Journal of Theoretical Biology 411 (2016) 1–5

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Analysis and comparison of lignin peroxidases between fungi and bacteria using three different modes of Chou's general pseudo amino acid composition

Mandana Behbahani, Hassan Mohabatkar*, Mokhtar Nosrati

Department of Biotechnology, Faculty of Advanced Sciences and Technologies, University of Isfahan, Isfahan, Iran

HIGHLIGHTS

- The classification using PseAAC showed over 80% accuracy.
- The binding free energy of bacterial LPs is more than fungi LP.
- The computational predictors were useful tools to compare LPs.

ARTICLE INFO

Article history:
Received 25 May 2016
Received in revised form
27 July 2016
Accepted 1 September 2016
Available online 8 September 2016

Keywords:
Lignin peroxidase
Binary classification
Bioremediation

ABSTRACT

Lignin peroxidases (LiPs) are important enzymes in the degradation process of lignin which are presented in different species of fungi and bacteria. In the present study, sequence and structure-based properties of LPs in fungi and bacteria are compared. These properties include pseudo amino acid composition (PseAAC), physicochemical properties and the secondary structure. Autodock 4 has been used for docking between LiPs and lignan. The motifs of LiP were predicted by MEME tool. Statistical analysis and Multinomial Naïve Bayes (MNB) algorithm were used for the classification of two LiP protein groups. The results demonstrated that molecular weight, isoelectric point, aliphatic, extinction coefficient and random coil percentage of LiPs in fungi and bacteria were significantly different between these two groups. The classification of these two groups based on the concept of PseAAC showed over 80% accuracy. The binding free energy between bacterial LiPs and lignan is significantly more than fungi LiP and ligand. The aliphatic and instability of most important motifs of bacteria and fungi were significantly different. In conclusion, the results indicated that computational techniques could provide useful information for comparing fungal and bacterial LiPs. These results can also explain that there is a relationship between efficacy and physicochemical properties of LiPs.

1. Introduction

Lignin is known as one of the most recalcitrant aromatic polymers. It is mostly derived from wood and abounds on the earth. Lignin degradation is done by both fungi and bacteria. Recently, some lignin degrading organisms and their enzymes are used for environmental bioremediation and various industries (Arora and Gill, 2001; Bugg et al., 2011). Lignin peroxidases (LiP) and versatile peroxidases are demonstrated as important enzymes in the degradation of lignin (Eom and Kim, 2014). So far, a lot of researches have been conducted on some lignin-degrading fungi such as white-rot and brown-rot fungi. The use of fungi to degrade lignin is widely studied due to their environmental importance and potential biotechnological applications (Janusz et al., 2013). However these fungi have some industrial limitations including particular culture conditions and substrates (Crawford and Crawford, 1976). There are a number of bacteria such as Streptomycetes that can break down lignin (Zimmermann, 1990; Ramachandra et al., 1988). The degradation of lignin by bacteria may be superior to fungi regarding to specificity, thermostability and mediator dependency. Some studies have been focused on the evolutionary patterns of fungi peroxidase (Janusz et al., 2013; Johansson and Nyman, 1993). However, the diversity of LiPs in fungi and bacteria has not been investigated yet. In the present study, the pseudo amino acid composition (pseAAC), physicochemical properties and secondary structure of LiPs in fungi and bacteria are compared.

* Corresponding author.
E-mail address: h.mohabatkar@ast.ui.ac.ir (H. Mohabatkar).

http://dx.doi.org/10.1016/j.jtbi.2016.09.001
0022-5193/

2. Materials and methods

2.1. Data collection

Amino acid sequences of LiPs of fungi and bacteria were fetched from NCBI (http://www.ncbi.nlm.gov). One thousand six hundred sixty LiP amino acid sequences from bacteria and 122 sequences of fungi were used as our datasets. The data were analyzed by CD-Hit (Fu et al., 2012) and redundant sequences (more than 95% similarity sequences) were removed from the datasets. After running CD-HIT, the number of LP sequences in bacteria and fungi was reduced to 105 and 25 respectively.

2.2. General pseudo amino acid composition

With the explosive growth of biological sequences in the post-genomic era, one of the most important but also most difficult problems in computational biology is how to express a biological sequence with a discrete model or a vector, yet still keep considerable sequence-order information or key pattern characteristic. This is because all the existing machine-learning algorithm can only handle vector but not sequence samples, as elucidated in a recent review (Chou, 2015). However, a vector defined in a discrete model may completely lose all the sequence-pattern information. To avoid completely losing the sequence-pattern information for proteins, the pseudo amino acid composition or PseAAC (Chou, 2005) was propose. Ever since the concept of pseudo amino acid composition or Chou's PseAAC (Du et al., 2012; Cao et al., 2013; Lin and Lapointe, 2013) was proposed in 2001, it has penetrated into many biomedicine and drug development areas (Zhong and Zhou, 2014; Zhou and Zhong, 2016) and nearly all the areas of computational proteomics (Kabir and Hayat, 2016; Dehzangi et al., 2015; Kumar et al., 2015; Mondal and Pai, 2014; Wang et al., 2015) as well as a long list of references cited in (Du et al., 2014). Because it has been widely and increasingly used, recently three powerful open access soft-wares, called 'PseAAC-Builder' (Du et al., 2012), 'propy' (Cao et al., 2013), and 'PseAAC-General' (Du et al., 2012), were established: the former two are for generating various modes of Chou's special PseAAC; while the 3rd one for those of Chou's general PseAAC (Chou, 2011, 2009), including not only all the special modes of feature vectors for proteins but also the higher level feature vectors such as "Functional Domain" mode (see Eqs. (9)–(10) of (Chou, 2015)), "Gene Ontology" mode (see Eqs. (11) and (12) of (Chou, 2015)), and "Sequential Evolution" or "PSSM" mode (see Eqs. (13) and (14) of (Chou, 2015)). Encouraged by the successes of using PseAAC to deal with protein/peptide sequences, three web-servers (Chen et al., 2014; Chen et al., 2015; Liu et al., 2015a, 2015b; Chen and Lin, 2015) were developed for generating various feature vectors for DNA/RNA sequences. Particularly, recently a very powerful web-server called Pse-in-One (Liu et al., 2015a, 2015b) has been established that can be used to generate any desired feature vectors for protein/peptide and DNA/RNA sequences according to the need of users' studies. In the current study, we are to use three different modes of the general PseAAC to analyze and compare the lignin peroxidases between fungi and bacteria.

2.2.1. Physicochemical property analysis

ProtParam is a primary structure prediction tool which computes different physicochemical properties of a protein. This tool is available at http://web.expasy.org/protparam (Gasteiger, 2005). In this study, five characteristics (molecular weight, theoretical pI, extinction coefficient, aliphatic index and grand average of hydropathicity) of LP proteins of fungi and bacteria were evaluated using ProtParam. The molecular weight, theoretical pI and aliphatic index are most studied and useful parameters that indicate physicochemical properties of a protein but extinction coefficient and grand average of hydropathicity are less studied in comparison to above mentioned parameters especially in comparative studies. The extinction coefficient indicates how much light a protein absorbs at a certain wavelength. It is useful to have an estimation of this coefficient for following a protein which a spectrophotometer when purifying it. The GRAVY value for a protein or a peptide is defined by the sum of hydropathy values of all amino acids divided by the protein length. Increasing positive score indicates a greater hydrophobicity.

2.2.2. Generating pseudo-amino acid composition (PseAAC)

The collected sequences of LiP proteins of fungi and bacteria were used to compute their PseAAC values. PseAAC was widely used in many earlier statistical methods for predicting various attributes of different proteins and peptides. The concept of PseAAC will describe protein sequences with quantitative representations while taking into account considerable sequence-order information (Chou, 2001). Therefore, PseAAC can provide a comprehensive combination with other properties to perform a reliable classification.

2.2.3. Secondary structure analysis

GOR IV is an efficient analysis tool for predicting the secondary structure of a protein from its amino acid sequence. This tool is available at http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_gor4.html (Kloczkowski et al., 2002). The frequency of secondary structure of LiPs of fungi and bacteria (alpha helix, extended strand and random coil) was computed and further analyzed.

2.3. Statistical analysis

The in silico comparison of physicochemical and secondary structure properties of LiPs in fungi and bacteria was evaluated using Receiver Operator Characteristic (ROC) curve analysis. Multinomial Naïve Bayes (MNB) classification algorithm was also used for evaluating the dissimilarity of datasets based on their PseAAC-generated values. ROC curve is a tool for organizing classifiers and visualizing their performance. The ROC graph analysis is usually used in machine learning and data mining research (Fawcett, 2006). ROC server can by computing accuracy (ACC) and Area under curve (AUC). Also ROC server can evaluate the differences between positive and negative classes of data.

2.4. Classification based on PseAAC

The dissimilarity between bacterial and fungal LiP protein sequences was characterized by the Multinomial Naive Bayes (MNB) algorithm in Weka Software version 3.7 (Hall et al., 2009). The performance of a binary classifier can be described by means of different parameters which are used here to analyze the classification. These are: ACC, precision or positive predictive value (PPV) and negative predictive value (NPV). PPV and NPV are the proportions of the two datasets results in binary classification and statistics and they are true positive and true negative results (Parikh et al., 2008). ACC is the most important parameter in describing a binary classification. When ACC reach over 0.80 (or 80%), the classification performance is evaluated to be acceptable.

2.5. Docking

The LiPs in fungi and bacteria also lignan were subjected to molecular docking study as target and ligand respectively. Lignan is a polyphenolic and a powerful competitive inhibitor of the Lip so can provide appropriate information from the active site of the