ProFAL: PROtein Functional Annotation through Literature

Francisco M. Couto[1], Mário J. Silva[1], and Pedro Coutinho[2]

[1] LaSIGE, Departamento de Informática, Faculdade de Ciências, Universidade de Lisboa, Campo Grande, 1749-016 Lisboa, Portugal, {fjmc,mjs}@di.fc.ul.pt
[2] UMR 6098, Architecture et Fonction des Macromolécules Biologiques, CNRS, 13402 Marseille CEDEX 20, France, pedro@afmb.cnrs-mrs.fr

Abstract. We introduce ProFAL (PROtein Functional Annotation through Literature), a new information system for automatic annotation of biological databases using Bioinformatics methods. The annotations are (gene-product, functional property) pairs, associating the attributes of a gene-product, stored in the database, to functional properties. The system retrieves documents related to each gene-product from online databases and extracts functional properties from their text. To validate these annotations, ProFAL implements heuristics based on a measure of correlation between annotations that we have introduced. To verify the validated annotations, ProFAL also provides a specific interface for manual curation. We evaluate the implementation and performance of ProFAL in a case-study.

1 Introduction

As for most fields of scientific research, relevant facts discovered in molecular biology have mainly been published in scientific journals [8]. Extracting knowledge from this large amount of unstructured information is an arduous task, even for human experts. An improvement was the creation and maintenance of structured databases that collect and distribute biological information. Examples are the GenBank or SwissProt databases that describe properties of common biological entities, such as genes and proteins.

In the past few decades, the explosion of available genomic data triggered an exponential growth of these databases, causing a lack of annotations for many recent entries [2]. However, a substantial amount of knowledge important to characterize each biological entity is spread through a vast set of heterogeneous sources. The integration of different data sources is a viable approach to correct and complete our knowledge about these biological entities [16].

Motivated by this fact, we introduce in this paper ProFAL (PROtein Functional Annotation through Literature), a novel system for automatic annotation of biological databases using knowledge extracted from the scientific literature. The literature

is retrieved using external biological web resources that link gene-products to related documents. From the retrieved literature, ProFAL extracts functional properties of the gene-products stored in the biological database. Next, it validates the information extracted to reduce the workload of manual curation. To verify the annotations, ProFAL provides a specific web interface that presents the extracted information integrated with the information stored in the biological database. The database curators used the web interface to conduct an initial evaluation of ProFAL usefulness.

The rest of this paper is structured as follows. Section 2 describes the ProFAL system in detail. In Section 3 we present the implementation of ProFAL, describing the sources of biological information used. Section 3.5 presents the achieved results. Section 4 discusses related work. Finally, in Section 5, we express our main conclusions and directions for future work.

2 ProFAL System

ProFAL is an automatic annotation system that identifies and integrates functional information of gene-products into biological databases.

ProFAL considers an annotation as a pair composed by a gene-product and a functional property. An annotation is valid only when its components have a biological relationship between them. Gene-product and functional property are generic concepts, which can be instantiated for various specific cases.

Before start, the biological database has to classify the gene products in families according to their structural information, and the biological properties have to be organized in an ontology structured as a graph.

Figure 1 presents a data flow diagram of ProFAL [4]. The *Biological DB* represents where gene-products are stored and classified in families according to their structure. The *Functional Ontology* represents a set of functional properties classified in a hierarchical taxonomy with specific relations between them. The *Literature* represents a collection of documents, from which can be extracted annotations between the database's gene-products and the ontology's properties. From this collection, the *Retrieval* process identifies a set of documents related to each gene-product stored in the database. Then, the *Extraction* process uses the text of the retrieved documents to annotate each gene-product with functional properties. Next, the *Validation* process automatically classifies the annotations based on heuristics. The annotations and their quantitative classification are integrated in the biological database. Finally, the *Verification* process discards misannotations by analyzing the annotations and their context in the biological database.

The first step of ProFAL is to retrieve only the relevant documents. Instead of extracting information from the entire available corpus, the Retrieval process only selects documents somehow related to each gene-product. Therefore, this process needs to analyze the gene-products attributes stored in the biological database. The result from the Retrieval process is a collection of (document, gene-products) pairs, where the gene-products represent the list of gene-products associated to the document. When available, meta-data about the referenced documents is retrieved, including the title, authors, date, journal, associated MeSH terms, abstract and full text. The meta-data is not es-

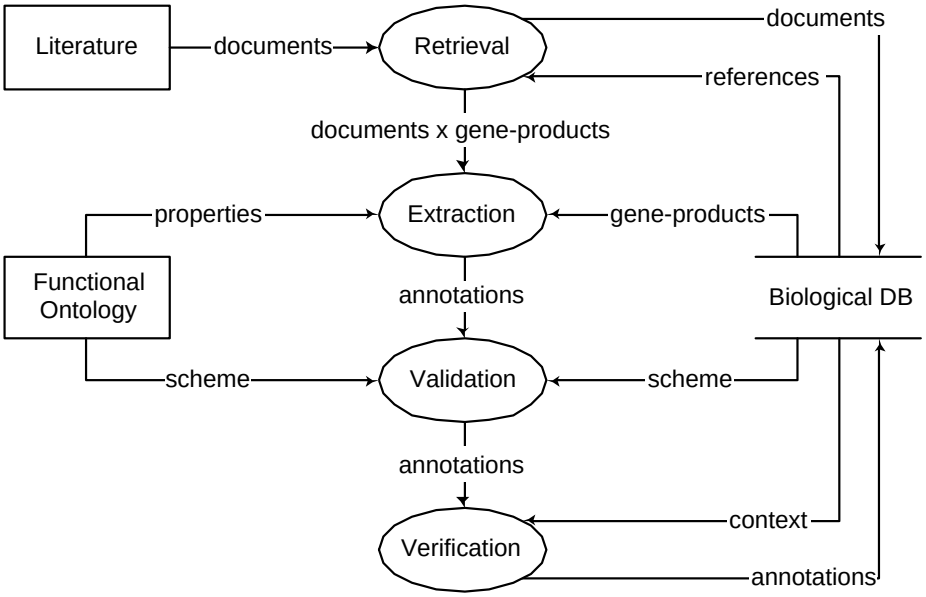

Fig. 1. ProFAL's DataFlow

sential to ProFAL but may be helpful to the user. All the information retrieved is then incorporated in the biological database.

The Extraction process identifies functional properties classified in a functional ontology in the text of the retrieved documents. This produces a collection of (document, functional property) pairs associating each document with the functional properties extracted form its text. The Extraction process can create the annotations by computing the transitivity between the two collections, i.e., if a document is associated with a gene-product and with a functional property, the Extraction process annotates the gene-product with the functional property. This is a naïve method, but can be integrated with known natural language processing techniques to avoid wrong annotations [19].

The Validation process classifies the annotations through a quantitative measure representing the confidence degree on the annotation's correctness. The quantitative measure is implemented using heuristics based on domain knowledge. The heuristics try to discard misannotations that do not satisfy some aspect of the domain.

To check the validation heuristics, the Verification process provides a web interface that shows all the available information about each annotation. A human expert uses the interface to verify the classification of each annotation through the information about its components and sources. ProFAL generates and maintains meta-data describing the sources from where each annotation was extracted, so the manual verification is not too arduous. Normally, the human expert just has to read the original passage from where the annotation was extracted to complete the process.

3 Implementation

We implemented ProFAL as a Java/XML/MySQL application [17]. ProFAL is currently being used in a configuration that uses the following information sources:

- CAZy (Carbohydrate Active enZYmes) is a database of carbohydrate-active enzymes identified and classified in various families by careful sequence and structural comparisons [11]. It describes the families of structurally-related catalytic and carbohydrate binding modules (or functional domains) of enzymes that degrade, modify, or create glycosidic bonds. It also links the sequences to GenBank(GenPept) [6], SwissProt/TrEMBL [3] and PDB [7] entries. These databases are repositories of gene and protein sequence and structural data used to characterize CAZy's enzymes.
- GO (Gene Ontology) provides a structured controlled vocabulary of gene and protein biological roles [10]. The three organizing principles of GO are molecular function, biological process and cellular component. Rison et al. discuss the reasons for choosing GO as the functional scheme in a survey about functional classification schemes [21]. They describe GO as "representative of the 'next generation' of functional schemes". Unlike other schemes, GO is not a tree-like hierarchy, but a directed acyclic graph (DAG), which permits a more complete and realistic annotation.
- PubMed is an online interface for the MEDLINE database [20]. MEDLINE provides a vast collection of abstracts and bibliographic information, which have been published in biomedical journals. In this paper, we consider a document as a bibliographic item whose citation is present in MEDLINE.

In this configuration, CAZy enzymes and GO terms assume the role of gene-products and functional properties, respectively, i.e., our annotations are (CAZy enzyme, GO term) tuples.

CAZy and GO data are available as MySQL databases, and PubMed abstracts as XML documents.

We now detail the implementation of each ProFAL process.

3.1 Retrieval

Most of the external databases (GenBank, SwissProt and PDB) entries contain bibliographic references to documents from where the information was extracted. Since each CAZy enzyme is linked to these entries, we associated the enzyme with the documents cited in its linked entries. These bibliographic references are human curated, so results of interest for each enzyme are certainly reported in its associated documents.

The methods used to extract bibliographic references were different for each external database, because their information is available in different formats. GenBank was until now the only database to publish XML formatted data. For the other databases, we had to develop custom parsers to extract their contents. We considered three types of extraction: (i) through the PubMed identifier, (ii) through the MEDLINE identifier and (iii) through the document's title and authors. When both PubMed and MEDLINE

are available, we check if they represent the same document; if not, the reference is considered incongruent and discarded. However, frequently none of the identifiers is provided and we have to identify the document in PubMed through its title and authors. This was implemented by searching in PubMed for all the authors' publications, and then selecting from the results the most similar title. We were forced to take this fuzzy approach since frequently the title mentioned in a database is not mentioned exactly in the same form as in PubMed.

Initially, we retrieved the complete XML file for each entry through the web. This has shown to be inefficient because of the large size of the GenBank entries (about 64 GB). However, as most entries contain information of no interest to ProFAL, we optimized the process to retrieve only the relevant XML elements.

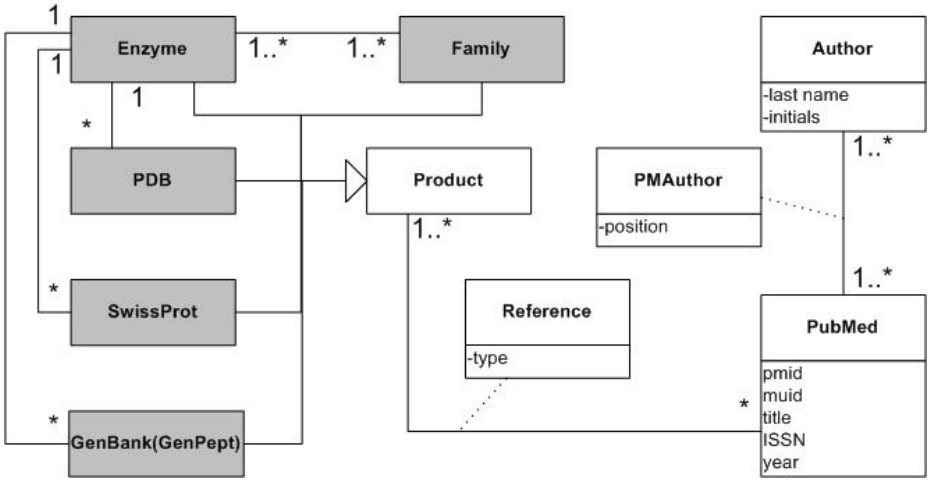

Fig. 2. CAZy's document class scheme

The UML class diagram presented in Figure 2 represents the extensions that ProFAL added to CAZy model. The classes on the left (gray filled) were already defined in CAZy. Each enzyme belongs to one or more families, and are linked to GenBank(GenPept), SwissProt and PDB entries. Each bibliographic reference associates an external database entry to a document in PubMed. The type of each reference defines how it is extracted. The PubMed class represents documents retrieved from PubMed. Its attributes are the PubMed and MEDLINE accession numbers, title, journal (ISSN) and year of publication. The documents' authors are represented in the class Author, with the information about their position in the article. A bibliographic reference can be also manually associated with an enzyme or a family. The retrieved abstracts of each document in XML and plain text are stored in a folder named with their PubMed identifier.

3.2 Extraction

The Extraction process relies on the computation of occurrences of functional terms mentioned in the text, a technique already used in similar projects [18, 14, 22]. We assume that if a document mentions a GO term, then there is an underlying biological relationship between the enzymes related to the document and the GO term, i.e., we *annotate* the enzymes with the GO term.

GO has three organizing principles represented as three orthogonal ontologies. We chose to start by extracting only molecular functional terms, given their greater importance to CAZy.

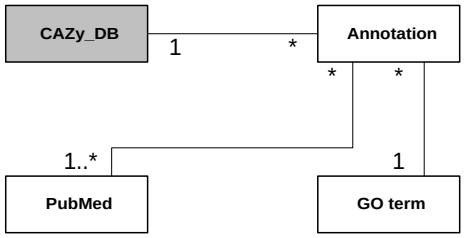

Fig. 3. CAZy's annotation class scheme

The UML class diagram presented in Figure 3 models the concept of annotation, which is an (enzyme, term) pair with the information about the documents from which it was extracted.

3.3 Validation

We have introduced a new heuristic for validating extracted information based on a measure of correlation between structure and function of gene-products [12]. The heuristic validates extracted information by checking if gene products from a common family match a common set of biological properties. Our previous evaluation has shown that it provides a valid approach to automatically classify the correctness of each annotation, enabling a reduction of the workload during the Verification process [13]. In ProFAL, we implemented this heuristic to increase the precision of the extracted annotations.

3.4 Verification

To be accepted, the extracted annotations have to be curated, i.e., it is necessary to verify that their components really have a biological relationship. We published ProFAL's results on CAZy's user interface, which can be used by a human expert to curate the information with a minimal effort.

An example of the bibliographic description of a specific enzyme, named 'k-carrageenase', is presented in Figure 4. The interface has two tables: the Publication table shows the

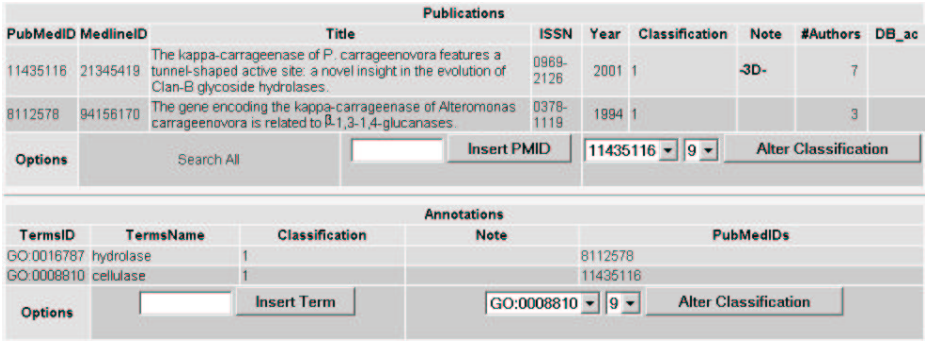

Fig. 4. CAZy's bibliographic Interface

bibliographic references, and the Annotation table shows the terms annotated with this enzyme. It has two bibliographic references. For each reference, the following information is presented: PubMed and MEDLINE accession numbers, title, journal (ISSN), year of publication, comments (note), authors and other referred enzymes. The "**-3D-**" symbol is automatically added into the note field when the article is a PDB reference. This alerts the curator for the fact that this article may contain important structural information. The authors column presents the number of authors through a link to information about them. The DB_ac column presents accession numbers of other CAZy's enzymes that are also referred by the article. The enzyme is annotated with 2 terms. For each reference the following information is presented: term's identifier number, term's type, term's name, comments (note), and the documents from where it was extracted. Both tables have a classification column for curating the entries. Its default value is 1, and its range goes from 0 to 9. An expert can replace the entry's classification according to its relevance. The last row of both tables has buttons to invoke administrative tools, for inserting new entries, and to reclassify the presented entries.

3.5 Results

	GenBank	SwissProt	PDB	Total
Bibliographic references	22849	8998	3561	
Distinct documents	4575	4006	785	6377

Table 1. Number of items retrieved

This section describes the results of our evaluation of the use of ProFAL with the January 2003 release of GO and CAZy databases. Table 1 presents the number of bibliographic references retrieved and the number of documents cited by them. The bibliographic references were linked to 17363 enzymes. ProFAL extracted 13869 annotations from the texts, associating 6918 enzymes with 1342 GO terms. Only about 40% of the

enzymes were annotated due to the lack of bibliographic references for most enzymes. This is not a limitation of ProFAL since it extracted, on average, 2.2 annotations per document.

The user interface for manipulating bibliographic annotations was designed and evaluated by CAZy's curators. After a few iterations in the development cycle, it stabilized its current design.

ProFAL has shown to be effective in finding new biologic annotations and providing a usable interface for their curation. A CAZy curator manually verified 173 extracted annotations related to 5 distinct families. This domain expert classified their relevance to the characterization of the functionality of the families as follows: 32 were classified as very important, 27 were classified as important, and 36 were classified as not so important. The remaining 78 annotations were classified as having no relevance. This gives a total of 95 correct annotations and 78 misannotations, representing a precision of 55%. However, some of the 78 misannotations could still be correct, since some enzymes could also belong to other families that have not been considered. Although the low precision, the feedback given after using ProFAL was that it reduces the workload of analyzing CAZy's data. As our heuristic has been used until now only to evaluate the extracted information, the precision will likely improve when the heuristic is also employed to check for misannotations. This will further shorten the time presently wasted on manual verification of the extracted annotations, and consequently improve the usefulness of ProFAL.

4 Related Work

Several projects have addressed the task of automatic annotation of biological databases. We describe in this section the projects with a closer similarity to our approach.

AbXtract is a system designed for automatic annotation of protein function [1]. Given a collection of protein families, relevant words to each family are extracted from abstracts based on their frequency. The word relevance in a protein family is measured by the difference between its frequencies in the family related set of abstracts and in the common background set of abstracts.

PubGene is a annotated gene network [18]. It is mainly composed by: a gene-article index, based on the genes names occurrences in text; a gene-gene network, based on genes common articles; and a gene-term map, based on the co-occurrence of gene names with GO terms in text, and based on the articles' MeSH terms. The network contains 13712 human genes extracted from the titles and abstracts of over 10 million documents.

GENIES [15] extracts and structures information about cellular pathways from biological literature [15]. The system is composed by several modules: a term tagger that identifies and tags genes and protein names; a preprocessor that identifies sentences, words and phrases, and performs lexical lookup; a parser that identifies relationships through grammar rules; and an error recovery process that uses multiple strategies to parse segments of a sentence.

Blaschke et al. used information extraction methods to identify protein interactions [9]. The goal was to identify interactions described in the Dictionary of Inter-

acting Proteins (DIP) database, in the literature. Their extraction system is based on pattern matching using predefined sentence structure rules and sub-rules (e.g. protein [verb]* [verb] [word]* protein). The list of nouns and verbs was manually generated.

Our system differs from the described projects in the type of information it deals, in its approach of integrating information from heterogenous sources, and in its techniques to evaluate the extracted annotations.

5 Conclusions and Future Work

We presented ProFAL, a novel system for annotation of gene-products stored in a biological database with functional properties. The system automatically extracts and validates the annotations from literature, which is retrieved using other online databases. In addition, it provides a user interface for browsing annotations and assist curators in their verification.

The system was implemented to automatically annotate CAZy enzymes with GO terms. The feedback from ProFAL users demonstrated its usefulness for analyzing biological data. Its main problem was the weak precision of the extracted annotations, which we aim to solve by improving our heuristic and upgrading ProFAL with known natural language processing techniques (such as stemming, part of speech tagging, named entity identification) [19].

Given the success of ProFAL in annotating the CAZy database, we are starting the development of a new configuration of the system for annotating a database with Arabidopsis gene expression data [5]. This will be used to evaluate how ProFAL performs on annotating gene-expressed data, since CAZy is restricted to a class of enzymes.

References

1. M. Andrade and A. Valencia. Automatic extraction of keywords from scientific text: Application to the knowledge domain of protein families. *BIOINF: Bioinformatics*, 14, 1998.
2. T. Attwood and D. Parry-Smith. *Introduction to Bioinformatics*. Longman Higher Education, 1999.
3. A. Bairoch and R. Apweiler. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, 28:45–48, 2000.
4. C. Batini, S. Ceri, and S. Navathe. *Conceptual Database Design: An Entity-Relationship Approach*. Benjamin/Cummings, 1992.
5. J. Becker and J. Feij. Study of gene expression patterns in developing pollen by arabidopsis genechips. *FEBS*, 2001.
6. D. Benson, I. Karsch-Mizrachi, D. Lipman, J. Ostell, B. Rapp, and D. Wheeler. GenBank. *Nucleic Acids Research*, 30:17–20, 2002.
7. H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, and P. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28:235–242, 2000.
8. C. Blaschke, R. Hoffmann, J. Oliveros, and A. Valencia. Extracting information automatically from biological literature. *Comparative and Functional Genomics*, 2:310–313, 2001.
9. C. Blaschke and A. Valencia. Can bibliographic pointers for known biological data be found automatically?: protein interactions as a case study. *Comparative and Functional Genomics*, 2:196–206, 2001.

10. T. G. O. Consortium. Creating the gene ontology resource: design and implementation. *Genome Res*, 11:1425–1433, 2001.
11. P. Coutinho and B. Henrissat. Carbohydrate-active enzymes: an integrated database approach. *Recent Advances in Carbohydrate Bioengineering*, pages 3–12, 1999.
12. F. Couto, M. Silva, and P. Coutinho. Curating extracted information through the correlation between structure and function. In *third meeting of the special interest group on Text Data Mining co-located with 11th Internacional Conference on Intelligent Systems for Molecular Biology (ISMB)*, Brisbane, Australia, June 2003.
13. F. Couto, M. Silva, and P. Coutinho. Improving information extraction through biological correlation. In *Data Mining and Text Mining for Bioinformatics Workshop co-located with 7th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, Dubrovnik-Cavtat, Croatia, September 2003.
14. J. D. et al. Mining MEDLINE: Abstracts, sentences, or phrases? In *PSB*, pages 326–337, 2002.
15. C. Friedman, P. Kra, H. Yu, M. Krauthammer, and A. Rzhetsky. GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*, 17(1):S74–S82, 2001.
16. M. Gerstein. Integrative database analysis in structural genomics. *Nature Structural Biology*, Structural genomics supplement:960–963, November 2000.
17. M. Grand. *Java Language Reference*. O'Reilly, 1997.
18. 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.
19. C. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, 1999.
20. MEDLINE. PubMed database at the National Library of Medicine. http://www.ncbi.nih.gov/PubMed.
21. S. Rison, T. Hodgman, and J. Thornton. Comparison of functional annotation schemes for genomes. *Funct. Integr. Genomics*, 1:56–69, 2000.
22. 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.