Journal of Theoretical Biology ■ (■■■■) ■■■–■■■

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Subcellular localization for Gram positive and Gram negative bacterial proteins using linear interpolation smoothing model

Q1 Harsh Saini [a], Gaurav Raicar [a], Abdollah Dehzangi [b], Sunil Lal [a], Alok Sharma [a,b]

Q2 [a] University of the South Pacific, Fiji
[b] Griffith University, Australia

HIGHLIGHTS

- We introduce a novel classifier, linear interpolation, for subcellular localization.
- Inspiration to use this technique came from natural language processing.
- The techniques tries to model dependencies between amino acids.
- We achieved good results on two bacterial datasets.

ARTICLE INFO

Article history:
Received 22 March 2015
Received in revised form
10 July 2015
Accepted 14 August 2015

Keywords:
Natural language processing
Hidden Markov models
Dependency models
Feature extraction

ABSTRACT

Protein subcellular localization is an important topic in proteomics since it is related to a protein's overall function, helps in the understanding of metabolic pathways, and in drug design and discovery. In this paper, a basic approximation technique from natural language processing called the linear interpolation smoothing model is applied for predicting protein subcellular localizations. The proposed approach extracts features from syntactical information in protein sequences to build probabilistic profiles using dependency models, which are used in linear interpolation to determine how likely is a sequence to belong to a particular subcellular location. This technique builds a statistical model based on maximum likelihood. It is able to deal effectively with high dimensionality that hinders other traditional classifiers such as Support Vector Machines or *k*-Nearest Neighbours without sacrificing performance. This approach has been evaluated by predicting subcellular localizations of Gram positive and Gram negative bacterial proteins.

1. Introduction

Subcellular localization of proteins is a very important research topic in molecular cell biology and proteomics since it is closely related to the protein's functions, metabolic pathways, signal transduction and other biological processes within a cell (Briesemeister et al., 2010; Imai and Nakai, 2010). Knowledge of a protein's subcellular localization also plays an important role in drug discovery, drug design and biomedical research. Determination of subcellular localizations experimentally is laborious, time-consuming and, in some cases, experimental means to determine some subcellular localizations of proteins is difficult using fluorescent microscopy imaging techniques (Mei et al., 2011).

In recent years, there has been significant progress in subcellular localization predication using computational means. There are approaches that extract features directly from the syntactical information present in protein sequences such as amino acid composition (AAC) (Tantoso and Li, 2008; Habib et al., 2008), N-terminus sequences (Höglund et al., 2006) and pseudo-amino acid composition (PseAAC) (Chou, 2011). Some approaches use the evolutionary information present in Position Specific Scoring Matrices (PSSM) to extract features (Xiao et al., 2011b). Features can also be generated from protein databases such as the annotations in Gene Ontology (GO), functional domain information, and/or textual information from the keywords in Swiss-Prot (Shen and Chou, 2010a, 2010b; Chou and Shen, 2010a; Chou et al., 2012; Li et al., 2014). Moreover, some researchers have utilized the information present in the physicochemical properties of amino acid residues to enhance prediction accuracy (Du and Li, 2006; Tantoso and Li, 2008). However, most prevalent techniques are a hybrid collection of various features to

E-mail addresses: saini_h@usp.ac.fj (H. Saini), raicar_g@usp.ac.fj (G. Raicar), abdollah.dehzangi@griffithuni.edu.au (A. Dehzangi), lal_s@usp.ac.fj (S. Lal), alok.fj@gmail.com (A. Sharma).

http://dx.doi.org/10.1016/j.jtbi.2015.08.020
0022-5193/

help identify discriminatory information for the classifiers to obtain an improved prediction accuracy (Tantoso and Li, 2008; Shen and Chou, 2010a, 2010b; Chou and Shen, 2010a; Chou et al., 2012; Briesemeister et al., 2010; Chou, 2011).

In proteomics, frequencies or probabilities of occurrence for amino acid subsequences in proteins have been used to extensively model proteins. Some features that can be considered as variants of such models include Amino Acid Composition (AAC) (Ding and Dubchak, 2001), Pairwise Frequency (PF1) and Alternate Pairwise Frequency (PF2) (Ghanty and Pal, 2009), bigram (Sharma et al., 2013), *k*-separated bigrams (Saini et al., 2014, 2015), and trigram (Paliwal et al., 2014). Although such models have been rigorously studied by researchers, they have mostly considered the probability distribution as an extracted feature for classification via means of another classifier such as Bayesian Classifiers, Artificial Neural Networks and Support Vector Machines (Saini et al., 2014; Sharma et al., 2013; Paliwal et al., 2014; Ghanty and Pal, 2009; Ding and Dubchak, 2001; Höglund et al., 2006).

Such probability models are also prevalent in other fields of study such as natural language processing (NLP), however, they have been deployed in a completely different manner. Instead of considering these models as features for input into other classifiers such as Support Vector Machines (SVM) or *k*-Nearest Neighbours (kNN), they are considered probabilistic dependency models that determine the likelihood of a protein belonging to a subcellular location. In this research, the linear interpolation smoothing model is proposed which extracts features using syntactical information from the protein sequences for predicting protein subcellular localizations. This approach is a basic approximation technique in NLP and its concepts have been applied in proteomics for this study. Linear interpolation builds probabilistic profiles for proteins based on the frequency information of amino acid subsequences extracted from proteins to perform subcellular localization. These probabilistic profiles may be following the independent or dependent model based on the probabilities being extracted. In this paper, the application of linear interpolation in proteomics is investigated and its ability to predict subcellular localizations of Gram positive and Gram negative bacterial proteins is analyzed.

2. Materials

For the purposes of comparison and benchmarking, publically available Gram positive and Gram negative bacterial protein databases were used. These databases have been widely used by researchers in recent literature (Dehzangi et al., 2014; Shen and Chou, 2010b; Pacharawongsakda and Theeramunkong, 2013; Huang and Yuan, 2013).

2.1. Gram positive dataset

This dataset comprises Gram positive bacterial proteins that contains both singleplex and multiplex proteins, which cover four subcellular locations. It contains 519 unique proteins where 515 proteins belong only to one location and 4 proteins belong to two locations. Similarly, it also has a pairwise sequence similarity threshold of 25% (Shen and Chou, 2010b). The details of the Gram positive dataset are provided in Table 1.

2.2. Gram negative dataset

In this dataset, Gram negative bacterial proteins covering eight subcellular locations are collected. It contains 1392 unique proteins where 1328 proteins belong only to one location and 64 proteins belong to two locations. Similarly, it also has a pairwise sequence similarity cut-off of 25% (Xiao et al., 2011b). The details of the Gram negative dataset are provided in Table 2.

Table 1
Summary of Gram positive bacterial protein dataset

Subcellular location	Number of samples
Cell membrane	174
Cell wall	18
Cytoplasm	208
Extracellular	123

Table 2
Summary of Gram negative bacterial protein dataset

Subcellular location	Number of samples
Cell inner membrane	557
Cell outer membrane	124
Cytoplasm	410
Extracellular	133
Fimbrium	32
Flagellum	12
Nucleoid	8
Periplasm	180

3. Method

Linear interpolation is a *backoff* model (Schölkopf et al., 2004), indicating that it aggregates information from different submodels to determine the likelihood of a protein belonging to a particular class. It builds probabilistic profiles for proteins based on the frequency information of amino acid subsequences extracted from proteins to perform subcellular localization. In this sense, linear interpolation is related to Hidden Markov Models (HMMs) and uses the Markov assumptions to build probabilistic profiles of varying dependencies for proteins, which are later used in this technique to determine the probability of a protein for belonging to a particular subcellular location (Caragea et al., 2010; Murphy and Bar-Joseph, 2011).

These probabilistic profiles are similar to amino acid subsequence models that are prevalent in the literature, however, their application in linear interpolation is completely different than those previously published. Additionally, there is an absence of techniques, in the literature, that aggregate information from various probabilistic models to form a consolidated prediction model. In this scheme, linear interpolation, an approach novel to proteomics, is used to consolidate information from dependent and independent probability distributions to identify the maximum likelihood of a query protein for belonging to a subcellular location.

3.1. Algorithm

Computationally, protein sequences and natural languages share many similarities. They both are ambiguous (similar structures can have different meanings), can be very large, are constantly changing, and are constructed by a combination of underlying set of constructs, amino acids for protein sequences and words for natural languages. Thus, there is a need to explore the applicability of some basic techniques that are prevalent in NLP for the field of proteomics.

Linear interpolation builds upon probabilistic models of varying dependencies from amino acid subsequences whereby it consolidates the information from these models in an approach

Please cite this article as: Saini, H., et al., Subcellular localization for Gram positive and Gram negative bacterial proteins using linear interpolation smoothing model. J. Theor. Biol. (2015), http://dx.doi.org/10.1016/j.jtbi.2015.08.020