ID AAH12399 (MTMR8 gene). These are two members of the same family and, hence, should be quite similar to each other. The GO terms associated with the above sequences are

$$G_1 = \{T_1 = 4721(\text{protein phosphatase activity}), T_2 = 6470(\text{protein amino acid dephosphorylation}), T_3 = 8270(\text{zinc ion binding})\}$$

and

$$G_2 = \{T_1 = 4721(\text{protein phosphatase activity}), T_2 = 6470(\text{protein amino acid dephosphorylation}), T_4 = 16787(\text{hydrolase activity})\}.$$

The sets of related densities are $\{g^{1i}\} = \{0.52, 0.57, 0.54\}$ and $\{g^{2i}\} = \{0.52, 0.57, 0.33\}$. Here, the set of common terms that supports the similarity of G_1 and G_2 is $\{T_1, T_2\}$.

To calculate the FMS, we need to build the two measures. The Sugeno measure for G_1 has $\lambda = -0.84$, resulting in the measure of the common set of $g_1(\{T_1, T_2\}) = 0.84$. The Sugeno measure for G_2 has $\lambda = -0.72$, resulting in $g_2(\{T_1, T_2\}) = 0.88$. Hence, the FMS similarity, s_{FMS}, is:

TABLE 2
Various Similarity Values between MTMR4 and MTMR8

Type	Fuzzy Similarity		'Bag of words' Similarity			Pair-wise Similarity		Sequence Similarity	
Measure	FMS	Weighted Jaccard	Jaccard	Dice	Cosine	Avg.	Max.	Smith-Waterman	BLAST
Similarity	0.86	0.57	0.5	0.67	0.75	0.28	1	0.83	0.85

TABLE 3
Various Similarity Values between MTMR4 and COL5A3

Type	Fuzzy Similarity		'Bag of words' Similarity			Pair-wise Similarity		Sequence Similarity
Measure	FMS	Weighted Jaccard	Jaccard	Dice	Cosine	Avg.	Max.	BLAST
Similarity	0.57	0.54	0.08	0.15	0.27	0.05	1	0.4

$$s_{FMS}(G_1, G_2) = \frac{g_1(\{T_1, T_2\}) + g_2(\{T_1, T_2\})}{2} = \frac{0.84 + 0.88}{2} = 0.86.$$

Other similarity measures for the same two proteins are given in Table 2.

The above two myotubularin genes should have high similarity since they belong to the same ENSEMBL myotubularin family. From Table 2, we see that the FMS value is closest to the BLAST and Smith-Waterman scores. The worst value is given by the pairwise average that grossly underestimates the similarity. From the above example, we see that the FMS is more sensitive to the elements that the two term sets have in common: If the common elements have a high information content, then the similarity is stronger. This fact agrees with our intuition about similarity. Another consequence of the same idea is that while, in the vector cosine similarity, the noncommon elements have no contribution (they are multiplied by zero), in FMS, they do contribute implicitly since the fuzzy measures are defined a priori for each term set.

Case 2: Similarity calculations for two gene products from different families. Consider the sequence G_1 with GenBank ID AAC12865 (MTMR2 gene) and the sequence G_2 with GenBank ID AAF59902 (COL5A3 gene). Since the first sequence is a member of the myotubularins and the second belongs to the alpha collagens, their similarity should be low. The GO terms associated with the above sequences are

$$G_1 = \{T_1 = 4722, T_2 = 4725, T_3 = 16787, T_4 = 6470, T_5 = 7517(\text{muscle development}), T_6 = 8151, T_7 = 8372, T_8 = 8138\}$$

and

$$G_2 = \{T_1 = 5201, T_2 = 7155, T_3 = 7397, T_4 = 7517(\text{muscle development}), T_5 = 5581, T_6 = 5588\}.$$

The set of related densities for G_1 are:

$$\{g^{1i}\} = \{0.6134, 0.8778, 0.3274, 0.5713, 0.5139, 0.2026, 0.3545, 0.7093\},$$

while those for G_2 are:

$$\{g^{2i}\} = \{0.8778, 0.7093, 0.3274, 0.5713, 0.2222, 0.2474\}.$$

Here, the only common term is $T_{15} = T_{24} = 7517$. Hence, the FMS similarity is $s_{FMS} = 0.5713$. Other similarity measures for the same two proteins are given in Table 3.

We see that, in this case, the weighted Jaccard and the FMS perform best while the pairwise maximum grossly overestimated the similarity value.

However, so far, the FMS has the same problem as the one previously mentioned for the vector cosine similarity and Jaccard similarity, that is, if $G_1 \cap G_2 = \emptyset$, then the similarity is zero. In this case, we have no information about the relation between the two sets. In the next section, we describe a method that solves this problem when the objects in the set belong to a taxonomy.

4 Augmenting the FMS for Sets of Ontology Objects

The Gene Ontology is a Directed Acyclic Graph where a child node is considered a more specialized object than the parent node. As in the previous section, assume that the objects in the GO have associated densities $\{g^i\}$, for example, the information content formed from studying a corpus, like SWISS-PROT. The key is that the further down one goes in the tree, the higher the associated densities are.

Consider the same sets $G_1 = \{T_{11}, \ldots, T_{1i}, \ldots, T_{1n}\}$ and $G_2 = \{T_{21}, \ldots, T_{2j}, \ldots, T_{2m}\}$, where $T_{1i}, T_{2j} \in GO$. The idea of the proposed method is to augment each set as:

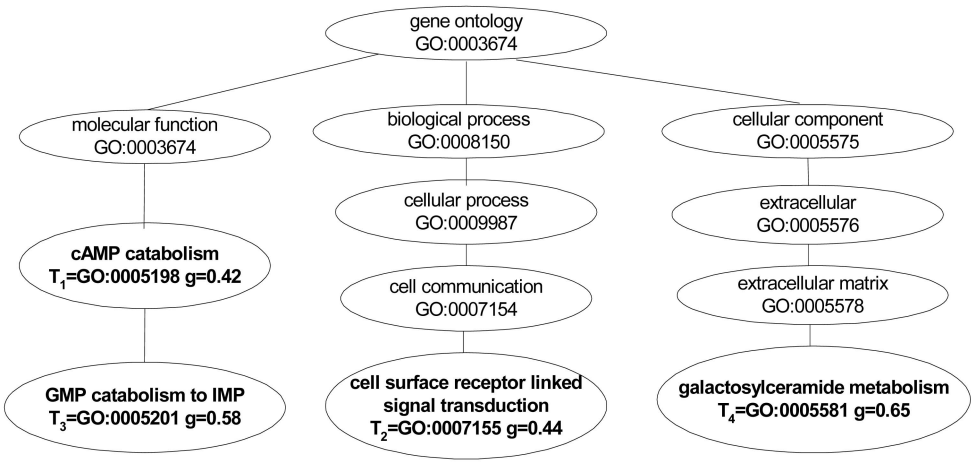

Fig. 2. View of GO containing the terms from Example 3 and their associated information-based weights.

$$G_1^+ = G_1 \cup \{T_{1i,2j}\} \text{ and } G_2^+ = G_2 \cup \{T_{1i,2j}\}, \quad (11)$$

where $\{T_{1i,2j}\}$ is the set of nearest common ancestors (NCA) of every pair (T_{1i}, T_{2j}). For each pair, the NCA is determined by the same Perl scripts that were used by Lord et al. in [6]. The resulting augmented intersection is:

$$[G_1 \cap G_2]^+ = [G_1^+ \cap G_2^+] = [G_1 \cap G_2] \cup \{T_{1i,2j}\}. \quad (12)$$

Using the augmented intersection, $[G_1 \cap G_2]^+$, the augmented FMS (AFMS), denoted by $s_{AFMS}(G_1, G_2)$, is defined as:

$$s_{AFMS}(G_1, G_2) = \frac{g_1^+([G_1 \cap G_2]^+) + g_2^+([G_1 \cap G_2]^+)}{2}, \quad (13)$$

where g_k^+ is the fuzzy measure computed on $G_k^+, k = \{1, 2\}$.

Example 3. AFMS calculation for reasonably similar gene products. Let us compute the GO similarity between the sequence with GenBank ID AAL02227 (COL21A1 gene) described by

$$G_1 = \{T_1 = 5198(\text{structural molecular activity}), \\ T_2 = 7155(\text{cell adhesion})\}$$

and the sequence BAB13947 (COL27A1 gene) described by

$$G_2 = \\ \{T_3 = 5201(\text{extracellular matrix structural constituent}), \\ T_4 = 5581(\text{collagen})\}.$$

We see that all of the Jaccard, Dice, cosine, and FMS similarity measures are 0 for this case. However, the two sequences are obviously similar since they are both in the collagen alpha 1 family. Also note that T_3 is a child in the GO of T_1.

The augmented sets are: $G_1^+ = \{T_1, T_2\}$ and $G_2^+ = \{T_1, T_3, T_4\}$. Since $NCA(T_3) = T_1$ (see Fig. 2) and the root node is ignored because its information content is 0 (common for all terms), the augmented intersection is $[G_1 \cap G_2]^+ = \{T_3\}$. Hence, the augmented FMS is:

$$s_{AFMS}(G_1, G_2) = \frac{0.42 + 0.42}{2} = 0.42.$$

We mention that, for the same case, the augmented Jaccard similarity is 0.25 and the augmented vector cosine similarity is 0.4. We conclude that the augmentation procedure works for all set-based similarity measures by taking advantage of the hierarchical structure of GO and adds value by taking advantage of the ontology structure.

5 Choquet Fuzzy Integral-Based Set Similarity Measure

Fuzzy integrals have been shown to be very useful for evidence fusion [3], [4], [5]. Fuzzy integrals combine the objective evidence supplied by each information source (the s-function in our scenario and discussed below) and the expected worth of each subset of information sources (via a fuzzy measure as above) to assign confidence to hypotheses and to rank alternatives in decision-making. This is a nonlinear combination of information and the worth of these information sources with respect to the decision is in dealing with the reliability in both forms of data.

For the purpose of comparing two gene products described by sets of GO terms, suppose that $G_1 = \{T_{11}, \ldots, T_{1i}, \ldots, T_{1n}\}$ and $G_2 = \{T_{21}, \ldots, T_{2j}, \ldots, T_{2m}\}$. Let $X = G_1 \times G_2$ and $s : X \rightarrow [0, 1]$ be a similarity function, i.e., $s_{ij}(T_{1i}, T_{2j})$ is the similarity between the pair of GO terms (T_{1i}, T_{2j}). To simplify the notation, we reorder the term pairs and label them by a single subscript so that $X = \{T_1, T_2, \cdots, T_{nm}\}$. The elements of X (pairs of GO terms) are considered to be sources of information that support the similarity of genes G_1 and G_2 to degree $s(T_k)$, where $T_k = (T_{1i}, T_{2j})$ for some i and j.

The confidence of the similarity is based on the confidence of each type of annotation (see Table 4). The confidence of a pair of terms, $c^{ij}(T_{1i}, T_{2j})$, could be assigned as $c^{ij}(T_{1i}, T_{2j}) = f(c(T_{1i}), c(T_{2j}))$, where f can be the maximum, average, or minimum operator and $c(T_{1i})$ and $c(T_{2j})$ are the respective reliabilities of assigning the annotations T_{1i} and T_{2j}. In the language of fuzzy measures, $c^{ij}(T_{1i}, T_{2j})$ represents the fuzzy density of the information source $T_k = (T_{1i}, T_{2j})$ for a fuzzy measure g over X. If X is a discrete set

TABLE 4
Numeric Values Chosen for the Reliability of the GO Annotation

Traceable author statement	Inferred from sequence similarity	Inferred from electronic annotation	Non-traceable author statement	Not documented	Not recorded
TAS	ISS	IEA	NAS	ND	NR
1	0.8	0.6	0.4	0.1	0.1

(as it is here), the Choquet similarity can be computed as follows:

$$s_{\text{Choquet}}(G,G_2) = \sum_{i=1}^{nm} \left[s(T_{(i)}) - s(T_{(i+1)})\right] \cdot g(S_i), \qquad (14)$$

where the function values are reordered so that

$$s(T_{(1)}) \geq s(T_{(2)}) \geq \cdots \geq s(T_{(nm)}),$$
$$s(T_{(nm+1)}) = 0,$$
$$S_i = \{T_{(1)}, \cdots, T_{(i)}\},$$

and g is the fuzzy measure generated by the set of densities $\{c^{ij}\}$.

Example 4. Choquet integral-based set similarity. Consider the same gene products as in Example 2, Case 1, $G_1 =$ AAH35609 (MTMR4 gene) and $G_2 =$ AAH12399 (MTMR8 gene). The GO terms associated with the above sequences together with their source codes are

$$G_1 = \{T_1 = 4721(\text{TAS}), T_2 = 6470(\text{IEA}), T_3 = 8270(\text{NR})\},$$
$$G_2 = \{T_1 = 4721(\text{ISS}), T_2 = 6470(\text{NAS}), T_4 = 16787(\text{NR})\}.$$

For each type of annotation, we associated a numeric value related to the reliability of that annotation (see Table 4).

We note that, in Table 4, we show only the six types of annotations that are present in our data set. For the extended set of types of GO annotations, the reader is referred to the GO Web site (www.geneontology.org). Also, the numeric values attached to each annotation type were chosen somewhat arbitrarily. They only obey the relation (www.geneontology.org): TAS > ISS > IEA > NAS. However, the Choquet integral framework allows for the computation of the above numeric values given target similarities for a set of gene products.

The reliability values for these two sets are $\{c^{1i}\} = \{1, 0.6, 0.1\}$ and $\{c^{2i}\} = \{0.8, 0.4, 0.1\}$. (Recall that the densities represent the importance of the single source of information in establishing similarity.). The pairwise similarity matrix and the related densities (combined using min) are given below:

$$s(T_{1i}, T_{2j}) = \begin{vmatrix} 0.52 & 0.33 & 0 \\ 0.1 & 0.1 & 0 \\ 0 & 0 & 0.58 \end{vmatrix},$$

$$c(T_{1i}, T_{2j}) = \begin{vmatrix} 0.8 & 0.4 & 0.1 \\ 0.6 & 0.4 & 0.1 \\ 0.1 & 0.1 & 0.1 \end{vmatrix},$$

where, for example, $c^{11} = \min(c(T_{11}), c(T_{21})) = 0.8$, and $s_{11} = s(T_{11}, T_{21}) = 0.52$. From the above matrices, the ordered pairwise term similarities are $\{s(T_{(i)})\} = \{0.58, 0.52, 0.33, 0.1, 0.1, 0, 0, 0, 0\}$ and the corresponding pairwise densities are

$$\{c^{(i)}\} = \{0.1, 0.8, 0.4, 0.4, 0.6, 0.1, 0.1, 0.1, 0.1\}.$$

In this paper, we use the following decomposable measure [28]: $g(\{c_{(1)}, c_{(2)}\}) = \min(1, g(\{c_{(1)}\}) + g(\{c_{(2)}\}))$. There are many choices for the form of the measure. We chose this one because of the ease of implementation. Hence, the Choquet similarity in this case becomes:

$$s_{\text{Choquet}} = [0.1(0.58 - 0.52) + 0.9(0.52 - 0.33) + 1(0.33 - 0.1) + 1(0.1 - 0.1) + 1(0.1 - 0) = 0.5.$$

The above value is between the average (0.28) and max (1) and depends on the reliability values assigned to the sources of annotation. The underlying hypothesis is that using annotation uncertainty (reliability) can help us model part of our uncertainty about the similarity of the two sequences. As the knowledge of various components of the GO annotations becomes more certain or changes with new experiments, then the weights of the evidence used to calculate the Choquet measure can be easily adjusted. This is particularly useful in a situation like gene function where the knowledge is changing rapidly.

An alternative similarity definition that accounts for the information reliability is the weighted Jaccard formula in which we multiply the information content of each term by its confidence factor (denoted as Reliability Weighted Jaccard, RWJ). In this case, the similarity from Example 4 becomes:

$$s_{\text{RWJ}} = (0.52 * \min(0.8, 1) + 0.57 * \min(0.4, 0.6))/ (0.52 * \min(0.8, 1) + 0.57 * \min(0.4, 0.6) + 0.33 * 0.1 + 0.54 * 0.1) = 0.88.$$

We note that the first two GO terms, although common to both gene products, do not have the same annotation confidence in both sets. Here, we used "min" to combine the confidences of the same term in different sets, although other operators such as average could be used.

In the next section, we validate the GO similarity measure introduced in this paper by investigating its correlation to sequence-based similarity measures.

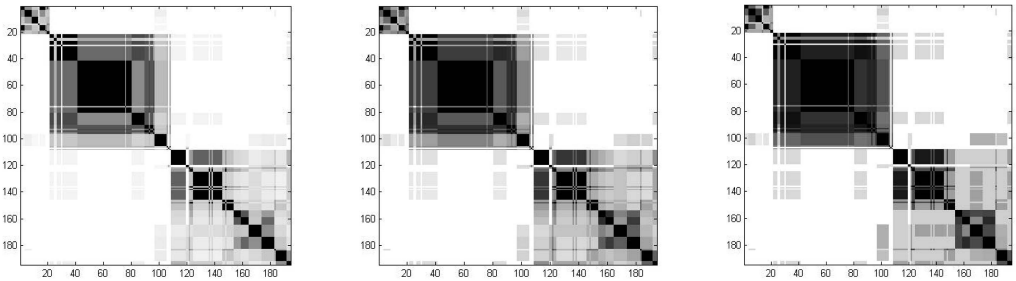

Fig. 3. GO similarity matrix for 194 human sequences. (a) Jaccard similarity. (b) Cosine similarity. (c) Fuzzy measure similarity.

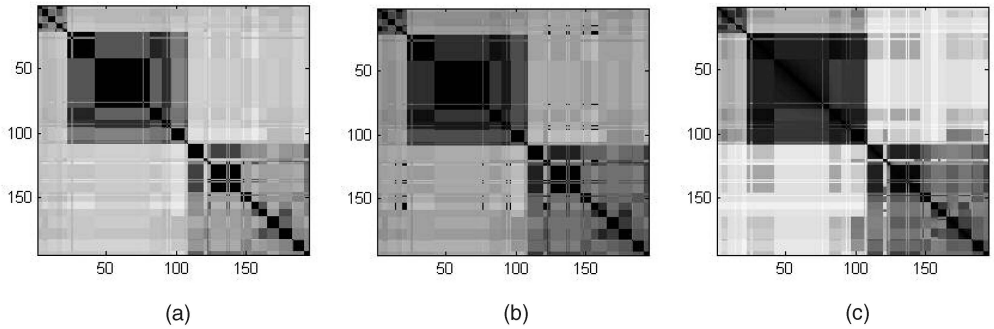

Fig. 4. Augmented GO similarity matrix for 194 human sequences. (a) Augmented Jaccard similarity. (b) Augmented cosine similarity. (c) Augmented fuzzy measure similarity.

6 GO Similarity between Myotubularin, Receptor Precursor, and Collagen Protein Families

The GO similarity for the above 194 sequences was computed as follows: First, the GO terms were extracted for each sequence using the ENSEMBL browser (http://www.ensembl.org/) in December 2003. Next, the densities for each applicable GO term were defined as normalized information content. The density, g^k, for term T_k is computed from (9), as in Lord et al. [1], [6]. The count of GO annotations of SWISS-PROT was performed in September 2003, where the total number of GO annotations in SWISS-PROT was found to be 83,468. Hence, the normalization factor was $-\ln(1/83,468) = 11.33$. Then, the complement of similarities was calculated on the term intersection sets using the information theoretic densities.

In Fig. 3, we show the three similarity matrices (Jaccard, cosine, and FMS) obtained for all 194 sequences. To more easily assess the similarity, the sequences were ordered in advance in the order of Table 1, that is, 1 to 21 are sequences from the myotubularin family, 22 to 108 are sequences from the receptor precursor family, and 109 to 194 are sequences from the collagen alpha family.

As we see from Fig. 3, in general, the similarity among the members of the same family is high while the similarity between families is low for all three similarity measures. Since most sequences that belong to the same gene have similar annotation, we expect to see dark squares on the diagonal of the similarity matrix.

In Fig. 4, we display the three similarity matrices obtained using the augmented version of the above similarities (augmented Jaccard, augmented cosine, and augmented FMS) for the $GPD194_{12.10.03}$ data set.

Comparing Fig. 4 to Fig. 3, more details appear in the upper right and lower left corners of the three family similarity matrices since the augmentation procedure replaces most of the zeros with nonzero values. This stronger within family similarity should produce more consistent agreement with the extracted ENSEMBL families during clustering. Note also that there is now more similarity between the three families since the augmentation procedure takes high-level genetic functions and processes into account.

To quantitatively assess the GO similarities, we compute the correlation between the GO similarities and the sequence similarity. Lord et al. [6] used a similar procedure to show that the average pairwise GO similarity is correlated to BLAST bit scores. To plot the GO similarity against the BLAST score, we used 10 BLAST bins (0.1 apart). For each bin, we averaged the GO similarity of the corresponding gene products. The value of the averages is then used by the plotting procedure. The average values for several GO similarities are shown in Fig. 5.

We can make several observations here. First, apparently the maximum pairwise aggregation has the highest correlation to BLAST. However, it also has the higher average standard deviation per bin (0.13 compared to 0.1 for FMS and about 0.7 for the others in Fig. 5). Second, it does not appear that the weighted Jaccard similarity is better than the unweighted Jaccard. Third, we see that the AFMS is better correlated to BLAST than FMS for values lower that than 0.7. This is due to the extra nonzero values that AFMS produces in those cases where the intersection of the two

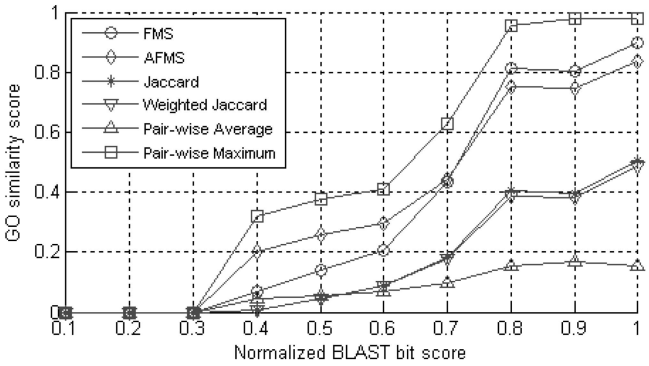

Fig. 5. GO average similarity score versus normalized BLAST bit score.

annotation sets is empty. At high values, this effect is not present anymore.

The above observations were performed on averaged GO similarities for each sequence similarly bin not considering the standard deviation of each bin. To assess the point-by-point correlation between GO similarities and the BLAST sequence similarity, one can use the correlation coefficient (see Table 5).

From Table 5, we see that maximum no longer correlates best to BLAST. In this case, the best correlation was obtained by the augmented fuzzy measure similarity (AFMS), although it is not striking. The average aggregation, together with the Jaccard measures, had the worst correlation. Overall, the correlation values in Table 5 are low. Looking at Fig. 1, we see that the BLAST scores between members of the collagen family (lower left corner) are not consistent. By contrast, in Fig. 3 and Fig. 4, the values in the same area are very high for the GO measures. This discrepancy produces a low overall correlation coefficient. We now use an alternative "target similarity" for comparison. Since we know the family assignment, we can compare the GO similarity to the ideal-case similarity matrix defined as:

$$s_{\text{ideal}}(i,j) = \begin{cases} 1 & \text{if i,j are in the same family} \\ 0 & \text{else.} \end{cases}$$

For this case, the correlation coefficients are given in Table 6.

The correlations are much higher in this case. Again, the fuzzy measure has the highest degree of correlation. The augmented fuzzy measure has a lower correlation than the nonaugmented version since, by design, it has nonzero elements for genes outside the same family.

We present the results of the correlations for the similarity measures that account for the reliability of the annotations in Table 7 and display them in Fig. 6.

The Choquet similarity correlates better with the BLAST sequence similarity than the reliability weighted Jaccard does. However, the Jaccard-based similarity has the advantage of being faster.

To provide supportive evidence that FMS mirrors human-expert opinion of close relationships between proteins that are scored as similar by it, consider the following: Myllyharju and Kivirikko [30] propose a division of the collagen superfamily into nine families. In light of their classification, we will give two examples of similar and dissimilar gene products from our data set. The similarities are computed using BLAST, Jaccard, and FMS. Myllyharju and Kivirikko group COL1A2 and COL24A1 as fibril

TABLE 5
Correlation Coefficient between GO Similarity and BLAST Similarity

GO similarity	**FMS**	**AFMS**	**Jaccard**	**Weighted Jaccard**	**Average**	**Maximum**
Correlation Coefficient (vs. BLAST)	0.52	0.54	0.44	0.44	0.44	0.47

TABLE 6
Correlation Coefficient between GO Similarity and the Ideal-Case Similarity

GO similarity	**FMS**	**AFMS**	**Jaccard**	**Weighted Jaccard**	**Pair-wise Average**	**Pair-wise Maximum**
Person's Coefficient (vs. BLAST)	0.9	0.86	0.72	0.7	0.82	0.84

TABLE 7
Correlation Coefficient for the Measures Using the Information Reliability

Similarity Measure/ Comparison target	**Reliability Weighted Jaccard**	**Choquet**
Correlation coefficient (BLAST)	0.41	0.49
Correlation coefficient (Ideal case 1-0 similarity)	0.65	0.85

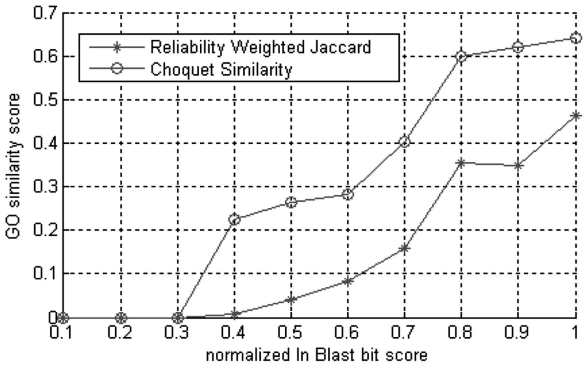

Fig. 6. Correlation between GO similarities that uses the reliability of the annotations and BLAST sequence similarity.

forming collagens in one group and COL21A1 in another group as fibril associated collagens. The results of the three similarity measures are given in Table 8.

Jaccard is clearly inconsistent since the value for the similar pair is smaller than that for the less similar pair. BLAST values are very small since no function is taken into account in its computation. FMS values are clearly consistent. First, the less similar pair is still "somewhat" similar since they belong to the same superfamily (collagen), while the value for the similar pair within the same family is very high, as it should be.

To demonstrate that the above examples are not accidental, we performed one more experiment. We were interested to find out how is the clustering of our collagen data set (85 gene product sequences) compared to the classification made by Myllyharju and Kivirikko. From the nine collagen families mentioned by Myllyharju and Kivirikko, we have in our data set representatives of three collagen families only: fibril forming collagens (FFC): {COL1A1, COL2A1, COL3A1, COL5A3, COL24A1, COL27A1}, type IV collagens {COL2A1, COL2A2, COL2A3, COL2A6}, and the fibril associated collagens with interrupted triple helices (FACIT) {COL9A1, COL9A2, COL21A1}. The clustering was performed using fuzzy c-means (using a suggestion from [31]) to group the 85 sequences into three clusters. When the FMS similarity was used, the members of the three families were classified perfectly, i.e., all members of the same family ended up in a single cluster. When the Jaccard similarity was used, the members of the FACIT family were merged into the type IV collagen family. With BLAST, the families were mixed together in the final three clusters and, hence, no family structure could be identified.

The relative execution time of the above similarities is given in Table 9, with Jaccard similarity being chosen as reference. Although the assessment is dependent of our implementation (MATLAB code), it offers an idea of the execution time versus accuracy trade-off.

While the Jaccard-based measures are the fastest, their correlations to the sequence-based similarities are the worst. If one desires (and has the computational means) to trade computation time for accuracy, the fuzzy measures are the method of choice.

7 Conclusions

In this paper, we investigated several novel measures that can be used to assess the similarity of two gene products based on the GO terms describing them. These similarity measures are intended to provide extra tools for the biologist in search of functional information about gene products. The methods presented here can be extended to other sets of annotations, such as motifs, domains, literature articles, etc.

The fuzzy measure similarity utilizes the Sugeno fuzzy measure with fuzzy densities calculated using an information theoretic approach. For the case when the intersection of the two sets is empty, we proposed an augmentation procedure that avoids forcing the resulting similarity to be zero by taking advantage of the structured nature of the ontology. We also proposed a method based on the Choquet integral to include the quality (reliability) of the annotation in the similarity measure. We showed that the proposed similarities correlate better to BLAST than the previously used approaches, average and maximum pairwise similarity.

TABLE 8
Comparison of Three Similarity Measures Values to the Expert's Opinion

Gene pair	Myllyharju	BLAST	Jaccard	FMS
COL24A1-COL21A1	Not similar	0.09	0.75	0.44
COL1A2-COL24A1	Similar	0.14	0.67	0.88

TABLE 9
Relative Execution Time of the GO-Based Similarity Procedures

Measure	Jaccard	Weighted Jaccard	Reliability Weighted Jaccard	FMS	AFMS	Choquet	Pair-wise Average	Pair-wise Maximum
Relative Execution Time	1	1.04	1.06	1.4	2	1.65	1.25	1.3

Since the FMS provides a different "look" at gene product similarity, we believe that it could be used in conjunction with other sequence-based similarity tools (such as BLAST) to improve clustering and knowledge discovery in gene product databases.

As future research, we plan to scale up our experiments to much larger gene product databases and to investigate the fusion of GO similarity with those calculated from sequence, expression, motif, etc., to produce more effective tools for knowledge discovery. We also intend to test our approach on microarray experimental data.

REFERENCES

[1] P.W. Lord, R.D. Stevens, A. Brass, and C.A. Goble, "Semantic Similarity Measure as a Tool for Exploring the Gene Ontology," *Proc. Pacific Symp. Biocomputing,* pp. 601-612, 2003.

[2] S. Raychaduri and R.B. Altman, "A Literature-Based Method for Assessing the Functional Coherence of a Gene Group," *Bioinformatics,* vol. 19, no. 3, pp. 396-401, Feb. 2003.

[3] *Fuzzy Measures and Integrals: Theory and Applications,* M. Grabisch, et al., eds. Springer-Verlag, 2000.

[4] M. Sugeno, "Fuzzy Measures and Fuzzy Integrals—A Survey," *Fuzzy Automata and Decision Processes,* pp. 89-102, 1977.

[5] J. Keller, P. Gader, and A.K. Hocaoglu, "Fuzzy Integrals in Image Processing and Recognition," *Fuzzy Measures and Integrals: Theory and Applications,* pp. 435-466, 2000.

[6] P.W. Lord, R.D. Stevens, A. Brass, and C.A. Goble, "Investigating Semantic Similarity Measures across the Gene Ontology: The Relation between Sequence and Annotation," *Bioinformatics,* vol. 19, no. 10, pp. 1275-1283, 2003.

[7] N. Speer, C. Spieth, and A. Zell, "A Memetic Clustering Algorithm for the Functional Partition of Genes Based on the Gene Ontology," *Proc. 2004 IEEE Symp. Computational Intelligence in Bioinformatics and Computational Biology (CIBCB 2004),* Oct. 2004.

[8] J.J. Jiang and D.W. Conrath, "Semantic Similarity Based on Corpus Statistics and Lexical Ontology," *Proc. Int'l Conf. Research on Computer Linguistics X,* 1997.

[9] S.L. Cao, L. Qin, W. He, Y. Zhong, Y. Zhu, and Y. Li, "Semantic Search Among Heterogeneous Databases Based on Gene Ontology," *Acta Biochemistry Biophysics Sinica,* vol. 36, no. 5, pp. 365-370, 2004.

[10] P. Resnik, "Semantic Similarity in a Taxonomy: An Information-Base Measure and Its Application to Problems of Ambiguity in Natural Language," *J. Artificial Intelligence Research (JAIR),* vol. 11, pp. 95-130, 1999.

[11] P. Ganesan, H. Garcia-Molina, and J. Widom, "Exploiting Hierarchical Domain Structure to Compute Similarity," *ACM Trans. Information Systems,* vol. 21, no. 1, pp. 64-93, Jan. 2003.

[12] J. Ontrup, T. Nattkemper, O. Gerstung, and H. Ritter, "A MeSH Term Based Distance Measure for Document Retrieval and Labeling Assistance," *Proc. 25th Ann. Int'l Conf. IEEE Eng. in Medical and Biological Societies (EMBC 2003),* Sept. 2003.

[13] C.D. Manning and H. Schutze, *Foundations of Statistical Natural Language Processing.* MIT Press, 2001.

[14] R. Kosala and H. Blockeel, "Web Mining Research: A Survey," *ACM SIGKDD Explorations Newsletter,* vol. 2, no. 1, June 2000.

[15] K.F. Aoki, A. Yamaguchi, Y. Okuno, T. Akutsu, N. Ueda, M. Kanehisa, and H. Mamitsuka, "Efficient Tree-Matching Methods for Accurate Carbohydrate Database Queries," *Genome Informatics,* vol. 14, pp. 134-143, 2003.

[16] A. Torsello, D. Hidovic, and M. Pelillo, "Four Metrics for Efficiently Comparing Attributed Trees," *Proc. 17th Int'l Conf. Pattern Recognition,* vol. 2, pp. 467-470, 2004.

[17] V.C. Bhavsar, H. Boley, and L. Yang, "A Weighted-Tree Similarity Algorithm for Multi-Agent Systems in E-Business Environments," *Proc. 2003 Workshop Business Agents and the Semantic Web,* pp. 53-72, June 2003.

[18] G. Sampson, R. Haigh, and E. Atwell, "Natural Language Analysis by Stochastic Optimization: A Progress Report on Project," *J. Experimental and Theoretical Artificial Intelligence,* vol. 1, pp. 271-287, Apr. 1989.

[19] J. Zhong, H. Zhu, J. Li, and Y. Yu, "Conceptual Graph Matching for Semantic Search," *Proc. 10th Int'l Conf. Conceptual Structures,* pp. 92-196, 2002.

[20] J. Wang, T.H. Bo, I. Jonassen, O. Myklebost, and E. Hovig, "Tumor Classification and Marker Gene Prediction by Feature Selection and Fuzzy C-Means Clustering Using Microarray Data," *BMC Bioinformatics,* vol. 4, no. 1, p. 60, Dec. 2003.

[21] T. Ando, M. Suguro, T. Hanai, T. Kobayashi, H. Honda, and M. Seto, "Fuzzy Neural Network Applied to Gene Expression Profiling for Predicting the Prognosis of Diffuse Large B-Cell Lymphoma," *Japan J. Cancer Research,* vol. 93, no. 11, pp. 1207-1212, Nov. 2002.

[22] H. Ressom, R. Reynolds, and R.S. Varghese, "Increasing the Efficiency of Fuzzy Logic-Based Gene Expression Data Analysis," *Physiological Genomics,* vol. 13, no. 2, pp. 107-117, Apr. 2003.

[23] C. Perez-Iratxeta, P. Bork, and M.A. Andrade, "Association of Genes to Genetically Inherited Diseases Using Data Mining," *Nature Genetics,* vol. 31, pp. 316-319, July 2002.

[24] J.C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms.* New York: Plenum Press, 1981.

[25] A.J. Enright, S. Van Dongen, and C.A. Ouzounis, "An Efficient Algorithm for Large-Scale Detection of Protein Families," *Nucleic Acids Research,* vol. 30, no. 7, 2002.

[26] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman, "Basic Local Alignment Search Tool," *J. Molecular Biology,* vol. 215, no. 3, pp. 403-410, 1990.

[27] J. Keller, M. Popescu, and J.A. Mitchell, "Taxonomy-Based Soft Similarity Measures in Bioinformatics," *Proc. IEEE Int'l Conf. Fuzzy Systems,* pp. 23-30, July 2004.

[28] D. Dubois and H. Prade, *Fuzzy Sets and Systems: Theory and Applications.* New York: Academic Press, 1980.

[29] E.M. Marcotte, M. Pellegrini, H.L. Ng, D.W. Rice, T.O. Yeates, and D. Eisenberg, "Detecting Protein Function and Protein-Protein Interactions from Genome Sequences," *Science,* vol. 285, pp. 751-753, 1999.

[30] J. Myllyharju and K.I. Kivirikko, "Collagens, Modifying Enzymes and Their Mutations in Humans, Flies and Worms," *Trends in Genetics,* vol. 20, no. 1, pp. 33-43, 2004.

[31] J.-M. Claverie, "Computational Methods for the Identification of Differential and Coordinated Gene Expression," *Human Molecular Genetics,* no. 8, pp. 1821-1183, 1999.

Mihail Popescu received the MS degree in medical physics in 1995, the MS degree in electrical engineering in 1997, and the PhD degree in computer science in 2003 from the University of Missouri at Columbia. He is now a National Library of Medicine postdoctoral fellow in Bioinformatics at the University of Missouri at Columbia. His main research interests are automatic functional annotation of gene products and microarray processing. He is a member of the IEEE.

James M. Keller received the PhD degree in mathematics in 1978. He is currently a professor in the Electrical and Computer Engineering Department at the University of Missouri-Columbia. He is also the R.L. Tatum Research Professor in the College of Engineering. His research interests center on computational intelligence: fuzzy set theory and fuzzy logic, neural networks, and evolutionary computation with a focus on problems in computer vision, pattern recognition, and information fusion, including bioinformatics, spatial reasoning in robotics, sensor and information analysis in technology for eldercare, and landmine detection. Dr. Keller has coauthored more than 225 technical publications. He is a fellow of the IEEE and a distinguished lecturer for the IEEE Computational Intelligence Society (CIS) and for the ACM. He is a past president of the North American Fuzzy Information Processing Society (NAFIPS), a past editor-in-chief of the *IEEE Transactions on Fuzzy Systems*, and was the general chair for the 2003 IEEE International Conference on Fuzzy Systems. He is currently the vice president for Publications for CIS.

Joyce A. Mitchell received the PhD degree in population genetics from the University of Wisconsin, Madison. She is a professor and chair of the Department of Medical Informatics at the University of Utah at Salt Lake City. She is an elected fellow of the American College of Medical Informatics and a founding fellow of the American College of Medical Genetics. Her current research interests are focused on clinical bioinformatics.