Journal of Theoretical Biology 364 (2015) 121–130

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Feature extraction by statistical contact potentials and wavelet transform for predicting subcellular localizations in gram negative bacterial proteins☆

G.A. Arango-Argoty[a,c,*], J.A. Jaramillo-Garzón[a,b], G. Castellanos-Domínguez[a]

[a] *Signal Processing and Recognition Group, Universidad Nacional de Colombia, s. Manizales, Campus La Nubia, km 7 via al Magdalena, Manizales, Colombia*
[b] *Research Center of the Instituto Tecnologico Metropolitano, Calle 73 No 76A-354, Medellín, Colombia*
[c] *Department of Computational and Systems Biology, University of Pittsburgh School of Medicine, 3501 Fifth Ave, Pittsburgh, PA 15260, USA*

HIGHLIGHTS

- We propose a model to predict protein subcellular localizations.
- We use statistical contact potentials with the continuous wavelet transform.
- Hidden Markov Models are used to model protein features.
- Classification is carried out using support vector machines.
- The proposed method increases the prediction performance over several methods.

ARTICLE INFO

Article history:
Received 8 August 2013
Received in revised form
27 August 2014
Accepted 28 August 2014
Available online 16 September 2014

Keywords:
Statistical contact potentials
Continuous wavelet transform
Hidden Markov models
Feature extraction
Support vector machines

ABSTRACT

Predicting the localization of a protein has become a useful practice for inferring its function. Most of the reported methods to predict subcellular localizations in Gram-negative bacterial proteins make use of standard protein representations that generally do not take into account the distribution of the amino acids and the structural information of the proteins. Here, we propose a protein representation based on the structural information contained in the pairwise statistical contact potentials. The wavelet transform decodes the information contained in the primary structure of the proteins, allowing the identification of patterns along the proteins, which are used to characterize the subcellular localizations. Then, a support vector machine classifier is trained to categorize them. Cellular compartments like periplasm and extracellular medium are difficult to predict, having a high false negative rate. The wavelet-based method achieves an overall high performance while maintaining a low false negative rate, particularly, on "periplasm" and "extracellular medium". Our results suggest the proposed protein characterization is a useful alternative to representing and predicting protein sequences over the classical and cutting edge protein depictions.

1. Introduction

Prediction of protein subcellular localizations can provide clues about the interactions and the different environments where the proteins are likely to reside, helping to elucidate their function and role in biological processes. Besides, identification of cellular compartments helps in the design of protein isolation experiments (Chou and Shen, 2008). For example, the detection of cell-surface-exposed proteins in a bacterial genome contributes in the discovery of therapeutic intervention points or diagnostic markers (Gardy and Brinkman, 2006). In particular, several experimental techniques have been developed for finding the cellular localization of specific proteins such as immunolocalization, direct fluorescent antibody, and tagged isotopes. Those methods are accurate, but their main issues include the high computational burden required to process experiments and intensive labor (Dunkley et al., 2004). However, due to the exponential growth of the protein databases holding a high number of sequences, but without annotations, it is impossible to carry out experimental validations on different organisms. To cope with this drawback, different

☆ *Availability*: The source code of the program has been developed in the Matlab and R environments public available at SourceForge http://sourceforge.net/projects/wamofi/?source=navbar.

* Corresponding author at: Signal Processing and Recognition Group, Universidad Nacional de Colombia, s. Manizales, Campus La Nubia, km 7 via al Magdalena, Manizales, Colombia.

E-mail address: gaarangoa@unal.edu.co (G.A. Arango-Argoty).

http://dx.doi.org/10.1016/j.jtbi.2014.08.051

computational methods have been developed as an alternative to predict subcellular localizations in the particular case of the Gram negative bacterial proteins. Here, the classification of a new bacterial protein is performed by finding similar sequences with experimentally determined cellular compartments.

The first approach to find similar sequences is the use of local alignments like BLAST or PSI-BLAST that search for homologous sequences among several public databases. Although in many cases these methods perform accurate inferences, a high sequence similarity does not guarantee that proteins play the same role in the cell (Yu et al., 2010; Bhasin et al., 2005; Bhasin and Raghava, 2004; Schäffer et al., 2001; Xie et al., 2005). Therefore, homology based annotations have significant drawbacks, namely: propagation of errors, threshold relativity, and low sensitivity/specificity (Sarac et al., 2008). As an alternative, feature-based machine learning approaches model differences between positive and negative samples, employing the extraction of protein properties arranged in a feature space, to be further used as the classifier input (Yu et al., 2006; Wang et al., 2005; Murray et al., 2002; Gardy and Brinkman, 2006; Jaramillo-Garzón et al., 2013). Furthermore, the feature representation can be carried out by inferring local information on the proteins, aiming to extract evolutionally conserved protein subsequences (known as *motifs*). However, this local extraction strategy has the following implications: not all proteins share the same common motifs, not all motifs are highly conserved, and the presence of noise induces errors reducing the classifier performance (Sarac et al., 2008; Blekas et al., 2005). To improve identification performance of subcellular compartments, hybrid approaches are developed that take advantage of the *homology-based strategies and global/local-feature* characterizations. To predict protein subcellular localizations in bacterial organisms, the following methods are commonly used: PSORTb v.3 (Yu et al., 2010), CELLO (Yu et al., 2006), PSLpred (Bhasin et al., 2005), LOCtree (Nair and BurkhardRost, 2005), P-CLASSIFIER (Wang et al., 2005), and GNeg-mPLoc (Chou and Shen, 2008); all these mentioned approaches cover different training algorithms including feature extraction based on amino-acid composition Bayesian networks for prediction inference, signal peptide descriptors, motif matching and scoring, homology representations based on hidden Markov models (HMM), and text labeling. Though these algorithms report low false positive rates, however, most of them produce high false negative scores. This deficiency is related to the protein feature representations. For instance, most of the applications (Psortb, CELLO, LOCtree, P-classifier and GNEg-mPloc) use feature vectors encoding different classes of amino acid compositions like n-grams, physio-chemical properties of the amino acids or the Gene Ontology annotation. In any way, this characterization misplaces the polypeptide chain distribution, leading to lose important information contained in neighboring segments of the amino acid sequence known as protein domains (Scott, 2000; Campbell and Kristina Downing, 1994).

In this work, we propose a method for predicting five distinct subcellular localizations in Gram negative bacteria. The approach makes use of features distributed along the protein sequence that we represent by profile motifs. In order to detect these patterns, we use the continuous wavelet transform that has been proved as a powerful characterization tool of protein motifs (Murray et al., 2002; Arango-Argoty, 2011; Li et al., 2004; Qiu et al., 2003, 2004). Moreover, with the aim to include structural information for representing protein primary structures, we introduce the use of the pairwise protein contact potentials described in the *AAindex* database (Kawashima and Kanehisa, 2000). These contact potentials are employed in different applications like protein structure prediction, fold recognition, molecular dynamics, protein ligand interactions, protein design and prediction of binding affinity (Kawashima and Kanehisa, 2000; Shen and Sali, 2009; Nakashima and Nishikawa, 1994; Miyazawa and Jernigan, 1996, 1985; Boniecki et al., 2003; Hamelryck et al., 2010). Statistical potentials are commonly based on the probability distributions of the measured pairwise distances between proteins and reference states, where contact potentials are expressed as approximations of free energy functions (Hamelryck et al., 2010). Thus, we propose to extract patterns from protein primary sequences using a simple numerical representation from each residue-to-residue contact potentials. However, obtained numerical representations can include high frequency components which produce noisy approximations. Therefore, the continuous wavelet transform is used to better encode those numerical series and stand out the amino acid interactions, allowing clustering of amino acids with similar *free energy* distributions. Afterwards, the protein is split into a set of subsequences. If a given protein group shares a related amino acid distribution, this sequence cluster is described by a similar pattern. In turn, all obtained clusters are modeled by profiles HMM

From the biological point of view, it is worth noting that proteins have key regions known as motifs being important for recognition and binding of different molecules in the cell environment. Also, proteins hold domains that are mostly conserved by evolution and effectuate different functions. Then, provided a protein set containing a common motif/domain properly characterized by the statistical contact potentials, the continuous wavelet transform can correctly identify and localize the motif/domain regardless if it is scattered over different positions among the protein set or even if the motif/domain includes amino acid mutations. Thus, we make use of this outcome to identify conserved motifs into a specific cellular compartment. Particularly, to avoid bias and classification overtraining, we make use of the free available datasets of gram negatives bacterial proteins: (i) A protein modeling dataset taken from ePsortdb that is used to detect the expressed motifs over the subcellular localizations (Nancy et al., 2011); (ii) while a control data set (reported in Gardy and Brinkman, 2006) is used for validation purposes, where the expressed profile-motifs from the modeling sequences are used as features into a single-class-SVM classifier for predicting the subcellular compartments. Lastly, for the sake of comparison, the influence of the wavelet transforms on the considered approach is evaluated in three different ways: (1) performance evaluation of three currently active softwares for subcellular localization prediction in Gram-negative bacteria: Psortb, CELLO, and SOSUIGramN; (2) assessment of classical protein representations; (3) a subsequence/profile based method that is close to the approach proposed here (the difference is the way the *motifs-features* are obtained). For (2) and (3) views, all evaluations are carried out following the same classification strategy. As a result, the performance prediction obtained by the wavelet method shows an improvement on three of the five considered subcellular localizations when it is compared to the other assessments. Thus, the wavelet method may be viewed as a reliable and efficient protein extraction alternative for improving the performed prediction of the of the five major protein subcellular localizations for Gram negative bacterial organisms.

This paper is organized as follows: in Section 2, a detailed explanation of the methodology is provided. In Section 3, the wavelet-based method is tested on Gram negative proteins against the state of the art approaches. Finally, in Section 4 the remarks and directions for further research are given.

2. Materials and methods

The proposed method appraises two principal stages: (A) *Motif descriptor* that models each protein set (i.e., modeling dataset) as a group of profiles. This HMM-based descriptor assumes there are recognizable amino-acid sequences (termed motifs) in different proteins that usually indicate biochemical function, for example, the so-called zinc finger motif (Klug, 2010). Furthermore, *motifs* are