

Gene ontology for measuring similarity

T-61.6070 Modeling of proteomics data, TKK
Seminar presentation 30.4.2008
Lauri Lahti

Based on:

Popescu, M., Keller, J., & Mitchell, J. (2006). Fuzzy measures on the gene ontology for gene product similarity. *IEEE/ACM Transactions on computational biology and bioinformatics*, vol. 3, no. 3, July-September 2006.
Online: <http://portal.acm.org/citation.cfm?id=1152960>

Sequencing

- Sequencing genomes has produced need for tools to assist in analysis of similarities between genes and among gene families
- Genes are grouped in various ways:
 - being part of gene families
 - being part of a metabolic pathway
 - being coregulated under various conditions
- Protein isoforms are produced from the same mRNA transcript but with alternate splicing

Similarity

- In analysis of similarity between gene products the DNA sequence and the expression values are obvious features to evaluate
- However, for many gene products, additional *symbolic* information is available:
 - Gene Ontology (GO) terms
 - terms from a thesaurus used to index the publications about the gene or gene product
- These symbolic features can be incorporated into *gene similarity functions*

Gene similarity functions

- Gene similarity functions should
 - *maximize* common supportive evidence (especially contained in the ontologic or taxonomic structure)
 - *minimize* the effect of ambiguity and/or incomplete annotations
- Popescu et al. (2006) describes novel similarity measures for gene product comparison: the FMS and Choquet
- They are based on *fuzzy measures* and *fuzzy set theory*
- Fuzzy theories have proved to be effective in many domains

Similarity based on taxonomy

Two categories of approaches to compute the similarity of two objects described by sets of terms belonging to a taxonomy:

- 1) The terms in the sets are considered *individually*
 - A) Pair-based approach
 - B) Setbased approach

- 2) The similarity measures use *graph similarity techniques*

These are discussed in the following

Approach 1A

Computing the similarity when the terms in the sets are considered individually:

Pair-based approach

- Aggregating the similarities between all pairs of terms from the two sets.
- The pairwise similarities are aggregated using a function such as maximum or average.

Approach 1A (cont.)

Example (Cao et al. (2004)):

- pairwise similarity between gene ontology terms was used to search multiple biological databases
- similarity was computed using *the information content* of a gene ontology term
- the information content of a concept c can be measured as negative the log likelihood, that is $-\log p(c)$
- as probability increases, informativeness decreases

Example (Ganesan et al. (2003)):

- Optimistic Genealogy Measure (OGM), a similarity measure involving combination between average and maximum
- based on the depth in the hierarchy, to compare different customers based on their buying behavior

Approach 1B

Computing the similarity when the terms in the sets are considered individually:

Setbased.approach

- So called “bag of words” approach
- The similarity is computed using set similarity measures such as Dice, Jaccard, or cosine
- Used actively in web content data mining

Approach 1B (cont.)

Example (Ganesan et al. (2003)):

- A generalization on the cosine measure based on the depth in the hierarchy

A general problem with the depth-based similarity:
the distance in a taxonomy is not uniform due to the variation in density of the various subtaxonomies

Approach 2

Computing the similarity with measures that use *graph similarity techniques*

- 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 typical for 3D structure matching, MESH-based
- document retrieval, 2D shape recognition, multiagent systems, natural language processing, database search etc.
- In the general case, this problem is NP-complete but various techniques allow computing the similarity in polynomial time

Fuzzy measures

- Fuzzy measures have not been used extensively in bioinformatics.
- For *microarray analyses* there has been use of fuzzy techniques such as fuzzy clustering, fuzzy neural networks, fuzzy rule systems, and fuzzy relations
- Also in bioinformatics in applications related to *document content analysis* similar techniques have been used

The article extends the earlier work (especially Lord et al. (2003)) who investigated semantic similarity measure to explore the gene ontology. The new fuzzy measures are compared 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 gene ontology.

The article tries to 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 Medical Subject Headings (MeSH).

The proposed fuzzy measure similarities address inconsistencies and inabilities of existent numeric comparisons used for gene products.

Basic Local Alignment Search Tool (BLAST) scores do not account for the functions of the proteins, as do annotation-based similarities.

Second, cardinality-based measures (that are based on number of elements of the set, such as Jaccard and Dice) ignore the information content of the annotation terms in their construction.

A frequently used annotation term (given by for example binding cassettes) could artificially make two gene products look more similar than they actually are.

Finally, the average of pairwise term information content 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 “challenging domain” are very similar.

The proposed new measures do not have these inconsistencies.

Given two gene products, G_1 and G_2 , we can consider them as being represented by collections of terms $G_1 = \{T_{11}, \dots, T_{1i}, \dots, T_{1n}\}$ and $G_2 = \{T_{21}, \dots, T_{2j}, \dots, T_{2m}\}$.

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

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

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

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

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

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 $v_i \in \mathbb{R}^{NT}$, 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:

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

where $v_1 \bullet v_2$ 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 vector 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.

However, if $NT \gg 0$ (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 $s_{ij}(T1_i, T2_j)$, and then the values for the pairs are aggregated using, for example, the average as:

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

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 ($s_{avg}(G_1; G_1) < 1$) if $m, n > 1$.

Without some kind of normalization 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.

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

$$SWJ(G_1, G_2) = \frac{\sum_{\{i|T_i \in G_1 \cap G_2\}} g^i}{\sum_{\{j|T_j \in G_1 \cup G_2\}} g^j}.$$

The measures proposed in this article 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).

The authors 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).

The gene products (and their) families were retrieved on 10 December 2003 using the ENSEMBL browser (<http://www.ensembl.org>).

These three gene families were chosen for several reasons.

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.

All in all, the sample had a range of similarities between genes and gene products.

To validate the similarity measures, the authors began by computing the correlation between the new measures and a sequence-based similarity using:

$$s(\text{seq}_i, \text{seq}_j) = \frac{\ln s_{\text{raw}}(\text{seq}_i, \text{seq}_j)}{\min\{s_{\text{raw}}(\text{seq}_i, \text{seq}_i), s_{\text{raw}}(\text{seq}_j, \text{seq}_j)\}},$$

where s_{raw} is the natural logarithm of the BLAST bit score between seq_i and seq_j . The outcome was indicating high similarity (low distance).

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 $G = \{T_1, \dots, T_n\}$ be a finite set of terms describing a gene 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 $A \subseteq B$.

Here 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 (1977). A fuzzy measure g is called a *Sugeno measure* (g_λ -fuzzy measure) if it additionally satisfies the following property:

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


$$g_\lambda(A \cup B) = g_\lambda(A) + g_\lambda(B) + \lambda g_\lambda(A)g_\lambda(B)$$

for some $\lambda > -1$.

If the densities are known, the value for any Sugeno fuzzy measure can be uniquely determined for a finite set G using

$$G = \bigcup_{i=1}^n \{T_i\} \text{ and } g_\lambda(G) = 1,$$

which leads to solving the following equation for:

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

For the current application, the set of fuzzy density values is constructed from the information sources in the set G

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.

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

Then, the density value is defined by

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

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

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

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

Similarity calculations for two gene products from the same family: Consider the sequence G1 with GenBank ID AAH35609 (MTMR4 gene) and the sequence G2 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

$G_1 = \{T_1 = 4721(\text{protein phosphatase activity}),$

$T_2 = 6470(\text{protein amino acid dephosphorylation}),$

$T_3 = 8270(\text{zinc ion binding})\}$

$G_2 = \{T_1 = 4721(\text{protein phosphatase activity}),$

$T_2 = 6470(\text{protein amino acid dephosphorylation}),$

$T_4 = 16787(\text{hydrolase activity})\}.$

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

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

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 G1 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 G2 has $\lambda = -0,72$, resulting in $g_2(\{T_1, T_2\}) = 0,88$. Hence, the FMS similarity, sFMS, is:

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

Similarity values can be compared 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

The above two myotubularin genes should have high similarity since they belong to the same ENSEMBL myotubularin family. From the table, 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 this 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.

Another example describing similarity calculations for two gene products from different families. Now the sequence G1 with GenBank ID AAC12865 (MTMR2 gene) and the sequence G2 with GenBank ID AAF59902 (COL5A3 gene).

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

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 $G1 \cap G2 = \emptyset$ 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.

The gene ontology is a directed acyclic graph where a child node is considered a more specialized object than the parent node.

Lets still assume that the objects in the gene ontology 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.

The idea of the proposed method is to augment each set as:

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

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

where $\{T_{1i,2j}\}$ is the set of nearest common ancestors (NCA) of every pair $\{T1i, T2j\}$

The augmented FMS (AFMS), denoted by $s_{AFMS}(G_1; G_2)$, is defined as:

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

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

Example

Augmented FMS 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

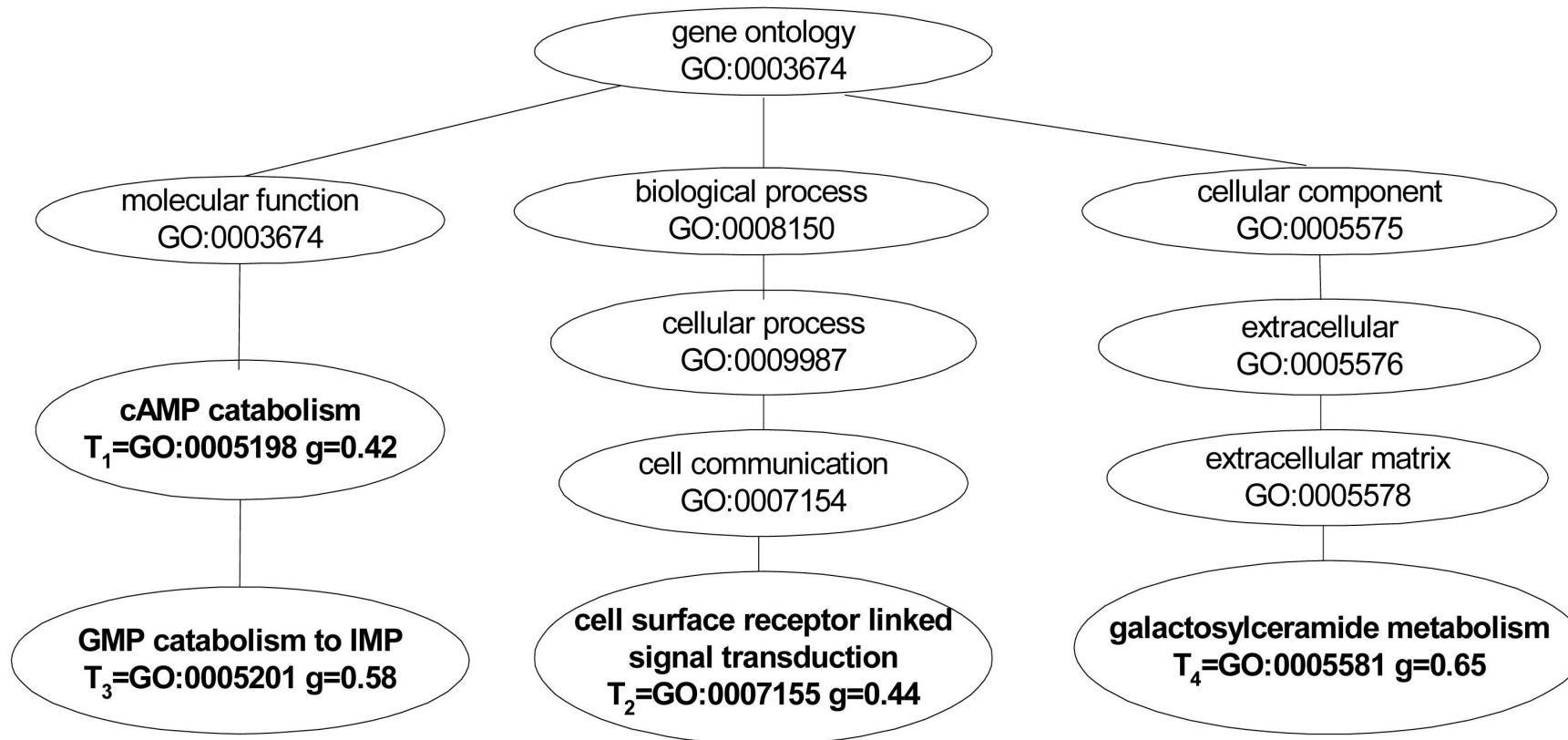
$$G_2 = \\ \{T_3 = 5201(\text{extracellular matrix structural constituent}), \\ T_4 = 5581(\text{collagen})\}.$$

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

$$G_2 =$$

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

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 T3 is a child in the GO of T1.



The augmented sets are: $G_1^+ = \{T1, T2\}$ and $G_2^+ = \{T1, T3, T4\}$
 Since nearest common ancestors $NCA(T3) = T1$ (see diagram)
 and the root node is ignored because its information content
 is 0 (common for all terms), the augmented intersection is
 $[G_1 \cap G_2]^+ = \{T3\}$

Hence, the augmented FMS is:

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

It can be calculated for the same case that 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.

Fuzzy integrals have been shown to be very useful for evidence fusion.

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 gene ontology terms, suppose that

$X = G1 \times G2$ and $s : X \rightarrow [0; 1]$ be a similarity function, i.e.,

$s_{ij}(T1_i, T2_j)$ is the similarity between the pair of gene ontology terms $(T1_i, T2_j)$.

To simplify the notation, we reorder the term pairs and label them by a single subscript so that $X = \{T1, T2, \dots, T_{nm}\}$.

The elements of X (pairs of gene ontology terms) are considered to be sources of information that support the similarity of genes $G1$ and $G2$ to degree $s(T_k)$, where $T_k = (T1_i, T2_j)$ for some i and j .

The *Choquet similarity* can be computed as follows:

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

where the function values are reordered so that

$$\begin{aligned} s(T_{(1)}) &\geq s(T_{(2)}) \geq \dots \geq s(T_{(nm)}), \\ s(T_{(nm+1)}) &= 0, \\ S_i &= \{T_{(1)}, \dots, T_{(i)}\}, \end{aligned}$$

and g is the *fuzzy measure* generated by the set of fuzzy densities $\{c_{ij}\}$ describing confidence of a pair of terms.

The above formulas can produce for example values that are between the average and maximum measures and depends on the reliability values assigned to the sources of annotation.

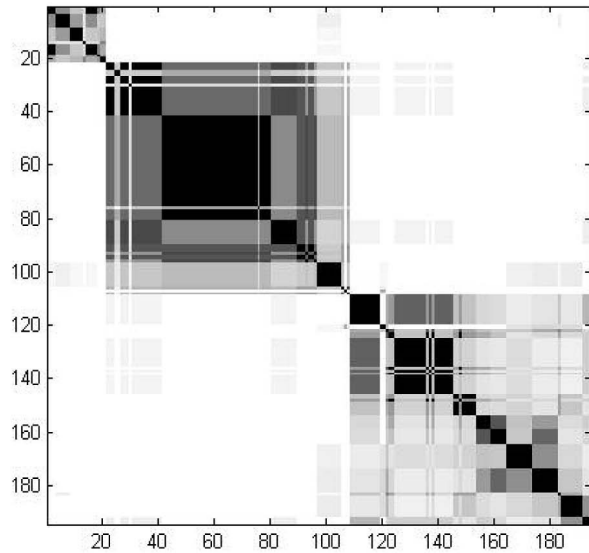
The *underlying hypothesis* is that using annotation uncertainty (reliability) can help us model part of our uncertainty about the similarity of the two sequences.

As the knowledge of various components of the *gene ontology 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.

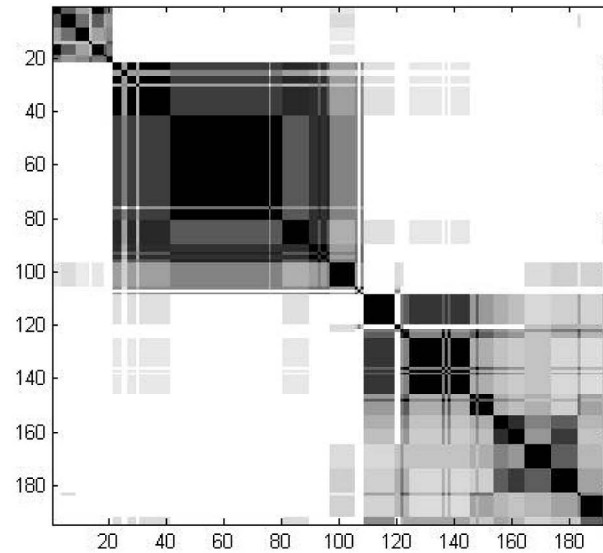
This is particularly useful in a situation like *gene function* where the knowledge is changing rapidly.

Next it is interesting to validate the proposed *gene ontology similarity* measures by investigating its correlation to sequence-based similarity measures.

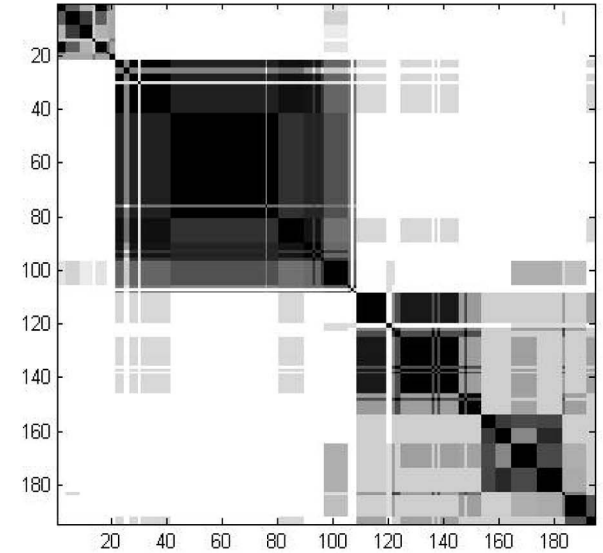
Gene ontology similarity measures were observed between Myotubularin, Receptor Precursor, and Collagen Protein families



(a)



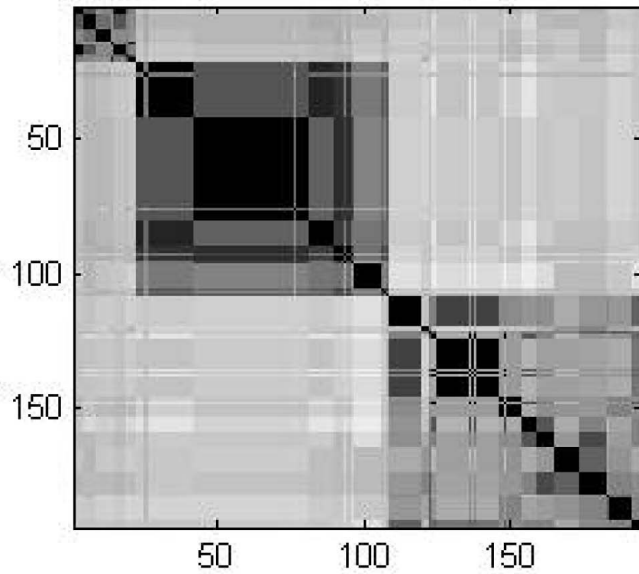
(b)



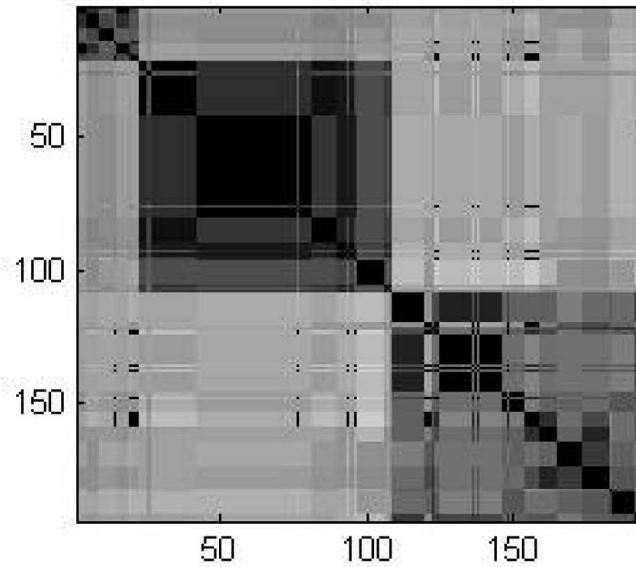
(c)

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

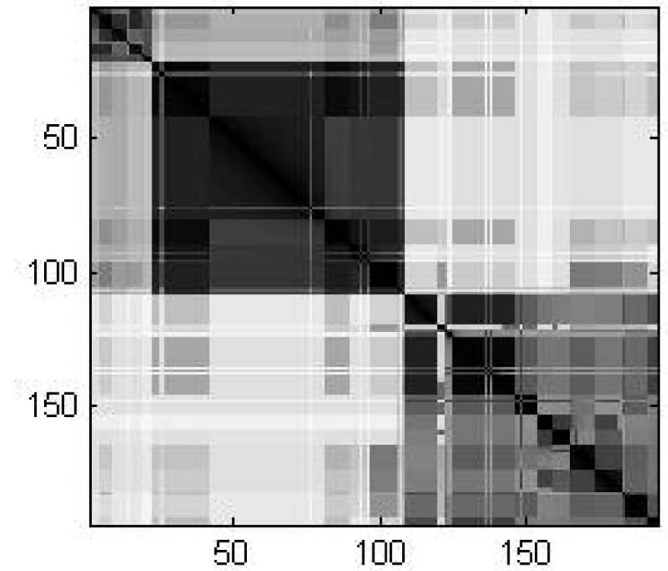
In general, the similarity among the members of the same family is high while the similarity between families is low for all three similarity measures. Since most sequences that belong to the same gene have similar annotation, we expect to see dark squares on the diagonal of the similarity matrix.



(a)



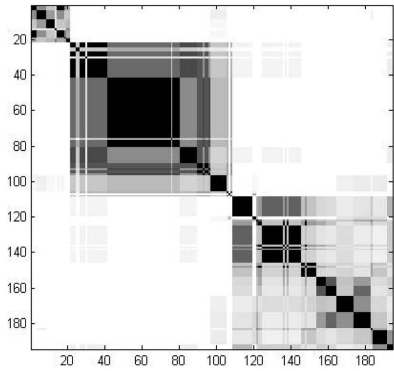
(b)



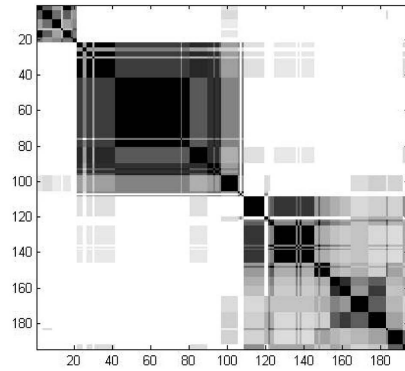
(c)

(a) Augmented Jaccard similarity. (b) Augmented cosine similarity.

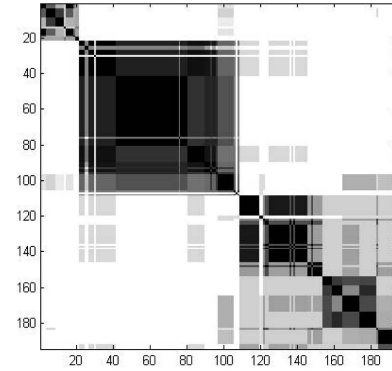
(c) Augmented fuzzy measure similarity.



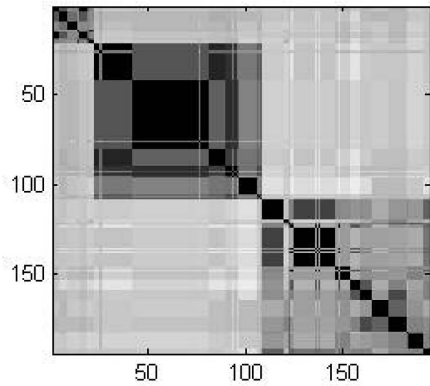
(a)



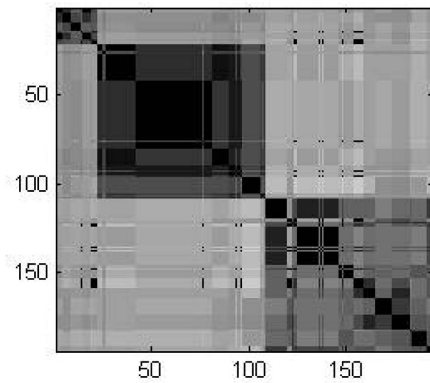
(b)



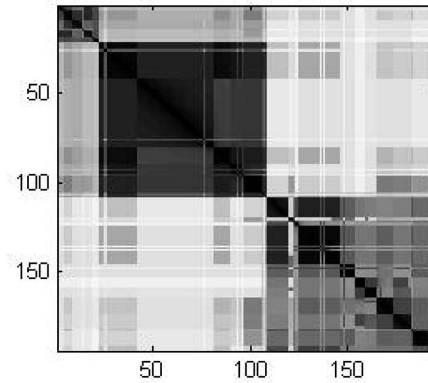
(c)



(a)



(b)



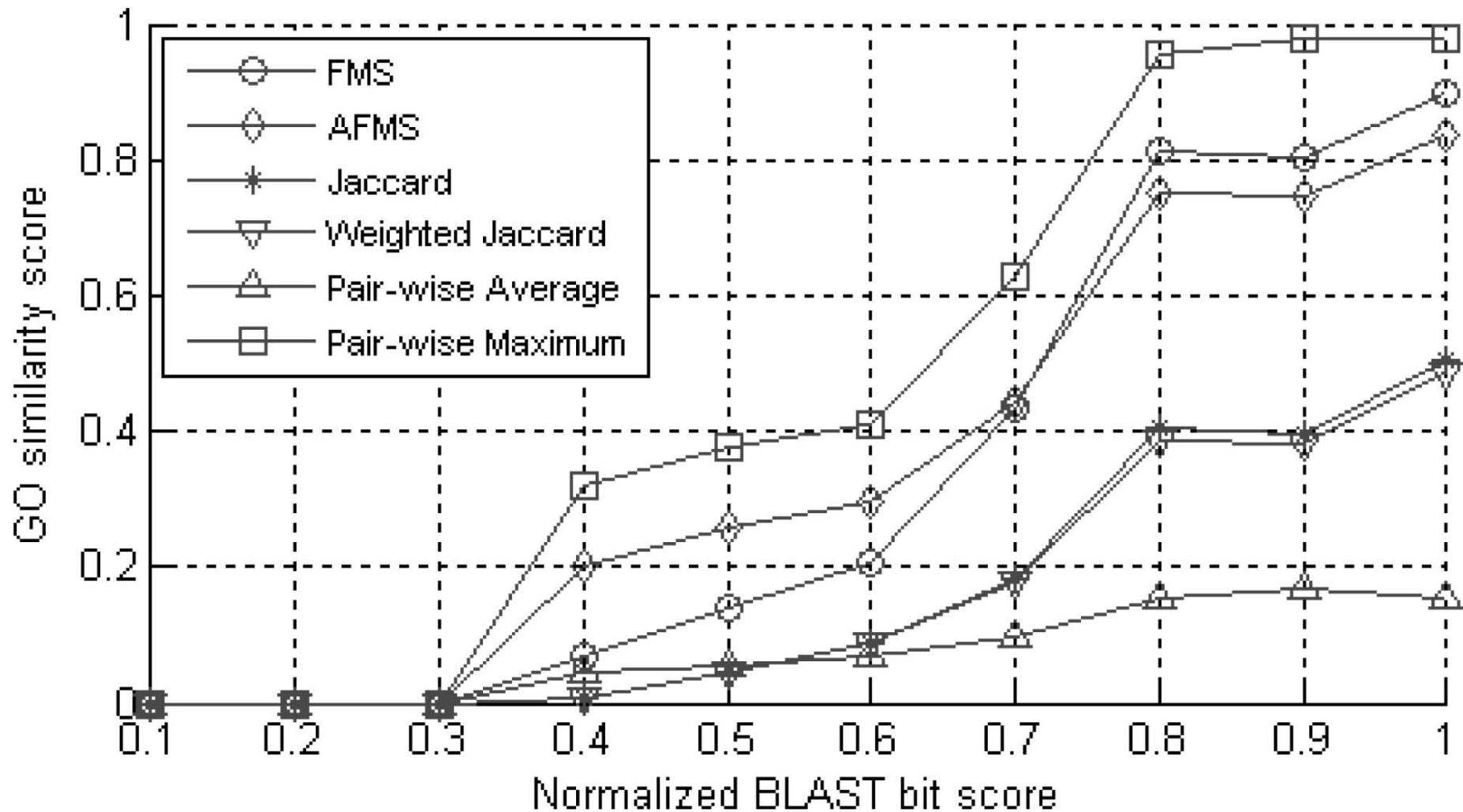
(c)

Comparing two previous figures, 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.

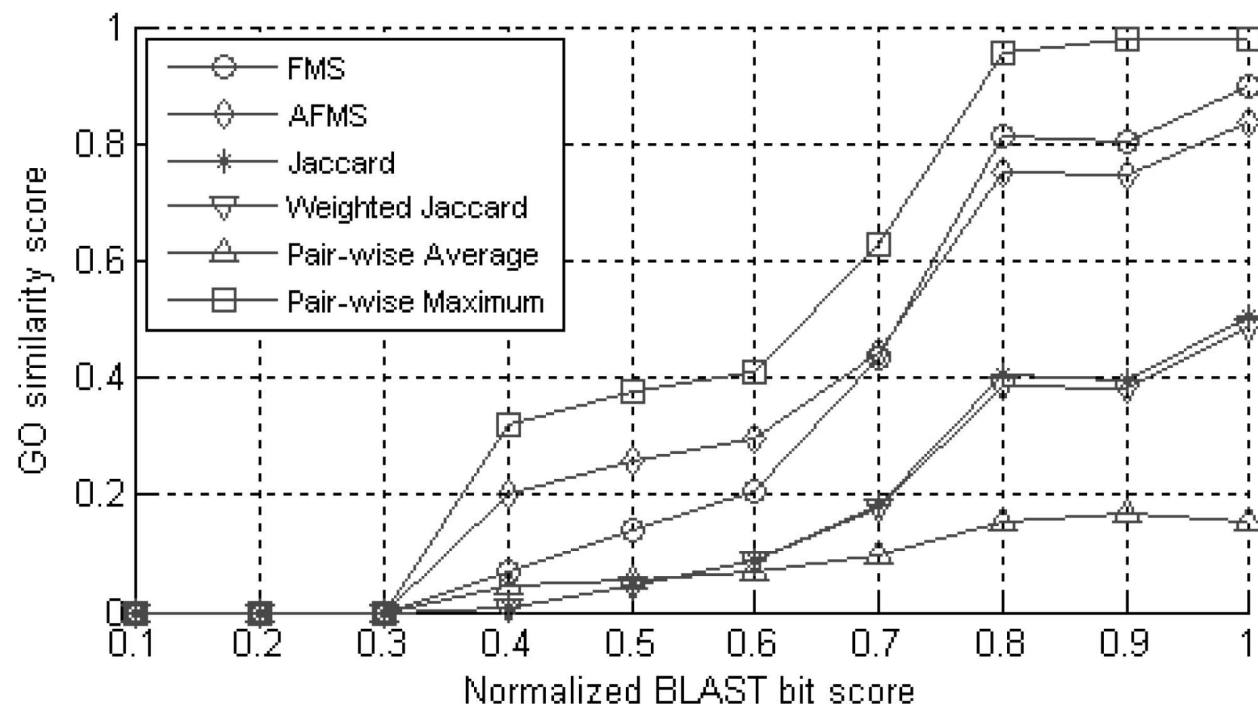
This stronger *within family similarity* should produce more consistent agreement with the extracted ENSEMBL families during clustering.

Also there is now more similarity between the three families since the augmentation procedure takes high-level genetic functions and processes into account.

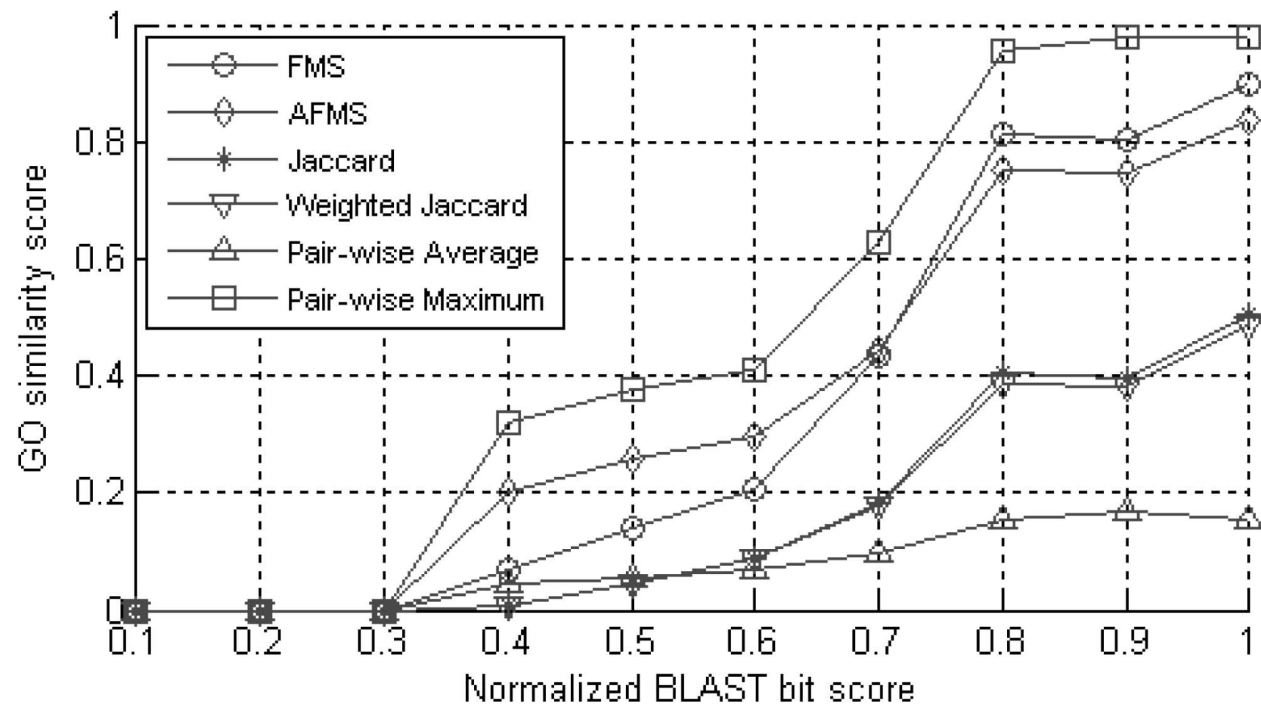
To quantitatively assess the GO similarities, we compute the correlation between the GO similarities and the sequence similarity.



Gene ontology average similarity score versus normalized BLAST bit score



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).



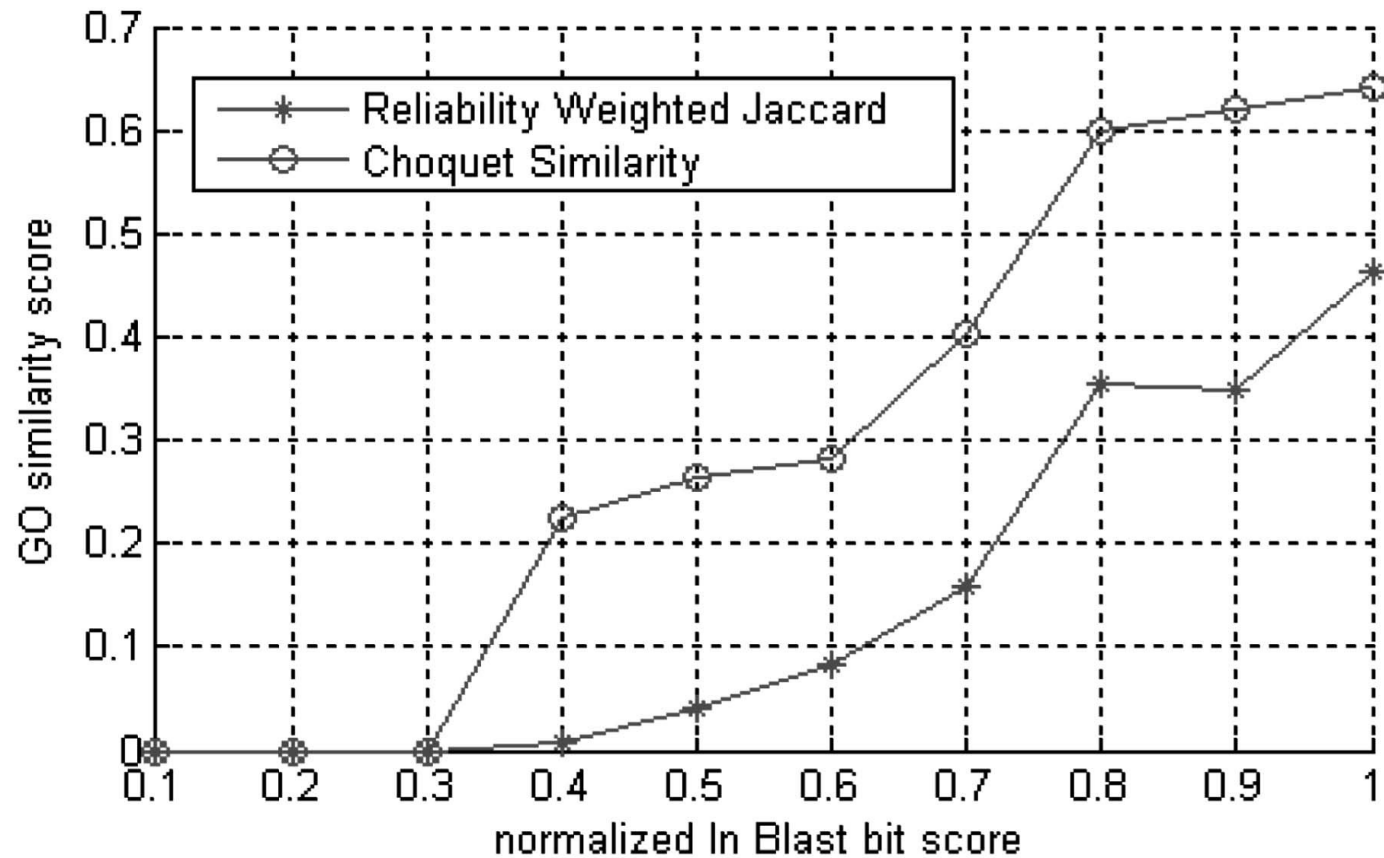
Second, it does not appear that the weighted Jaccard similarity is better than the unweighted Jaccard. Third, we see that the AFMS is better correlated to BLAST than FMS for values lower than 0.7. This is due to the extra nonzero values that AFMS produces in those cases where the intersection of the two annotation sets is empty. At high values, this effect is not present anymore.

The best correlation was obtained by the augmented fuzzy measure similarity (AFMS), although it is not striking.

Since we know the family assignment, we can compare the GO similarity to the ideal-case similarity matrix defined as:

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

For this case, the correlation coefficients can be presented as follows.



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

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.

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

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

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.

Conclusions

The article investigated several novel measures that can be used to assess the similarity of two gene products based on the GO terms describing them.

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.

Authors also proposed a method based on the Choquet integral to include the quality (reliability) of the annotation in the similarity measure. They showed that the proposed similarities correlate better to BLAST than the previously used approaches, average and maximum pairwise similarity.

The FMS could be beneficially used in conjunction with other sequence-based similarity tools (such as BLAST) to improve clustering and knowledge discovery in gene product databases.

Questions:

1. What is the role of fuzzy methods in this article?
2. What is the significance of the gene ontology in this approach?
3. Go through all examples in the article and see if you agree. Try to summarize one of them.
4. Please name some other alternative approaches for measuring similarity of genes.

The article is available online at

<http://portal.acm.org/citation.cfm?id=1152960>

Thanks for your time!