Appears in Proceedings of the 3th Meeting of the Special Interest Group on Text Mining at ISMB2003

CURATING EXTRACTED INFORMATION THROUGH THE CORRELATION BETWEEN STRUCTURE AND FUNCTION

Francisco M. Couto, Mário J. Silva
Dept. of Informatics
Faculdade de Ciências da
Universidade de Lisboa
E-mail: {fjmc,mjs}@di.fc.ul.pt

Pedro Coutinho
UMR 6098, Architecture et Fonction des
Macromolécules Biologiques
Centre National de la Recherche Scientifique
E-mail: pedro@afmb.cnrs-mrs.fr

Abstract

We propose to apply the correlation between structure and function of gene products to curate information automatically extracted from biological literature. This can be achieved by automatically validating extracted information that satisfies the correlation, since it has strong evidence of being correct.

We applied a semantic similarity measure (SSM) to identify a correlation between the modular structures of glycoside hydrolases (GHs) and functional terms extracted from associated literature. The source of GHs was CAZy, a database of carbohydrate-active enzymes classified in various families by their modular structure. We retrieved literature associated with each GH. From this literature, we extracted Gene Ontology (GO) functional terms. We implemented a SSM on GO to measure the relatedness between the GO terms extracted. Finally, we identified the correlation by comparing the probability of extracting similar terms inside with outside a family.

Keywords: semantic similarity measure, text mining, gene ontology, modular structure

1 Introduction

Most of text mining systems evaluate their results by human curation or by exploiting existing databases [9, 1, 5]. In this paper, we propose another approach to curate extracted information that relies on the correlation between structure and function. Our approach is supported by the dogma of molecular biology that structure should be correlated with biological activity. For instance, when similar functional terms are automatically annotated with a significant number of gene products that are structurally related, we have strong evidence that these annotations are correct since gene products from a common family tend to share a common set of biological activities. Therefore, extracted information that satisfies the correlation has a large likelihood of being correct. This evidence can be used to automatically filter information before being human curated. To assure the feasibility of our approach, we measured the correlation in extracted information

To calculate the relatedness between functional terms, we implemented a semantic similarity measure (SSM). A SSM determines the degree of relatedness between two concepts expressed in a semantic network. Two kinds of approaches are prevalent in these measures: information content (node based) and conceptual distance (edge based). Information content considers the similarity between two terms the amount of information they share, where a term contains less information when it occurs very often. Conceptual distance is based on the shortest topologic distance between two terms in the scheme taxonomy. Five different proposed SSMs were experimentally compared in WordNet [2]. The comparison shows that Jiang and Conrath's SSM provides the best results overall [7]. This SSM is a hybrid approach, i.e., it combines information content and conceptual distance with some parameters that can be used to control the degree of each factor's contribution.

The information was extracted from biological literature that was associated with glycoside hydrolases (GHs). The source of GHs was CAZy database [3]. CAZy attributes each carbohydatre-active enzyme to one or more families of catalytic and carbohydrate-binding modules

	bibliographic references	distinct documents
GenBank	22849	4575
SwissProt	8998	4006
PDB	3561	785
Total		6377

Table 1: Number of items retrieved

according to its modular structure. From the literature, we extracted Gene Ontology (GO) terms and computed the relatedness between them using a SSM. GO structures a controlled vocabulary of gene and protein biological roles, held in a Directed Acyclic Graph (DAG). The three organizing principles of GO are the molecular function, the biological process and the cellular component. However, in this study, we have only used the molecular function principle. By having a set of GO terms associated with each family, we were able to identify a correlation by comparing the probability of extracting similar terms inside with outside a family.

2 Retrieval & Extraction

We developed an automatic process to identify bibliographic references and to extract their contents. The bibliographic references were identified from external databases (e.g. GenBank, SwissProt, PDB), since in CAZy each enzyme is associated with a set of entries to these external databases. Thus, each enzyme was associated with a set of bibliographic references. Table 1 presents the number of bibliographic references and the number of distinct documents they refer. We obtained the abstracts of these documents from PubMed.

The extraction method of GO terms is based on their occurrences in text, which is similar to the method used in some projects [6, 4, 10]. We assumed that if a document associated with an enzyme mentions a GO term in its abstract then there is an underlying biological relation between the enzyme and the GO term. Thus, we associated each enzyme with the GO terms extracted from its associated literature.

3 Semantic Similarity

To compute the relatedness between two GO terms based on Jiang and Conrath's SSM, we had to calculate the following factors: their closest common ancestor; the shortest path between each term and their common ancestor; and for each term in the paths its information content, its depth and the number of its direct descendents (i.e. local density). The common ancestor was identified through the transitive closure of the DAG. The information content computation is based on the terms' frequency. We computed the number of occurrences of each term in the corpus. However, if a term occurs then all its ancestor terms also occur. Thus, we propagated the term occurrences throughout the hierarchy, reaching a frequency for the root node (*GO:0003674 molecular_function*) equal to the sum of all the occurrences, i.e. it has no relevant information.

We considered the GO terms associated to a family as the union of all terms associated to its GHs. We measured the relatedness of terms inside and outside a family by computing the semantic similarity between terms associated to a family with each other and with all the extracted terms, respectively. We used a set of similarity thresholds to decide whether two terms were considered similar or not, given their semantic similarity value. This way, we were able to know the number of similar terms inside and outside a family for a given similarity threshold. These numbers were converted to probabilities, i.e., we computed the probability of extracting similar terms inside ($\mathcal{P}_{in}$) and outside ($\mathcal{P}_{out}$) a family.

4 Results

This section describes the results of our last analysis performed on the January 2003 release of GO and CAZy databases. We computed the relatedness between 343 GO terms, which were related to 9831 GHs classified in 90 families. The graphic in figure 1 shows the ratio of $\mathcal{P}_{in}$ over $\mathcal{P}_{out}$. Each value represents the average of this ratio for the families analyzed, using the similarity threshold that maximized the difference between $\mathcal{P}_{in}$ and $\mathcal{P}_{out}$. The parameters α and β control the degree of how much the node depth and density factors contribute to semantic similarity computation. These contributions become

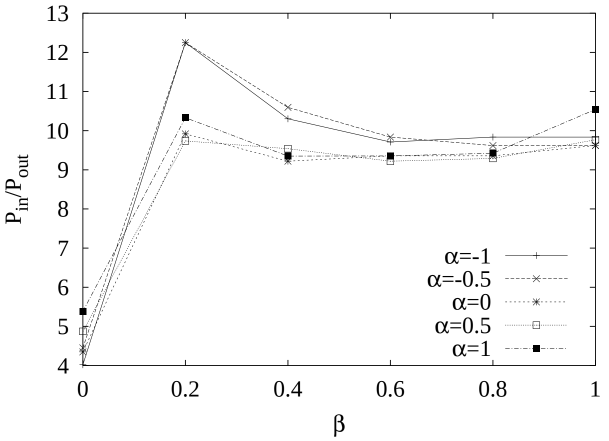

Figure 1: $\mathcal{P}_{in}$ over $\mathcal{P}_{out}$

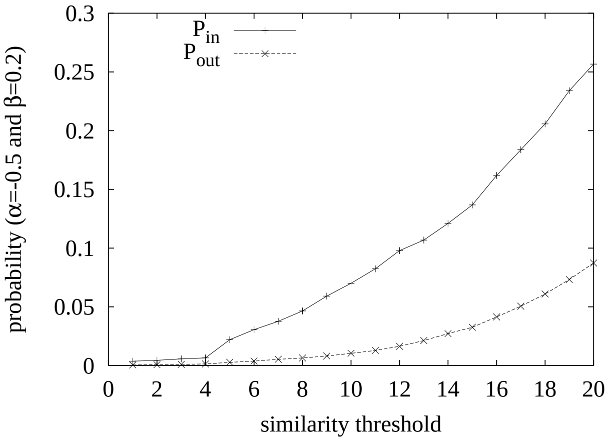

Figure 2: $\mathcal{P}_{in}$ against $\mathcal{P}_{out}$

less significant when α approaches 0 and β approaches 1. The values achieved show that the probability of extracting similar terms inside a family is significantly larger than outside it, as anticipated. The best result, more than 12 times larger, is when $\alpha{=}-0.5$ and $\beta{=}0.2$. This means that the density of the DAG and the depth of each node are important conceptual distance factors to amplify the correlation. In figure 2, the graphic shows $\mathcal{P}_{in}$ against $\mathcal{P}_{out}$ for different similarity thresholds when $\alpha{=}-0.5$ and $\beta{=}0.2$. As it was expected, both probabilities are proportional to the similarity threshold since a larger similarity threshold implies also a larger number of similar terms. The relevant fact in the graphic is that $\mathcal{P}_{in}$ is always significantly larger than $\mathcal{P}_{all}$, which shows that enzymes with similar modular structure tend to be annotated with similar functional terms.

5 Conclusions & Related Work

We automatically annotated GHs with GO terms extracted from biological literature. We applied a SSM to compute the probability of extracting similar terms inside and outside a GH family. The result was a larger probability inside than outside a GH family, which shows a correlation between structure and function. This quantitative measure of the correlation supports our approach of using it as an effective tool to curate information automatically extracted from biological literature. Because correlated information has strong evidence of being correct.

In related work, Lord also applied SSMs to GO [8], with the following main differences to our work: we presented results of an hybrid measure instead of an information content measure, concluding that information content can improve their results when integrated with conceptual distance; we correlated not the sequence but a modular structure classification; and we extracted automatically the GO terms from free text instead of using human curated annotations.

References

[1] C. Blaschke, L. Hirschman, and A. Valencia. Information extraction in molecular biology. *Briefings in BioInformatics*, 3:1–12, 2002.

[2] A. Budanitsky and G. Hirst. Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures. Workshop on WordNet and Other Lexical Resources, in the North American Chapter of the Association for Computational Linguistics (NAACL-2000), Pittsburgh, PA, June 2001.

[3] P. Coutinho and B. Henrissat. Carbohydrate-active enzymes: an integrated database approach. *Recent*

Advances in Carbohydrate Bioengineering, pages 3–12, 1999.

[4] J. D. et al. Mining MEDLINE: Abstracts, sentences, or phrases? In *PSB*, pages 326–337, 2002.

[5] L. Hirschman, J. Park, J. Tsujii, L. Wong, and C. H. Wu. Accomplishments and challenges in literature data mining for biology. *Bioinformatics*, 18(12):1553–1561, 2002.

[6] T. Jenssen, A. L. greid, J. Komorowski, and E. Hovig. A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics*, 28:21–28, may 2001.

[7] J. Jiang and D. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the 10th International Conference on Research on Computational Linguistics (ROCLING X)*, Taiwan, 1997.

[8] P.W.Lord, R. Stevens, A. Brass, and C.A.Goble. Semantic similarity measures as tools for exploring the Gene Ontology. In *Pacific Symposium on Biocomputing*, pages 601–612, 2003.

[9] S. Ray and M. Craven. Representing sentence structure in hidden markov models for information extraction. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence*, Seattle, 2001. WA. Morgan Kaufmann.

[10] B. Stapley and G. Benoit. Biobiblimetrics: Information retrieval and visualization from co-occurrences of gene names in MEDLINE abstracts. In *PSB*, pages 326–337, 2002.