Methods xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Text as data: Using text-based features for proteins representation and for computational prediction of their characteristics

Hagit Shatkay [a,b,e,*], Scott Brady [c,e], Andrew Wong [d,e]

[a] Dept. of Computer and Information Sciences, University of Delaware, Newark, DE 19716, USA
[b] Delaware Biotechnology Institute, University of Delaware, Newark, DE 19711, USA
[c] School of Medicine, University of Toronto, Toronto, ON M5S 1A8, Canada
[d] Office of Personalized Genomics & Innovative Medicine, Mount Sinai Hospital, Toronto, ON M5G 1X5, Canada
[e] Computational Biology and Machine Learning Lab, School of Computing, Queen's University, Kingston, ON K7L 3N6, Canada

ARTICLE INFO

Article history:
Received 12 April 2014
Received in revised form 21 September 2014
Accepted 21 October 2014
Available online xxxx

Keywords:
Biomedical text mining
Machine learning
Text classification
Protein subcellular location
Protein function prediction
Protein annotation
Text mining
Protein representation
Protein location prediction

ABSTRACT

The current era of large-scale biology is characterized by a fast-paced growth in the number of sequenced genomes and, consequently, by a multitude of identified proteins whose function has yet to be determined. Simultaneously, any known or postulated information concerning genes and proteins is part of the ever-growing published scientific literature, which is expanding at a rate of over a million new publications per year. Computational tools that attempt to automatically predict and annotate protein characteristics, such as function and localization patterns, are being developed along with systems that aim to support the process via text mining. Most work on protein characterization focuses on features derived directly from protein sequence data. Protein-related work that does aim to utilize the literature typically concentrates on extracting specific facts (e.g., protein interactions) from text. In the past few years we have taken a different route, treating the literature as a source of text-based features, which can be employed just as sequence-based protein-features were used in earlier work, for predicting protein subcellular location and possibly also function. We discuss here in detail the overall approach, along with results from work we have done in this area demonstrating the value of this method and its potential use.

1. Introduction

The era of large-scale genome-based biology has been marked by an unprecedented number of sequenced genes and proteins, accompanied by a tremendous growth in the number of biomedical publications. High-throughput sequencing technology provides fast and relatively easy means to obtain the sequence information for a multitude of proteins. Naturally, traditional experimental methods for studying these proteins lag behind, resulting in a rapid increase in the number of proteins whose sequence is available but whose role within biological processes remains unknown. Much research is thus dedicated to characterizing proteins, identifying their structure, function, location and interactions, as well as to making such information available through public databases such as SwissProt and UniprotKB [59] or the Protein Data Bank [42].

As a lot of the information pertaining to genes and proteins is (and has been) published throughout the scientific literature, there is a surge of interest in biomedical text mining methods [52], aiming to accelerate the acquisition and the structuring of information obtained from unstructured text. Simultaneously, computational methods for predicting and deducing protein function, structure and location are also being developed. Here we discuss work that is in the intersection of these two directions, namely, the utilization of text as a component within computational methods for predicting protein subcellular location and function.

Computational methods for predicting proteins' characteristics typically utilize features derived from protein sequence, possibly along with structure or interaction networks [5,20,46]. For instance, the function-prediction systems GOtcha [30], OntoBLAST [63], and BLAST2GO [12] rely on sequence similarity, PHUNCTIONER [39] and ConFunc [60] use similarity between protein structures, while GeneMANIA [33] and an earlier system by Chua et al. [10] rely on protein-interaction networks. Similarly, quite a few location prediction systems use sequence motifs, sequence similarity, or more refined sequence-based features to predict the

* Corresponding author at: Dept. of Computer and Information Sciences, University of Delaware, Newark, DE 19716, USA.
E-mail address: shatkay@udel.edu (H. Shatkay).

http://dx.doi.org/10.1016/j.ymeth.2014.10.027
1046-2023/

subcellular location of proteins [e.g., [3,9,18,23,24,34,55]]. Notably, computational prediction of a protein's function or location, as discussed here, is often framed as a classification task. The class-labels are the possible functions or the organelles within the cell, and the goal is to take a protein – typically represented as a feature vector based on sequence properties – and assign it with a correct class-label.

An alternative approach to the sequence-based representation of proteins is text-based representation. The underlying idea is that if a passage of text is relevant to a protein, there is often information therein that can be used to help deduce a protein's class (i.e., its subcellular location or its bio-molecular role). In the context of protein characterization, text can be put to use through two distinct approaches: *Information Extraction* and *Text-based classification*. While we focus on the latter (i.e., classification), we briefly discuss the former. *Information extraction* systems aim to identify and extract phrases or terms within the text that explicitly describe the protein's characteristics. That is, rather than *predict* yet unknown information, extraction systems aim to *find out what has already been discovered and reported* in the literature about the protein in terms of function, process or location. AbXtract [1], which was one of the earliest extraction systems in the biomedical domain, aimed to identify and rank sentences discussing protein function based on statistical properties of words in the sentence. Craven and Kumlien [14] have used a hidden Markov model of sentence structure to *extract* protein subcellular location from documents discussing it. Several later systems have used extraction strategies to identify text passages discussing protein function. For instance, Pérez et al. [40] introduced a dictionary-based system that extracts keywords from the literature or from databases and associates them with GO categories; Other systems used pattern matching and sentence structure to retrieve sentences containing a protein along with Gene Ontology (GO) terms denoting function [8,26]. A recent function prediction system [56] identifies pairs of GO terms and proteins within abstracts, and uses them as part of an integrative similarity measure (kernel) employed in classifying proteins by function. Additional information extraction systems have been used in a variety of knowledge discovery tasks within the biomedical domain (see surveys, e.g., [11,25,52]).

In contrast to textual information extraction systems, *classification systems* represent genes and proteins using features that are derived from text sources – regardless of whether the text explicitly discusses the proteins' function/location. The idea underlying this approach, which is rooted in probabilistic information retrieval and language models [47–49], is that the language or, more explicitly, *the distribution of words* used within the text to discuss the protein (or the gene) can provide cues about its function, process or location. We can thus make use of sets of proteins whose characterization is already known, represent them based on text-features, and train machine-learning classifiers that can then assign class-labels to yet-unannotated proteins (where the latter are also represented using text-based features).

For instance, in an early work Raychaudhuri et al. [45] classified published abstracts into *biological process* GO categories (using 21 categories). Proteins that are mentioned in each abstract are then assigned the GO categories associated with the abstract. In our own early work on using text for characterizing gene's function, we have introduced the use of probabilistic topic models applied to PubMed abstracts for representing sets of genes sharing a common function [53]. Van Driel et al. [16] later use a similar idea for grouping and characterizing genes, by identifying similarities among the text describing their respective phenotypes, obtained from OMIM; Groth et al. [21,22] also approach phenotype-based study of genes by applying a clustering technique to the text-descriptions of phenotypes, and associating text and keywords within it with GO categories. A text-based classification system by Stapley et al. [57] used support vector machines to assign yeast proteins to subcellular locations; Nenadic et al. [36] used a similar approach to annotate proteins with one of *11 biological process* terms from the upper levels of the GO hierarchy. In both cases, proteins were represented as vectors of words occurring in abstracts that mentioned the protein's name. More recent work in the area of text-assisted functional annotation [37,58] examined the classification of biomedical abstracts (rather than of proteins) into functional categories, tagging the abstracts themselves with relevant GO codes.

Another source of text considered for use in automated characterization of proteins consists of the descriptive terminology (typically GO terms) appearing within protein annotations in public databases, such as SwissProt/UniProtKB. Eisenhaber and Bork's rule-based Meta_A(nnotator) [17] used functional annotation terms from the protein's SwissProt entry to deduce the protein's location. Nair and Rost [35] used text from the same source to associate proteins with selected functional keywords and develop the LOCkey classifier for predicting subcellular location. Utilizing only such functional keywords for protein representation greatly limited the coverage of the system to proteins already annotated with these keywords. Eskin and Agichtein [19] expanded on LOCkey by utilizing as part of the classification scheme more of the annotation terms associated with the proteins, as well as protein sequence features, albeit without demonstrating improved performance. More recent systems for protein subcellular location prediction such as Proteome Analyst [28] and YLoc [3], while relying primarily on sequence-based features for representing proteins, also employ text-features obtained from protein annotation (e.g., GO terms annotating the proteins) to aid in the prediction. Notably, having a prediction system use such features for protein-representation implies that the protein in question has already been manually curated and annotated, which limits the utility of the system to aid in de novo annotation of proteins that have not yet been characterized.

The methods we discuss throughout the rest of this paper aim to take advantage of the available published text for protein representation and classification, without relying on manually-curated annotation terms (such as GO terms assigned to the protein). We thus focus on text obtained from the published literature, specifically from PubMed abstracts [43], that can be associated with proteins and utilized by automated systems. These ideas have been put to use in specific systems we have developed to address the two tasks discussed before, namely protein subcellular location prediction [2,4] and protein function prediction [61]. Here we present a complete framework for using text as a basis for representing and characterizing proteins; moreover, the *function prediction* work and results discussed here employ *support vector machine classifiers (SVM)*, as opposed to *k*-nearest neighbor classifiers that were used before (the latter were reported in [61]). The approach and the methods, the results of applying them – and the lessons learned from these applications, are presented and discussed in detail throughout the following sections.

2. Methods: from proteins to text and back

To use text as a form of data for characterizing proteins, one must first identify a source of text pertaining to proteins, along with a strategy for associating each protein with its related text. Next, one needs to represent proteins as feature vectors based on the associated text, possibly making use of additional aspects of the protein (such as sequence-based information) in the representation. Once proteins are represented as feature vectors, machine-learning methods for training and testing classifiers can be applied and used for protein characterization. In this section we focus primarily on the first two steps, namely association of proteins