Analytical Biochemistry xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Analytical Biochemistry

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

Predicting peroxidase subcellular location by hybridizing different descriptors of Chou' pseudo amino acid patterns

Yong-Chun Zuo [a,*], Yong Peng [b], Li Liu [b], Wei Chen [c], Lei Yang [d,*], Guo-Liang Fan [b,*]

[a] The Key Laboratory of Mammalian Reproductive Biology and Biotechnology of the Ministry of Education, Inner Mongolia University, Hohhot 010021, China
[b] Laboratory of Theoretical Biophysics, School of Physical Science and Technology, Inner Mongolia University, Hohhot 010021, China
[c] Center of Genomics and Computational Biology, College of Sciences, Hebei United University, Tangshan 063000, China
[d] College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, China

ARTICLE INFO

Article history:
Received 23 February 2014
Received in revised form 22 April 2014
Accepted 25 April 2014
Available online xxxx

Keywords:
Peroxidase proteins
Chou' pseudo amino acid patterns
GO-homology annotation
Prediction performance

ABSTRACT

Peroxidases as universal enzymes are essential for the regulation of reactive oxygen species levels and play major roles in both disease prevention and human pathologies. Automated prediction of functional protein localization is rarely reported and also is important for designing new drugs and drug targets. In this study, we first propose a support vector machine (SVM)-based method to predict peroxidase subcellular localization. Various Chou' pseudo amino acid descriptors and gene ontology (GO)-homology patterns were selected as input features to multiclass SVM. Prediction results showed that the smoothed PSSM encoding pattern performed better than the other approaches. The best overall prediction accuracy was 87.0% in a jackknife test using a PSSM profile of pattern with width = 5. We also demonstrate that the present GO annotation is far from complete or deep enough for annotating proteins with a specific function.

Peroxidases are ubiquitous enzymes that catalyze a number of oxidative reactions by using various peroxides as electron acceptors [1,2]. These peroxidase proteins are central elements of the antioxidant defense system, which are extremely widespread in almost all microorganisms and higher organisms. They are essential for the regulation of reactive oxygen species levels and for the promotion of various substrates' oxidation [3–5]. There has been increased interest in them over the past few years; for example, the mammalian heme peroxidase enzymes play major roles in both disease prevention and human pathology defense [6,7]. Therefore, knowing the localization of peroxidase proteins will be important for disease prevention and human pathologies.

Proteins in various subcellular locations play distinct roles in biological processes, such as triggering programmed cell death. Protein localization may be used as a starting point for function prediction systems. Knowing a protein's localization is an important step toward understanding its function [8,9]. Experimental and computational methods are two very important methods for annotating protein functional information. During the past 2 decades, a substantial amount of bioinformatics work for predicting protein subcellular location has been carried out and rapidly developed; significant progress has been achieved with the establishment of various organism-specific benchmark datasets [10–15]. However, to the best of our knowledge, there are few theoretical methods for localization prediction for proteins of specific function.

Therefore, it is becoming crucial to develop a reliable automatic subcellular localizer for identifying the locations of functional proteins. In this study we first attempted to annotate the subcellular localization of a specific oxidoreductase, peroxidase, by using a computational method based on state-of-the-art features. Several different descriptors of the Chou' pseudo amino acid pattern have been discussed for localization prediction [16–21], including amino acid composition (AAC) [22], dipeptide composition (DC) [23,24], split amino acid composition (SAAC) [25], evolutionary information (PSSM) [10,26–28], and gene ontology (GO) of homologous proteins [29–32]. All of the above features were selected as input parameters to established an automatic subcellular classifier. The best overall prediction accuracy achieved 87.0% in a jackknife test for eight locations by using a PSSM profile with width = 5. The GO-homology annotation with different sequence identities was also discussed; the evaluation results showed the present GO annotation is far from complete or deep enough for accurately annotating the localization of peroxidase proteins.

* Corresponding authors. Fax: +86 471 5227683.
E-mail addresses: yczuo@imu.edu.cn (Y.-C. Zuo), yanglei_hmu@163.com (L. Yang), eeguoliangfan@sina.com (G.-L. Fan).

http://dx.doi.org/10.1016/j.ab.2014.04.032
0003-2697/

Materials and methods

Benchmark datasets

The data of peroxidase proteins used in this research were extracted from the PeroxiBase database [33]. PeroxiBase is a unique specialized database, which is devoted to established comprehensive peroxidase families and superfamilies from both eukaryotes and prokaryotes. More than 10,000 peroxidase-encoding sequences come from 940 organisms, and each sequence is individually annotated in this database. Since the number of multiplex proteins in the existing database is not large enough to construct a statistically meaningful benchmark dataset for studying a case of multiple locations, only the proteins with singleplex locations were used in this experiment, and every protein is characterized by an expert sequence annotation procedure, with manual curation, which is a guarantee of quality necessary for performing subcellular localization analysis. After the redundant sequences were removed using the CD-HIT algorithm [34], 586 nonredundant peroxidase proteins were obtained. According to the annotation information, these defensin sequences can be classified into eight subcellular locations: apoplastic (30), chloroplastic (44), cytosolic (265), mitochondrial (44), peroxisomal (107), secreted (23), stromal (37), and thylakoid (37). After measuring by the CD-HIT program, most of the protein similarity scores in each family were lower than 80%.

Features and modules

Support vector machine (SVM), as a strong machine learning technique, is used to evaluate various alternative features of our work. SVM is a machine learning algorithm based on statistical learning theory, which has been successfully used for classification [35]. The basic idea of SVM is to transform the data into a high-dimensional feature space and then determine the optimal separating hyperplane by using a kernel function. In this work, we used the free software LIBSVM to predict peroxidase protein location. A radial basis function (RBF) was chosen as the kernel function. For multiclassification, SVM uses a one-versus-one strategy and constructs $k \times (k-1)/2$ classifiers and voting strategy to assign the class for an arbitrary protein sequence. Here various features of a protein sequence were utilized to perform a comprehensive study and achieve maximum accuracy.

PSSM profile of patterns

Evolutionary conservation usually reflects important biological function. An amino acid at a conserved site of a protein is preferred to locate at a functionally important region [36]. PSI-BLAST is a robust measure of residue conservation in a given location. Evolutionary information on protein sequences like PSSM can be created using a PSI-BLAST search. Compared to the compositional information, the PSSM profile provides more important information of evolutionary significance about residue conservation at a given position in a protein sequence [31,37]. In this study, the PSSM was generated using the PSI-BLAST search with a cutoff E value of 0.001 against the Swiss-Prot database.

The PSSM provides a matrix of dimension L rows and 20 columns for a protein chain with L amino acid residues, where 20 columns represent the occurrence/substitution of each type of 20 amino acids [38]. We summed all of the rows in the PSSM corresponding to the same amino acid in the sequence and then divided each element by the length of the sequence. In the prediction of peroxidase location, we used PSSM profiles with different similarities to generate 400 dimension (20×20 residue pairs) input vectors as parameters.

Composition profile of patterns

The aim of calculating the protein composition is to transform the variable lengths of the protein sequence to fixed-length vectors. This is an important and crucial step for protein classification using a computational approach because it requires a fixed-length pattern.

Amino acid and dipeptide compositions

The AAC representation of a given sequence is composed of 20 different amino acids with a variety of shapes, sizes, and chemical properties. A protein can be represented as a 20-dimensional (20D) vector according to AAC [22]. DC is the occurrence frequency of each of 2 adjacent amino acid residues. It is used to encapsulate the global information of each protein sequence, and a protein can be represented as a 400D vector by means of DC [39–41]. In this study, the AAC and DC of the N-part split amino acid composition were selected as classification vectors.

Split amino acid composition

In simple amino acid-, dipeptide-, and pseudo amino acid-based compositions, the composition is taken at once for the whole sequence, whereas in the split amino acid composition model, the protein sequence is divided into different parts and the composition of each part is calculated separately [25]. The composition is taken independently for the N parts of the protein sequence [42]. Hence, the advantage of SAAC over standard AAC is that it provides a greater weight of compositional biasness to proteins that have a signal at different sequence regions. In our SAAC model each protein is divided into 1 to 10 parts to train the optimal parameter combination for the SVM program.

Gene ontology profile of patterns

Gene Ontology is one of the databases that describes molecular function, and the molecular function of the GO database is correlated to the subcellular location [43]. Accordingly, protein sequences formulated in the GO database space would be clustered in a way that better reflects their subcellular locations [29]. However, to incorporate more information, instead of using only 0 and 1 element, as done in Ref. [44], here let us use a different approach as described below.

First, we searched for the homologous proteins of protein **P** from the Swiss-Prot database (released on 5 September 2012) using the PSI-BLAST method, with the expected value $E \leqslant 0.001$ for the BLAST parameter [31]. Second, we collected those proteins that had $\geqslant 60\%$ pairwise sequence identity with protein **P** into a subset, $\mathbf{P}^{\text{homo}}$, called the "homology set" of **P**. All the elements in $\mathbf{P}^{\text{homo}}$ could be deemed the "representative proteins" of **P,** sharing some similar attributes such as structural conformation and biological function. These representative proteins retrieved from the Swiss-Prot database must each have their own accession number. Third, we searched each of the accession numbers collected in the second step against the GO database to find the corresponding GO number. Last, we statistically analyzed each coordinate of the vector and found that many of the coordinates were equal to 0. This denoted that certain GOs did not belong to any protein; these GOs were eliminated, and the dimension of the GO feature vector was decreased in this manner.

Please cite this article in press as: Y.-C. Zuo et al., Predicting peroxidase subcellular location by hybridizing different descriptors of Chou' pseudo amino acid patterns, Anal. Biochem. (2014), http://dx.doi.org/10.1016/j.ab.2014.04.032