PF-Words: Biomedical Literature Based Protein Function Search

Wei Wei, Mike Wong, Anagha Kulkarni

Department of Computer Science,
San Francisco State University,
San Francisco, CA 94132
(weiw,mikewong,ak)sfsu.edu

Abstract

Proteins are large biological molecules whose functions support cellular processes necessary for life. The goal of this work is to leverage biomedical literature to identify the set of protein functions (PF) that are relevant to a free-text English query. This objective is beset by many challenges: paucity of relevant data, rapid growth of related but noisy data, multiple competing databases with incompatible nomenclature and classifications, and lack of a single authoritative ground truth, are but a few.

To address the data scarcity and noise challenges we develop a classification based document selection approach that assembles a high quality document set that is used to infer the set of relevant PFs for the query. The actual PF relevance is estimated using two algorithms that we develop by adapting well established information retrieval and topic modeling techniques. To address the lack of single authoritative ground truth we propose an automated method that generates two definitions of the ground truth which reflect the upper and lower bounds on PF semantics. The conducted empirical evaluation confirms the efficacy of the proposed approaches (F1-score of 0.75 or higher), and illustrates the value of the developed supporting framework.

1 Introduction

Proteomics is the study of the structure, function, and interactions of proteins. Proteins support cellular life processes by means of protein functions (PF) such as digestion, immune response recognition, and DNA regulation and replication. Improper protein function is the root cause for many pathologies. Thus, better understanding of existing PFs, and discoveries of new PFs is critically important for advances in medical research.

At a high level, the goal of the presented work is to support open-ended exploration of protein function semantics. Specifically, given a query in plain English (e.g. binding site), our objective is to identify the set of protein functions that are relevant to the query. Although, several public databases for protein function such as PROSITE, Pfam, and InterPro provide free-text search functionality [4, 8], the effectiveness of these systems is subpar. There are multiple reasons why these search tools perform poorly.

One reason is the problem of knowledge acquisition; characterizing and curating protein function is difficult, expensive, and time-consuming work. Of the hundreds of cataloged protein functions, few have more than a handful of definitive publications. This data scarcity accentuates the classic problem of vocabulary gap between the query and the documents. For example, insulin is important to blood glucose regulation, however, the entries in Pfam or PROSITE for insulin do not mention blood glucose regulation. As a result, if the user query is blood glucose or glucose, the search algorithm does not retrieve insulin as one of the results. This degrades the recall of the search system, where recall is the fraction of total relevant PFs for the query that the system could retrieved.

Also the rapid growth of tangentially related documents (noise) which is fueled by advances in sequencing, structure resolution techniques, and biopathway discovery, has an unfortunate side effect of making it difficult to search for relevant PFs using simple word matching search approaches. Specifically, the noise degrades the precision of the search results because unrelated PFs get flagged as being relevant to the user query. A third issue is the lack of an authoritative nomenclature and categorization. Multiple protein function clustering approaches proliferate and compete in the field. Counterintuitively, they often have different names for clusters and slightly different memberships that are otherwise strongly similar (e.g. ASP PROTEASE vs. Asp).
Motivated by these observations we propose a family of approaches (PF-Words) for identifying relevant protein functions for a query that hinges on three key ideas. First, to address the vocabulary gap problem the sources of information need to be expanded beyond the original entries for the proteins. We leverage the scholarly articles that are available through the popular repository, PubMed Central1 for this purpose. Second, to minimize the noise in the expanded sources of information, they need to be assembled discriminatively, selecting documents that add inferential value. Third, to accommodate multiple protein function clusterings, the ground truth is defined as a range: the minimal set of protein functions, and the maximal set of PFs that are relevant to the query.

These ideas form the core of the PF-Words approaches where we adapt well-established retrieval models and topic modeling techniques to identify the set of relevant PFs for a free-text query. For empirical evaluation of PF-Words approaches we introduce a query set that was defined by biomedical professionals. The experimental results demonstrate that the proposed approaches support a consistent performance of 75% or higher in F1 score. When compared to a strong baseline approach, the PF-Words approaches are at least 8% more precise, and provide at least 39% higher recall. The query set, and the other data used in this work, along with the codebase, will be made available under Creative Commons License.

2 Related Work

Recent achievements in automated knowledge discovery from biomedical literature include limited but improving successes in: biopathway discovery, protein ontology development, database curation, and protein function identification [6, 7, 1, 5]. All of these related problems attempt to overcome inconsistencies in protein naming. Protein nomenclature can reflect the name of the transcribing gene, protein function, sequence or structural motifs, among other choices. This lack of consistency has long foiled attempts to map similar and related concepts, hindering communication.

While knowledge discovery has made some progress, automated method of search and retrieval of protein functions using plain English queries is still developing. Existing search interfaces to many of the online protein function databases, such as PROSITE, Pfam and InterPro, are based on Apache Lucene[2]. In these applications, Lucene indexing is limited to text fields in each respective database and limited cross-references. For example, Pfam’s keyword search currently covers text in Pfam entries, sequence and species descriptions, PDB header information, GO terms, and abstract text from InterPro entries. Relevant PubMed literature is not included in the search, so users of these interfaces must rely on entry text, which only summarizes from the literature, and thus search failures due to vocabulary gap are prevalent on these search systems.

3 PF-Words

We propose a family of algorithms, PF-Words, with the objective of addressing the problem of estimating the relevant protein functions for a user query. The common theme across the proposed approaches is that they use PubMed articles that are directly or indirectly associated with the protein functions to obtain a descriptive representation of the PFs. These textually-rich sources of information are leveraged by each approach in a unique manner to infer the set of relevant protein functions for the query. Below we first describe our methodology for compiling the different sources of information that are used to represent a PF during the relevance estimation process.

3.1 Sources of Information

We assemble three sets of PubMed articles that are used to infer the relevant PF set for a query. The smallest set, Core Docs, comprises of the PF description provided on PROSITE’s entry page for the PF, and the content of the PubMed references that are listed on the entry page. These references are often the PF’s inaugural characterization, or a reformulation of the PF written by a consortium of experts. For each PF there are usually only a handful of these references. Although high-quality, the Core Docs set suffers from the data scarcity problem. Often, terms that are strongly related to the protein function (such as blood glucose regulation for insulin) might not be mentioned in these articles.

We address this information paucity problem by extending the Core set with supporting documents which define our largest set, All Docs. The supporting articles for each PF are identified using a two step process as follows. In addition to primary references, the PROSITE entry for a protein function also references the entry pages for two other widely used protein databases: Swiss-Prot2 and PDB3. These entry pages provide a list of references to PubMed articles that are about the protein, and protein structure. We harvest these references to compile the All References set. As previously mentioned, some of these articles (such as

---

1www.ncbi.nlm.nih.gov/pmc/
2www.uniprot.org
3www.rcsb.org/pdb/home/home.do
those focused on a protein’s role in a biopathway) are only tangentially relevant to the protein function of interest, and might add noise to the set.

To reduce noise, we create a middle set Expanded Docs by selectively adding only those articles from All Docs set that have been estimated to be relevant to the Core PF articles using the following methodology. Given \( N \) PFs, we learn \( N(N - 1) \) pairwise binary naïve Bayes classifiers, each distinguishing between a unique PF pair. The Core articles of each PF are used as the training set to learn the classification models. At test time, we assign a PF to an article if at least \( \left\lfloor \frac{N - 1}{2} \right\rfloor \) of its \( N - 1 \) pairwise classifiers vote in favor of that PF for the given article. Each of these three sets is used by the algorithms that are introduced next to select relevant PFs for a given query.

Each of these three sets, Core Docs, Expanded Docs, and All Docs, is used by the PF-Words algorithms, which are introduced next, to select the relevant PFs for the given query.

### 3.2 PF-Words::Indri

The first approach starts by evaluating the query against one of the PubMed article sets described above. Specifically, the Indri Search Engine\(^4\) is used to rank the articles based on their estimated relevance to the query. Since each article in our dataset is associated with a unique PF, the articles retrieved for a query can be used to deduce the set of relevant PFs for the query. Each retrieved articles \((d)\) assigns a vote to the PF it represents. This vote is computed using the following function.

\[
vote(d) = rel(d) \cdot e^{-\lambda \cdot rank(d)} \tag{1}
\]

where \( rank(d) \) is the article’s rank, \( rel(d) \) is the relevance score assigned by Indri to the article, and \( \lambda \) is the decay factor. This function models our intuition that articles retrieved at early ranks are more reliable representatives of the associated PFs than the articles at lower ranks. As the rank value increases the second component of the equation exponentially dampens the vote assigned by the article. The vote of \( 10^{-4} \) or lower is considered to have converged to zero. The distribution of document relevance scores \((rel(d))\) determines the rate at which this convergence occurs. If relevance scores drop rapidly as we go down the document ranking then the convergence will happen sooner, but if the relevance scores reduce gradually then the convergence will be slower. The convergence, in effect, determines how many top ranking articles are used to determine the set of relevant PFs for the query.

Specifically, all the protein functions represented until the convergence are considered relevant to the query. We refer to this method of rank cutoff selection as the Dynamic approach, since the cutoff may be different for every query. In addition to this method we also employ a simple approach that assumes that the PFs represented by a fixed number of top ranked articles in the Indri results are relevant to the query. We refer to this second technique as the Fixed approach.

### 3.3 PF-Words::LDA

For our second method, we use topic modeling based document ranking. Specifically, we employ an approach (LBDM) [9] that is based on a popular topic modeling technique, latent Dirichlet allocation (LDA) [3]. The central idea behind LBDM is to use topic models in addition to the traditional document models to improve document ranking effectiveness. The topic models are learned from the document collection using LDA as part of a pre-processing step which is performed before any query is evaluated. During query time, the probability of a document \( D \) generating the query term \( q \) is then estimated using the below formulation.

\[
P(q|D) = (1 - \lambda)P_{LDA}(q|D) + \\
\lambda \left( \frac{N_d}{N_d + \mu} P_{MLE}(q|D) + \left( 1 - \frac{N_d}{N_d + \mu} \right) P_{MLE}(q|coll) \right) 
\]

(2)

where \( P_{LDA}(q|D) \) is the term generation probability under the learned topic models, \( P_{MLE}(q|D) \) is the maximum likelihood estimate (MLE) of term probability under document language model, and \( P_{MLE}(q|coll) \) is the MLE of term probability under the collection language model (LM). The second component in the above equation uses document modeling with Dirichlet smoothing using complete collection, as proposed by Zhai and Lafferty [10].

The probability of a document \( D \) generating the complete query is estimated simply using the classic query likelihood model where \( P(q|D) = \prod P(q_i|D) \), and the documents are ranked based on the computed \( P(q|D) \) probabilities. In our approach the next and final step is to estimate the set of relevant protein functions for the query using the document ranking obtained with the LDA-based model. We experiment with the same two cutoff selection approaches described in Section 3.2, Dynamic and Fixed, which determine how many of the top ranked documents are used to select the relevant PF set for the query.

\(^4\)www.lemurproject.org/indri/
4 Experimental Methodology

The data that was used for the empirical evaluation of the proposed approaches, and the baseline search approach are described next.

4.1 Data

Twenty PROSITE protein functions were chosen for this study, based on the number of member proteins and available structures. The list of these PFs along with the sizes of the three information sources is provided in Table 1.

4.2 Queryset and Ground Truth

The following set of 25 queries, that was assembled by biomedical professionals, was used for empirical evaluation: apoptosis, atm, binding site, calcium, cofactors, disulfide bond, drug, growth factor, heme, hiv, hydrolases, iron, kinase, nucleotide binding, oxidoreductases, protease, receptor, serine, signaling, threonine kinase, toxins, transmembrane, tyrosine kinase, vitamins, and zinc.

Defining the corresponding ground truth, the set of relevant PFs for each query, is a challenging task due to lack of consensus among PF researchers. The descriptions provided in the PROSITE PF definition (entry page), along with supporting cross-referenced literature may be considered as a sound foundation (Core Docs set), and used to define the ground truth. However, these documents are few and brief, providing poor coverage of the broader PF semantics. Richer protein function semantics can be uncovered by extending a protein function corpus to include the cross-referenced literature (All Docs set). However, these documents might not all be relevant to protein function (e.g. protein structure resolution methodology documents or biopathway discovery documents, where a particular protein function is among many other protein functions working in concert). As a solution we define two ground truths: Conservative and Expressive. The former considers a PF relevant to the query if the query terms occur in the Core Docs set of that PF. The expressive ground truth identifies a PF as relevant to the query if the query terms occur in the All Docs set of that PF. On an average, the conservative and expressive ground truths identify 2.67 and 3.71 relevant protein functions per query, respectively.

4.3 Baseline PF search approaches

We model the baseline approach after the typical search system used by protein function repositories (PROSITE, Pfam). The common approach is to search for entry pages with exact matches of the query terms. Since the Core Docs set contains the entry pages we use this set as the information source for the baseline approach. However, note that Core set also contains the primary references for each PF, as a result, our baseline has access to more information than the typical existing
The baseline approach uses a fixed cutoff of 10, that is the top 10 documents are used to identify the set of relevant PFs for the query.

5 Results and Analysis

The three sources of information (Core, Expanded, and All Docs), and the two protein function selection algorithms (Indri based and Topic modeling based), together provide six different configurations that we experiment with. A 5-fold cross validation methodology was adopted to set the following algorithm parameters: decay rate $\lambda$ for the exponential decay approach of document cutoff selection, and smoothing parameter $\mu$ and mixing parameter $\lambda$ for LDA-based PF selection approach. The experimental results for the baseline, and the other configurations are reported in Table 2. The performance is quantified using standard metrics for classification tasks: precision, recall, and F1-measure. The difference in the performances of any two configurations was tested for statistical significance using paired T-test ($p < 0.05$).

When using the conservative ground truth definition to characterize Indri or LDA-based retrieval performance, Expanded and All Docs sets report substantially lower performance than Core Docs. That is, adding supporting articles to the primary references hurts the inference process. The short explanation for this trend is: incomplete ground truth. Because the conservative ground truth is derived from the primary sources their coverage of PF semantics is limited. The supporting articles often identify relevant PFs that get labeled as false-positives because they are not present in the primary references. As an example, for query *nucleotide binding* the Indri-based approach identifies five PFs as relevant to the query (PROTEIN KINASE ST, PROTEIN KINASE TYR, RNASE PANCREATIC, EF HAND 1, and ATPASE ALPHA BETA), but the conservative ground truth only labels one of them (ATPASE ALPHA BETA) as being relevant. PROTEIN KINASE ST, and PROTEIN KINASE TYR functions transfer phosphate group from a nucleoside triphosphate to another protein, and are therefore relevant. RNASE PANCREATIC functional class are responsible for degradation of RNA to small molecules, and RNA is composed of nucleotides. Also, protein candidates of EF HAND 1 functional class can bind to a nucleoside triphosphate. Thus all four PFs identified as irrelevant by the conservative ground truth are actually valid selections for this query. For these reasons, henceforth, we focus on analyzing the results with the expressive ground truth definition.

For both, Indri and LDA-based approaches, expanding the Core Docs with supporting documents (Expanded and All) improves F1 scores substantially. Using the Dynamic rank cutoff selection approach performs consistently better than using a fixed cutoff.
of top 10 ranks. The recall supported by the Dynamic
cutoff selection is almost always higher than the recall
with top 10 ranks, which indicates that the former
allows deeper ranks than 10 to contribute. Overall,
the PF-Words::Indri approach with Expanded Docs set,
and Dynamic rank cutoff selection provides the best
performance, F1-score: 0.81.

6 Conclusions and Future work

This work studied an important problem that is
at the intersection of proteomics and text analytics.
Given a free-text query, how to identify relevant pro-
tein functions for the query, using biomedical liter-
ature. We proposed two solutions for this problem
that successfully adapt existing IR and topic modeling
techniques, and support strong empirical performance
(F1-score of 0.75 or higher). We illustrate that one
of the crucial components of the proposed approaches
is the information source (document sets) that is used
to identify the relevant PFs. A document selection
approach that addresses the data scarcity problem
without introducing noise to the set supports the best
performance. We also studied two different definitions
of ground truth to capture the lower and upper bounds
of PF semantics, and demonstrate their value.

There are four immediately interesting future direc-
tions for this research: (1) developing a user-interface
for interactive, exploratory search for PF researchers;
(2) integration of the developed approaches with a
protein function search algorithm such as Stanford
FEATURE to search for relevant functions in an unan-
notated protein structure; (3) fuzzy searches using
query expansion to find related relevant protein func-
tions; and (4) a more thorough investigation with all
available protein functions in PROSITE.

7 Acknowledgments

This work was supported in part by Center for
Computing for Life Sciences at San Francisco State
University. The authors would like to thank Dr.
Dragutin Petkovic for his inputs and support.

References

[1] Hagit Shatkay Andrew Wong. Protein Function
Prediction using Text-based Features extracted
from the Biomedical Literature: The CAFA
Challenge. BMC Bioinformatics, 14(Suppl 3),
2013.

[2] Andrzej Białecki, Robert Muir, and Grant Inger-
soll. Apache lucene 4. In SIGIR 2012 workshop
on open source information retrieval, pages 17–24,
2012.

[3] David M Blei, Andrew Y Ng, and Michael I
Jordan. Latent Dirichlet Allocation. 3:993–1022,
2003.

[4] Robert D. Finn, Alex Bateman, Jody Clements,
Penelope Cogoll, Ruth Y. Eberhardt, Sean R.
Eddy, Andreas Heger, Kirstie Hetherington, Liisa
Holm, Jaina Mistry, Erik L L Sonnhammer,
John Tate, and Marco Punta. Pfam: The
protein families database. Nucleic Acids Research,

as data: Using text-based features for proteins
representation and for computational prediction of
their characteristics. Text mining of biomedical
literature, 74:54–64, 2015.

Accomplishments and Challenges in Literature
Data Mining for Biology 1 Introduction 2 Recent
Accomplishments : Increasing Depth. Assessment,

Automatic extraction of gene/protein biological
functions from biomedical text. Bioinformatics,

[8] Franck Valentin, Silvano Squizzato, Mickael
Goujon, Hamish McWilliam, Juri Paern, and
Rodrigo Lopez. Fast and efficient searching of
biological data resources-using EB-eye. Briefings

[9] Xing Wei and W Bruce Croft. LDA-Based
Document Models for Ad-hoc Retrieval. pages

[10] Chengxiang Zhai and John Lafferty. A study of
smoothing methods for language models applied
to information retrieval. ACM Transactions of