Computational Biology and Chemistry 65 (2016) 1–7

Contents lists available at ScienceDirect

Computational Biology and Chemistry

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

Research Article

Predicting protein subcellular localization based on information content of gene ontology terms

Shu-Bo Zhang[a,*], Qiang-Rong Tang[b]

[a] Department of Computer Science, Guangzhou Maritime Institute, Room 803 Building 88, Dashabei Road, Huangpu District, 510275, Guangzhou, PR China
[b] Department of Shipping, Guangzhou Marine Institute, Room 205 Shipping Building, Hongshan NO.3 Road, Huangpu District, 510275, Guangzhou, PR China

ARTICLE INFO

Article history:
Received 11 April 2016
Received in revised form 10 July 2016
Accepted 11 September 2016
Available online 14 September 2016

Keywords:
Protein subcellular localization
Gene ontology
Information content
Support vector machines

ABSTRACT

Predicting the location where a protein resides within a cell is important in cell biology. Computational approaches to this issue have attracted more and more attentions from the community of biomedicine. Among the protein features used to predict the subcellular localization of proteins, the feature derived from Gene Ontology (GO) has been shown to be superior to others. However, most of the sights in this field are set on the presence or absence of some predefined GO terms. We proposed a method to derive information from the intrinsic structure of the GO graph. The feature vector was constructed with each element in it representing the information content of the GO term annotating to a protein investigated, and the support vector machines was used as classifier to test our extracted features. Evaluation experiments were conducted on three protein datasets and the results show that our method can enhance eukaryotic and human subcellular location prediction accuracy by up to 1.1% better than previous studies that also used GO-based features. Especially in the scenario where the cellular component annotation is absent, our method can achieved satisfied results with an overall accuracy of more than 87%.

1. Introduction

Predicting the location where a protein resides within a cell is a fundamental and crucial issue in cell biology, as it contributes to our knowledge about the molecular function of a protein, as well as the biological pathway in which it is involved (Cocco et al., 2004; Park et al., 2005), which will in turn provide insights into drug targets identification and drug design (Cai and Chou, 2005). Conventionally, experimental methods have been regarded as gold standard for protein subcellular localization prediction. However, wet-lab biological experiments are time-consuming, laborious and costly (Wan et al., 2015). In particular, with the completion of large-scale sequencing project, the number of protein sequences has grown exponentially, which brings enormous challenges for traditional experimental methods to determine the subcellular localization for such huge number of proteins. This makes it extraordinarily desirable to develop efficient automated methods to determine protein localization accurately.

As a matter of fact, computational approaches to the prediction of subcellular localization of proteins have been attracting increasing attentions from the community of bioinformatics, and extensive studies have been conducted in the past two decades (Du and Wang, 2014). As pointed in a recent review (Chou, 2015), in the last decade or so, a number of web-servers were developed for predicting the subcellular localization of proteins with both single site and multiple sites based on their sequences information alone. They can be roughly classified into two series (Chou, 2015). One is the "PLoc" series and the other is "iLoc" series. The "PLoc" series contains the six web-servers (Shen and Chou, 2009a, 2009b, 2010a, 2010b; Chou and Shen, 2010a, 2010b) to deal with eukaryotic, human, plant, Gram positive, Gram negative, and virus proteins, while the "iLoc" series contains the seven web-servers (Xiao et al., 2011a, 2011b; Chou et al., 2012; Wu et al., 2011, 2012; Lin et al., 2013) to deal with eukaryotic, human, plant, animal, Gram positive, Gram negative, and virus proteins, respectively. Studies have shown that most significant enhancement in prediction system is achieved by developing feature extraction method rather than improving the classifiers (Tantoso and Li, 2008; Sharma et al., 2015; Dehzangi et al., 2015), in the past two decade, a wide range of features has been extracted from the protein sequences and other resources to characterize protein for their subcellular localization prediction. The features in the

* Corresponding author.
E-mail addresses: 845996912@qq.com (S.-B. Zhang), tangqiangrong@hotmail.com (Q.-R. Tang).

http://dx.doi.org/10.1016/j.compbiolchem.2016.09.009
1476-9271/

literature can be categorized into two groups: sequence-based features and knowledge-based features. The former characterizes a protein based on its amino acid sequence, which come in three forms: (1) compositions, (2) sorting signals, (3) sequence homology, and (4) Position Specific Scoring Matrix (PSSM). The composition-based methods derive the information embedded in the whole amino acid sequences statistically. The compositions in this group include amino acid compositions (Nakashima and Nishikawa, 1994), dipeptide compositions (amino acid pair compositions) (Huang and Li, 2004), gapped amino acid compositions (Park and Kanehisa, 2003), and pseudo amino acid compositions (PseACCs) (Chou, 2001). It is noticeable that Chou's pseudo amino acid composition is probably the most widely adopted composition feature in this field. The sorting-signal-based features are established on the appearance of the signal peptides such as N-terminal sorting signals, C-terminal sorting signals, or nuclear localization signals (NLS), which resides in special position of a protein sequence. The homology-based features are established on the assumption that homologous sequences are more likely to reside in the same subcellular location (Nair and Rost, 2002a), and this kind of compositions include functional domain composition (Jia et al., 2007) and motif compositions (Tang et al., 2013). PSSM provides a substitution probability of a given amino acid based on its position along with the protein sequence (Chou and Shen, 2007). In (Dehzangi et al., 2015) and (Sharma et al., 2015), the authors proposed a normalization approach to construct a normalized PSSM using the information from original PSSM.

Knowledge-based features are derived from some high-level semantic annotations of proteins across various resources, such as text from the titles and abstracts in the literature, protein to protein interaction databases, and Gene Ontology (GO) annotation database(Ashburner et al., 2000), which is the focus of this study. Nair and Rost derived a kind of lexical feature from the SWISS-PROT keywords based on lexical analysis (Nair and Rost, 2002b). Other features proposed by Brady (Brady and Shatkay, 2008) and Fyshe (Fyshe et al., 2008) fall into this category too. Protein-interaction-based features are established on the fact that interacting proteins tend to localize within the same cellular compartments (Schwikowski et al., 2000). Du et al. pointed out that protein–protein interaction information is useful in predicting protein subcellular locations; and they analyzed the protein localization based on protein–protein interaction network (Du and Wang, 2014). Mintz-Oron et al. (2009) used metabolic networks for enzyme localization prediction using constraint based models.

GO-based features come in three forms: (1) the 0–1 value feature that was firstly used by Cai and Chou (2003), where they constructed a feature vector to characterize a protein with each element in it indicating whether the protein is annotated with a predefined GO term; (2) the term-frequency feature that introduced by Wan et al. (2013), where they replaced each element in the 0–1 feature vector with the frequency that a corresponding GO term occurs in the annotation dataset; and (3) implicit feature that based on GO-based similarity measurement (Huang et al., 2008), this kind of feature was embedded into the classification algorithm in the form of semantic similarity between two proteins, which was derived by a kernel function. The 0–1 value feature was widely used and has been shown to be superior to other features for protein localization prediction(Wan et al., 2013; Shen and Chou, 2006a). However, most of researches only take into account the information that whether a protein is annotated with the predefined GO terms or the frequency that a term annotates proteins in a given dataset, while the information hiding in the intrinsic structure of the GO graph is often ignored, which may limits the discriminate power of a prediction method.

As demonstrated by a series of recent publications (Qiu et al., 2016; Xiao et al., 2016; Chen et al., 2016; Jia et al., 2016; Liu et al., 2016a, 2016b) in compliance with Chou's 5-step rule (Chou, 2011), to establish a really useful sequence-based statistical predictor for a biological system, we should follow the following five guidelines: (a) construct or select a valid benchmark dataset to train and test the predictor; (b) formulate the biological sequence samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted; (c) introduce or develop a powerful algorithm (or engine) to operate the prediction; (d) properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor; (e) establish a user-friendly web-server for the predictor that is accessible to the public. Below, we are to describe how to deal with these steps one-by-one.

Inspired by previous work in protein location prediction, we propose a novel method, called GOICPSL, to predict the subcellular localization of protein in this study. GOICPSL is established on the information content (IC) of GO terms, which is derived from the intrinsic structure of GO graph. Our method differs from other GO-based methods in that: (1) it explores information from the intrinsic structure of the GO graph to characterize the GO terms that annotate proteins investigated; (2) it constructs the feature vector with each element in it denoting how informative a GO term is, not just the presence or absence of some predefined GO terms; and (3) the feature that only consist of information from the molecular function (MF) and the biological process (BP) sub-ontologies can still achieve high prediction accuracy in the scenario where the cellular component (CC) information is absent. Evaluation experiments were conducted on a eukaryotic protein dataset and two benchmark datasets, and the results demonstrate that our method can enhance eukaryotic and human subcellular location prediction accuracy by up to 1.1% better than previous studies that also used GO-based features. Especially in the scenario where the cellular component annotation is absent, our method can achieved satisfied results with an overall accuracy of more than 87%. The algorithm GOICPSL was implemented in Matlab and is freely available from http://ejl.org.cn/bio/GOICPSL.zip

2. Materials and methods

2.1. Materials

Two protein datasets provided by Chou (Shen and Chou, 2006a, 2006b) and a new eukaryotic protein dataset were used to test the performance of our approach. Chou's datasets EUK16 (Shen and Chou, 2006a) and HUM12 (Shen and Chou, 2006b) were created from Swiss-Prot 48.2 in 2005 and Swiss-Prot 49.3 in 2006, respectively. The EUK16 contains 4150 eukaryotic proteins with 16 classes (from 16 kinds of subcellular localizations), and the HUM12 comprises 2041 proteins with 12 classes. Each protein in the two datasets has no more than 25% sequence identity to any other proteins in the same subcellular localization.

The new eukaryotic proteins was constructed in the same way as Chou (Shen and Chou, 2006a), where the protein dataset was cut off at 25% sequence similarity by a culling program (Wang and Dunbrack, 2003). The proteins with a single subcellular localization that falls within the 16 classes of localizations and being annotated with at least one GO term were selected. After limiting the sequence identity among protein sequences in the same location to 25%, 798 eukaryotic proteins distributed in 14 subcellular localizations (See Table 1 for the details) were used. The GO datasets released in July 2015 were used to test our approach, it contains 27985 BP, 3826 CC, and 9954 MF terms. The gene annotation dataset was retrieved from http://www.ebi.ac.uk/QuickGO/