01 Oct 2007

Determining Domain Similarity and Domain-Protein Similarity using Functional Similarity Measurements of Gene Ontology Terms

Lisa Michelle Guntly

Jennifer Leopold
Missouri University of Science and Technology, leopoldj@mst.edu

Anne M. Maglia
Missouri University of Science and Technology

Recommended Citation

L. M. Guntly et al., "Determining Domain Similarity and Domain-Protein Similarity using Functional Similarity Measurements of Gene Ontology Terms," *Proceedings of the IEEE 7th International Symposium on Bioinformatics and Bioengineering (2007, Boston, MA)*, pp. 1209-1213, Institute of Electrical and Electronics Engineers (IEEE), Oct 2007.
The definitive version is available at https://doi.org/10.1109/BIBE.2007.4375717

Determining Domain Similarity and Domain-Protein Similarity Using Functional Similarity Measurements of Gene Ontology Terms

Lisa M. Guntly
Department of Computer Science
University of Missouri-Rolla
Rolla, MO 65409, USA

Jennifer L. Leopold
Department of Computer Science
University of Missouri-Rolla
Rolla, MO 65409, USA

Anne M. Maglia
Department of Biological Sciences
University of Missouri-Rolla
Rolla, MO 65409, USA

***Abstract*—Protein domains typically correspond to major functional sites of a protein. Therefore, determining similarity between domains can aid in the comparison of protein functions, and can provide a basis for grouping domains based on function. One strategy for comparing domain similarity and domain-protein similarity is to use similarity measurements of annotation terms from the Gene Ontology (GO). In this paper five methods are analyzed in terms of their usefulness for comparing domains, and comparing domains to proteins based on GO terms.**

Keywords-protein; domain; Gene Ontology

I. INTRODUCTION

Protein domains are highly functional sites in a protein that perform very specific functions across multiple proteins. Sequence comparison methods and structural comparison methods are often used to discover domains within a set of protein sequences. Despite the fact that domain sequences can vary widely in length (e.g., from 25 to 500 amino acids), the results of these two methods are usually in agreement [1].

Because domains are identified based on high similarity in sequence or structural measurements, it is not surprising that specific instances of a domain are relatively similar to each other in sequence, while different domains tend to demonstrate very little sequence similarity. Thus, comparing domains based on sequence does not provide much novel information. Improvement in sequence comparison and alignment tools has increased significantly the number of known domains. Given the increasing amount of data on domains and proteins that is being discovered, efficient computational methods are needed to investigate domains and their connections to proteins. Comparing functional data is a useful way to look at differing domains and the proteins that contain them. One such approach is to measure the similarity between annotation terms that describe proteins, an approach that has previously been used [2] to examine gene product similarity.

The Gene Ontology (GO) is one of the most useful sources for functional annotation information about genes and gene products. This controlled vocabulary has been extended to describe other functional biological units, such as domains and motifs. A given domain or protein may be annotated with multiple terms, reflecting the ability of domains and proteins to be involved in multiple processes, perform a variety of alternative functions, and have varying locations within a cell. The information and organizational structure provided by this ontology, and the extensive annotations of domains and proteins, stands as a useful foundation for considering domains and domain-protein similarity. Also of use in measuring the similarity of annotation terms assigned to proteins is the InterPro database, which includes links to other database records, structural data, GO annotations, and publication data.

In this paper we examine the usefulness of employing various similarity measurement methods that utilize GO annotation terms (including those from InterPro) to compare domains, and to compare domains to proteins. The advantages and disadvantages of these methods are discussed with respect to a data set of three zinc finger domains for domain comparison, and a separate zinc finger domain with two associated proteins for domain-protein comparisons.

II. BACKGROUND

Techniques to compare similarity of groups of words have largely developed from, and been applied to, text document similarity and searching techniques. For example, Internet searching takes advantage of similarity measurements in ranking results. Many researchers have begun to augment such similarity measurements to take into account the value of documents [3]. Since the value, or information content, of a GO term could affect similarity measurements, similarly augmented or different measurements should be explored in connection to GO data.

Traditional similarity measurements have several problems that need to be considered when applying them to domain and domain-protein relationships. The first issue is inaccuracy in the form of underestimation or overestimation. Pairwise similarity measurements exhibit this flaw most noticeably.

The second issue with standard similarity measurements in relation to GO terms is the lack of consideration for the information content of a term. GO terms that appear rarely in a set of annotations can be considered to have higher information content than terms that are common. Both pairwise and set similarity measurements are lacking in this respect. However, fuzzy measure similarity can be used to incorporate the information content of a GO term into the similarity calculation [2]. Other methods that include (or can be adjusted to include) information content as a component of the similarity measurement have been studied as well [4].

1-4244-1509-8/07/$25.00 ©2007 IEEE

A third concern with traditional similarity measurements is the inability to exploit the organizational structure of GO to determine if two terms are closely related, but not an exact match. This leads to zero similarity measurements in cases where some similarity should exist. Modifying fuzzy measure similarity to address this problem in connection to gene products has shown the usefulness of incorporating such information [2]. But inaccurate zero similarity results have yet to be thoroughly investigated.

A fourth concern is the reliability of a given annotation. Because all of the InterPro annotations are done manually, no difference exists in their reliability. However, if these similarity measurements are expanded to other databases that use various methods to determine GO annotations, reliability of the annotation becomes another information source that should be incorporated into the similarity measurement.

The information provided by GO has been used successfully to calculate similarity of gene products through a number of similarity measurements in other studies [2, 4]. Of the similarity measures tested on gene products, fuzzy measure similarity was found to correlate the best with BLAST sequence similarity [2]. Herein we examine the use of these measurements for domain comparisons and domain-protein comparisons.

III. METHODS

Domains and protein data used for our comparison were taken from the InterPro database. Domains were selected from records classified as domains, and corresponding proteins were selected from the list of entries in which the domain was found. Only domains and proteins with GO annotations available in the InterPro database were considered for this initial data set. A set of three zinc finger domains was selected for preliminary testing in domain comparison, and a separate zinc finger domain with two associated proteins was selected to test domain-protein similarity. Two sample zinc finger domains and their GO terms are shown below. These will be used as examples for the similarity measurements that will be examined in this section.

IPR000967 (Zinc finger, NF-X1-type)
D_1 = {T_1: 6355 (regulation of transcription),
T_2: 3700 (transcription factor activity),
T_3: 8270 (zinc ion binding),
T_4: 5634 (nucleus)}.

IPR000197 (Zinc finger, TAZ-type)
D_2 = {T_1: 6355 (regulation of transcription),
T_2: 3712 (transcription cofactor activity),
T_3: 8270 (zinc ion binding),
T_4: 5634 (nucleus)}

Both D_1 and D_2 are zinc finger domains, suggesting that some similarity in function, and thus in GO terms, should exist.

A. Pairwise Similarity

Pairwise similarity considers two sets of terms in pairs, $s_{ij}(T_{1i}, T_{2j})$. The average pairwise similarity measurement is:

$$s_{AVE}(D_1, D_2) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} s_{ij}}{mn} \quad (1)$$

For the example set of domains, s_{AVE} was computed as:

$$\sum_{i=1}^{n}\sum_{j=1}^{m} s_{ij} = 3$$
$$mn = 16$$
$$s_{AVE} = .19$$

The result of an average pairwise similarity measurement is actually an underestimation of similarity. This can be seen clearly when considering self-similarity, which is less than one if the number of terms in the sets is greater than one. A maximum pairwise estimate could be used to compute similarity. However, it would overestimate the similarity to the extent that its value is negligible. For example, in two sets of terms with at least one term in common, the maximum pairwise similarity would always be 1.

B. Set Similarity

Set similarity measurements differ from pairwise similarity measurements by considering the terms as an entire set, rather than in pairs. These methods largely avoid the underestimation and overestimation issues of pairwise similarity, suggesting that such methods would be of more use in considering domain and domain-protein similarity. One such method, Jaccard similarity, is defined as:

$$s_J(D_1, D_2) = \frac{|D_1 \cap D_2|}{|D_1 \cup D_2|} \quad (2)$$

For the example set of domains, s_J was computed as:

$$D_1 \cap D_2 = \{T_1, T_3, T_4\}$$
$$|D_1 \cap D_2| = 3$$
$$D_1 \cup D_2 = \{T_1, T_3, T_4, T_{21}, T_{22}\}$$
$$|D_1 \cup D_2| = 5$$
$$s_J(D_1, D_2) = .6$$

Another such method, Dice Similarity, is defined as:

$$s_D(D_1, D_2) = \frac{2|D_1 \cap D_2|}{|D_1| + |D_2|} \quad (3)$$

For the example domains, s_D was computed as follows:

$$2 * |D_1 \cap D_2| = 6$$
$$|D_1| = 4, \quad |D_2| = 4$$
$$s_D(D_1, D_2) = .75$$

While Jaccard and Dice similarity improve on the overestimation problem of pairwise similarity measurements for these data, they still fail to take into account useful information from the GO organizational structure.

C. Vector Space-Based Similarity

Vector space-based similarity does not consider pairs of terms, but instead converts the sets of terms to vectors. Cosine similarity can be used to translate the term sets into binary vectors. The size of the vectors is determined by the number of terms in $D_1 \cup D_2$. If a term is present in a domain's term set, the vector value is 1; otherwise, the value is 0. Other variations on this similarity method use weighted vectors, which could

facilitate dealing with the information content of a term, but this is not explored here.

Cosine similarity, is defined as:

$$s_V(D_1, D_2) = \frac{v_1 \bullet v_2}{|v_1||v_2|} \quad (4)$$

For the example set, the combined terms set is {T_1, T_2, T_3, T_4, T_5}. The vectors v_1 and v_2, determined from D_1 and D_2, respectively, are then translated as:

$$v_1 = [1 \quad 1 \quad 1 \quad 1 \quad 0]$$
$$v_2 = [1 \quad 0 \quad 1 \quad 1 \quad 1]$$

Thus, the cosine similarity is calculated as:

$$v_1 \bullet v_2 = 3$$
$$|v_1| = 2, \quad |v_2| = 2$$
$$s_V(D_1, D_2) = .75$$

D. *Fuzzy Measure Similarity*

The inclusion of information provided by an ontology structure allows fuzzy measurement similarity to include data on the information content of a term, and it can be adjusted based on the ontology structure to avoid inaccurate zero-similarity calculations. Fuzzy measure similarity is based on a generalization of a probability measure called a fuzzy measure, *g*, which must satisfy the following properties:

1. $g(\varnothing) = 0$ and $g(D) = 1$
2. $g(A) \le g(B)$ if $A \subseteq B$

The mapping of a term to a fuzzy measure is called a fuzzy density function and can be interpreted as the importance of the term or information source of the data. This mapping can be subjective, but with the incorporation of GO and the InterPro database for this problem, the densities can be determined from data on annotations of domains and proteins. A particularly useful category of fuzzy measure is the Sugeno fuzzy measure [5], which must satisfy the following additional property [6]:

3. For all $A, B \subseteq D$ with $A \cap B = \varnothing$.
$$g_\lambda(A \cup B) = g_\lambda(A) + g_\lambda(B) + \lambda g_\lambda(A) g_\lambda(B) \quad . \quad (5)$$
for some $\lambda > -1$

The subscript λ will be omitted from this equation in future references unless needed for clarification. With known densities, a unique λ for a Sugeno measure can be determined from (5), and the fact that D is the set of terms associated with a domain with $g(D) = 1$. The resulting equation for λ is:

$$(1 + \lambda) = \prod_{i=1}^{n} (1 + \lambda g^i) . \quad (6)$$

This equation has a unique solution for $\lambda > -1$ [2]. To ensure that two domains or proteins that share multiple GO terms are more similar than those that share one, $g(T_1) = g^1$ instead of $g(T_1) = 1$ is used if n = 1.

The fuzzy density values are determined using the approach described in [2], using the InterPro database and a script that computes the number of occurrences of a term and its children in the Gene Ontology. The following equation defines the fuzzy density calculated in this fashion [2]:

$$g^k = \frac{-\ln(p(T_k))}{\max_{T_j \in GO}\{-\ln(p(T_j))\}}, \quad (7)$$

The probability of finding a GO term or its children in the InterPro database was calculated as:

$$p(T_k) = \left(\frac{\text{count}(T_k) + \text{count}(\text{children of } T_k)}{\text{count}(\text{all GO terms in InterPro})} \right)$$

Fuzzy measurement similarity can then be defined as:

$$s_{FMS}(D_1, D_2) = \frac{g_1(D_1 \cap D_2) + g_2(D_1 \cap D_2)}{2} \quad (8)$$

For the example domains, these preliminary calculations were performed to collect the needed numbers for the fuzzy measurement similarity calculation. The first calculation determines the probability of finding the GO term or its children in an InterPro entry. The probabilities associated with the five unique GO terms in the example domains are listed in Table I.

TABLE I. PROBABILITIES AND DENSITIES OF GO TERMS

GO id	GO term	p(Tk)	Density (g^k)
6355	regulation of transcription, DNA-dependent	0.0131	0.44
3700	transcription factor activity	0.0076	0.5
8270	zinc ion binding	0.0077	0.5
5634	Nucleus	0.0196	0.4
3712	transcription cofactor activity	0.0005	0.78

The GO term 3712 (transcription cofactor activity) was the most specific of the terms in this set, as demonstrated by the low probability of finding the term or its children in the GO annotations of InterPro entries. The other GO terms had probabilities corresponding to their specificity as well. The densities (i.e., information content) of each of these terms was calculated using (7) and is displayed in Table I. As should be expected, the GO term 3712 had the highest density of the five terms, since it was the most specific. A low probability of finding the GO term or its children in the InterPro database indicates the term has a high information content/density. The other four terms had fairly similar information content. None of the GO terms in this set were highly unspecific, so the information content of all the terms was relatively high.

After determining the densities of the associated GO terms, the λ values for D_1 and D_2 were calculated using (6):

D_1: $(1 + \lambda) = (1+.44\lambda) * (1+.5\lambda) * (1+.4\lambda)$
$\lambda_{D1} = -.87$

D_2: $(1 + \lambda) = (1+.44\lambda) * (1+.5\lambda) * (1+.4\lambda) * (1+.78\lambda)$
$\lambda_{D2} = -.95$

With the λ values of D_1 and D_2 known, the Sugeno measures, g^1 and g^2 can be calculated on the intersection of D_1 and D_2. The Sugeno measure on a single term is equal to the density. The intersection of terms for the examples domains is $\{T_1, T_3, T_4\}$. For the example, the calculations to find the values of g^1 ($\{T_1, T_3, T_4\}$) and g^2 ($\{T_1, T_3, T_4\}$) are as follows:

$$g^1(\{T_1\}) = 0.44$$
$$g^1(\{T_3\}) = 0.5$$
$$g^1(\{T_4\}) = 0.4$$

$$g^1(\{T_1, T_3\}) = g^1 + g^3 + \lambda g^1 g^3$$
$$= 0.44 + 0.50 + -0.87(0.44 * 0.5) = 0.75$$
$$g^1(\{T_1, T_3, T_4\}) = g^{13} + g^4 + \lambda g^{13} g^4$$
$$= 0.75 + 0.4 + -0.87(0.75 * 0.4) = 0.89$$

$$g^2(\{T_1\}) = 0.44$$
$$g^2(\{T_3\}) = 0.5$$
$$g^2(\{T_4\}) = 0.4$$

$$g^2(\{T_1, T_3\}) = g^1 + g^3 + \lambda g^1 g^3$$
$$= 0.44 + 0.50 + -0.95(0.44 * 0.5) = 0.73$$
$$g^2(\{T_1, T_3, T_4\}) = g^{13} + g^4 + \lambda g^{13} g^4$$
$$= 0.73 + 0.4 + -0.95(0.73 * 0.4) = 0.85$$

The values of g^1 ($\{T_1, T_3, T_4\}$) and g^2 ($\{T_1, T_3, T_4\}$) can then be used in conjunction with (8) to determine the fuzzy measure similarity of the example domains as follows:

$$s_{FMS}(D_1, D_2) = \frac{g^1(\{T_1, T_3, T_4\}) + g^2(\{T_1, T_3, T_4\})}{2}$$
$$= \frac{.89 + .85}{2} = .87$$

IV. DOMAIN SIMILARITY

The five similarity measurements described in the previous section were applied to the following three domains:

IPR000967 (Zinc finger, NF-X1-type)
D_1 = {T_1: 6355 (regulation of transcription),
T_2: 3700 (transcription factor activity),
T_3: 8270 (zinc ion binding),
T_4: 5634 (nucleus)}

IPR002694 (Zinc finger, CHC2 type)
D_2 = {T_1: 6260 (DNA replication),
T_2: 3677 (DNA binding),
T_3: 8270 (zinc ion binding),
T_4: 3896 (DNA primase activity)}

IPR000197 (Zinc finger, TAZ-type)
D_3 = {T_1: 6355 (regulation of transcription),
T_2: 3712 (transcription cofactor activity),
T_3: 8270 (zinc ion binding),
T_4: 5634 (nucleus)}

Each of these is a zinc finger domain type, so some similarity in function should exist. The average pairwise similarity, Jaccard similarity, Dice similarity, cosine similarity, and fuzzy measure similarity of the zinc fingers are shown in Tables II-VI, respectively. All of the similarity measures indicate that D_1 and D_3 show the highest similarity, while the other pairs show a lesser amount of similarity. The average pairwise similarity for the domains is underestimated, as was expected from earlier analysis of the method. Corrections of the underestimation issues involved in average pairwise similarity could be attempted; however, the other methods provide similar enough information that the usefulness of a corrected average pairwise similarity is questionable. Results from Jaccard, Dice, and cosine similarities were fairly consistent for these data. The fuzzy measure similarity results for these zinc finger domains were higher than other similarity estimates. This measurement is conceivably more accurate though, due to the relatively high information content of multiple terms that were similar.

V. DOMAIN-PROTEIN SIMILARITY

The five similarity measurements were also applied to the following zinc finger domain (D_4) and two proteins (G_1 and G_2) in which the domain is found:

IPR001841 (Zinc finger, RING-type)
D_4 = {T_1: 5515 (protein binding),
T_2: 8270 (zinc ion binding)}

IPR011364 BRCA1
G_1 = {T_1: 6281 (DNA repair),
T_2: 3677 (DNA binding),
T_3: 8270 (zinc ion binding),
T_4: 5634 (nucleus)}

IPR012227 TNF receptor-associated factor TRAF
G_2 = {T_1: 7165 (signal transduction),
T_2: 42981 (regulation of apoptosis),
T_3: 8270 (zinc ion binding)}

The RING-type zinc finger domain is a special type of zinc finger that is thought to be involved in protein-protein interactions. The domain appears in six InterPro entries, only two of which are annotated with GO terms. One of these, the BRCA1 protein family, is a DNA damage repair protein. The other is the TNF receptor-associated factor TRAF protein family, which is significantly involved in a signal transduction pathway in cells. The RING-type zinc finger domain is found near the beginning of both of the proteins. Because the zinc finger domain is found in the proteins, some similarity should exist between the GO terms associated with each protein.

The results of the five similarity calculations between the zinc finger domain and proteins in which it is found are shown in Table VII. The similarity measures indicate that the zinc finger domain has a nearly equal similarity to both of the proteins. Again the average pairwise measurements are low, although closer to the other measurements than in the domain comparison. Jaccard similarity agrees with cosine similarity, showing fairly low similarity values. Fuzzy measure similarity is again the highest, because the GO term that the groups had in common is 8270 (zinc ion binding). This term has a relatively high density value which increases the fuzzy measure similarity. Because of the consideration of information content of the GO terms in fuzzy measure similarity, it is reasonable to expect results that differ from other measurements and are likely to be more accurate.

VI. DISCUSSION

Several trends from the domain and domain-protein similarity measurements that were calculated can be used to identify useful characteristics of the measurements. The first noticeable trend is that fuzzy measure similarity gave higher similarity scores in all cases. This is indicative of the relatively high information content in the similar GO terms, demonstrating the usefulness of included information content in a similarity measurement. Dice similarity was the closest of the other measurements to the high fuzzy similarity measurement scores. This suggests that Dice similarity may be the most useful of the set similarity measurements to consider for similarity of GO terms. However, Dice similarity does not scale with information content as fuzzy measure similarity does, meaning that the Dice scores can be artificially high for similar terms with low information content.

Average pairwise similarity showed little usefulness due to underestimation of similarity. Adjusting the average pairwise similarity with self-similarity data would improve the accuracy of the scores, but would provide no new information that set and vector-based similarity could not provide. The results indicate that the inclusion of information content in a similarity calculation can alter significantly the results. Thus, fuzzy measure similarity would be the most useful measurement for application of these types of data.

As the results show for our initial data set, differences exist between zinc finger domains. These domains were expected to have some similarity in function, which should be seen through similarity of GO terms. The similarity measurements calculated for the three domains confirmed this expectation. However, the similarity measurements also demonstrated that some domains classified as zinc finger domain types are more similar in function than others. This type of information could be of use for grouping domains.

Because domains often are identified originally as groups based on sequence, the lack of sequence similarity between various zinc finger domains is reasonable. However, this makes it difficult to compare domains. If domains related in function like zinc finger domains fail to demonstrate any sequence similarity, comparison of sequence similarity between domains is unlikely to produce useful results. Therefore, computing similarity based on function is of greater use to biologists interested in similarities between domains.

Comparison of domains to the proteins in which they are found provides additional information. The similarity between a domain and a protein in which it is found can be considered a measurement of the domain's significance to the protein's function. For an individual protein, the significance of a domain can be used to examine the importance and role of one domain in comparison to others. Since many proteins contain multiple domains, comparing their significance can give an idea of which sections of a protein are the most important to examine in more detail. The usefulness of measuring a domain's significance to protein function extends to comparing the same domain in multiple proteins. Information on how significant a domain is on average and across multiple proteins gives a more thorough understanding of which domains are particularly significant to protein function.

VII. FUTURE WORK

In future work we plan to explore methods that address considerations of inaccurate zero similarity scores and variable reliability in GO annotations. By including the most closely related parent term to all pairs of GO terms in each of the term sets, closely related terms that do not add to a similarity score in standard methods will show some similarity. The benefit of implementing this approach with a fuzzy measure similarity is that parent terms are less specific, and thus, have lower information content. As a result, the similarity between two related terms will be lower than if the terms were equivalent.

TABLE II. PAIRWISE SIMILARITY

	D1	D2	D3
D1	0.25	—	—
D2	0.06	0.25	—
D3	0.19	0.06	0.25

TABLE III. JACCARD SIMILARITY

	D1	D2	D3
D1	1	—	—
D2	0.14	0.1	—
D3	0.6	0.14	0.1

TABLE IV. DICE SIMILARITY

	D1	D2	D3
D1	1	—	—
D2	0.25	1	—
D3	0.75	0.25	1

TABLE V. COSINE SIMILARITY

	D1	D2	D3
D1	1	—	—
D2	0.25	1	—
D3	0.75	0.25	1

TABLE VI. FUZZY MEASUREMENT SIMILARITY

	D1	D2	D3
D1	1	—	—
D2	0.5	1	—
D3	0.87	0.5	1

TABLE VII. VARIOUS SIMILARITIES WITH DOMAIN 4 (ZINC FINGER RING-TYPE)

	Avg. Pairwise	Jaccard	Dice	Cosine	Fuzzy Measure
G1	0.13	0.2	0.33	0.2	0.5
G2	0.18	0.25	0.4	0.25	0.5

REFERENCES

[1] A. Marchler-Bauer, A. Panchenko, N. Ariel, and S. Bryant, "Comparison of Sequence and Structure Alignments for Protein Domains, " *Proteins: Structure, Function, and Genetics*, Vol. 48, No. 3, pp 439-446, 2002.

[2] M. Popescu, J. Keller, and J. Mitchell, "Fuzzy Measures on the Gene Ontology for Gene Product Similarity," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. Vol. 3, Issue 3, pp. 263 - 274, 2006.

[3] A. Paepcke, H. Garcia-Molina, G. Rodriguez-Mula, and J. Cho, "Beyond Document Similarity: Understanding Value-Based Search and Browsing Technologies," *SIGMOD Records*, Vol. 29, No.1, March 2000.

[4] P. Lord, R. Stevens, A. Brass, and C. Goble, "Semantic Similarity Measures as Tools for Exploring the Gene Ontology," *Pacific Symposium on Biocomputing*, Vol. 8, pp. 601-612, 2003.

[5] *Fuzzy Measures and Integrals: Theory and Applications,* M. Grabisch, T. Murofushi, and M. Sugeno, eds. Springer-Verlag, 2000.

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