BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 7 2013, pages 960–962
doi:10.1093/bioinformatics/btt072

Systems biology

Advance Access publication February 19, 2013

propy: a tool to generate various modes of Chou's PseAAC

Dong-Sheng Cao[1], Qing-Song Xu[2] and Yi-Zeng Liang[1,*]

[1]Research Center of Modernization of Traditional Chinese Medicines, Central South University, and [2]School of Mathematics and Statistics, Central South University, Changsha 410083, People's Republic of China

Associate Editor: Trey Ideker

ABSTRACT

Summary: Sequence-derived structural and physiochemical features have been frequently used for analysing and predicting structural, functional, expression and interaction profiles of proteins and peptides. To facilitate extensive studies of proteins and peptides, we developed a freely available, open source python package called protein in python (propy) for calculating the widely used structural and physicochemical features of proteins and peptides from amino acid sequence. It computes five feature groups composed of 13 features, including amino acid composition, dipeptide composition, tripeptide composition, normalized Moreau–Broto autocorrelation, Moran autocorrelation, Geary autocorrelation, sequence-order-coupling number, quasi-sequence-order descriptors, composition, transition and distribution of various structural and physicochemical properties and two types of pseudo amino acid composition (PseAAC) descriptors. These features could be generally regarded as different Chou's PseAAC modes. In addition, it can also easily compute the previous descriptors based on user-defined properties, which are automatically available from the AAindex database.

Availability: The python package, propy, is freely available via http://code.google.com/p/protpy/downloads/list, and it runs on Linux and MS-Windows.

Contact: yizeng_liang@263.net

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on September 16, 2012; revised on January 15, 2012; accepted on February 7, 2013

1 INTRODUCTION

Sequence-derived structural and physicochemical features have been widely used in the development of machine learning models for predicting protein structural and functional classes (Chou, 2001, 2009), protein–protein interactions (Shen *et al*., 2007), protein–ligand interactions (Yu *et al*., 2012), subcellular locations and peptides of specific properties (Chou and Shen, 2008). These features are highly useful for representing and distinguishing proteins or peptides of different structural, functional and interaction profiles. Currently, these structural and physicochemical features of proteins and peptides were routinely used to characterize target proteins in drug–target pairs and predict new drug–target associations to identify potential drug targets (He *et al*., 2010), following the spirit of chemogenomics.

Several programs for computing protein structural and physicochemical features have been developed (Du *et al*., 2012; Holland *et al*., 2008; Li *et al*., 2006); however, they are not comprehensive and can only be limited to a certain kind of features. Additionally, these are not freely and easily accessible.

We implemented a selection of sophisticated protein features and provide them as a package for the free and open source software environment python. The propy package aims at providing the user with comprehensive implementations of these descriptors in a unified framework to allow easy and transparent computation. To our knowledge, propy is the first open source package computing a large number of protein features based on user-defined structural and physicochemical properties. We recommend propy to analyse and represent the proteins or peptides under investigation. Further, we hope that the package will be helpful when exploring questions concerning the structures, functions and interactions of proteins and peptides in the context of systems biology.

2 PACKAGE DESCRIPTION

The propy package can compute a large number of structural and physicochemical features from amino acid sequence. A list of features for proteins and peptides covered by the current version of propy is summarized in Table 1. These features can be divided into five groups, each of which has been independently predicting protein- and peptide-related problems by using machine-learning methods. The first group includes three features, amino acid composition, dipeptide composition and tripeptide composition, with three descriptors and 8420 descriptor values. The second group consists of three different autocorrelation features: normalized Moreau–Broto autocorrelation, Moran autocorrelation and Geary autocorrelation. The autocorrelation features describe the level of correlation between two protein or peptide sequences in terms of their specific structural or physicochemical property. Each of these features has eight descriptors and 240 descriptor values. The third group contains three feature sets: composition, transition and distribution with 21 descriptors and 147 descriptor values. They represent the amino acid distribution pattern of a specific structural or physicochemical property along a protein or peptide sequence. Seven types of physicochemical properties have been used for calculating these features (Supplementary Material). The fourth group includes two sequence-order feature sets, one is sequence-order-coupling number with two descriptors and 60 descriptor values, and the other is quasi-sequence-order with two descriptors and 100 descriptor values. These features are derived from both Schneider–Wrede physicochemical distance matrix and Grantham chemical distance matrix. The fifth group contains two types of pseudo-amino acid compositions

*To whom correspondence should be addressed.

Table 1. List of various Chou's PseAAC modes of proteins and peptides by propy

Feature groups	Features	No. of descriptors
Amino acid composition	Amino acid composition	20
	Dipeptide composition	400
	Tripeptide composition	8000
Autocorrelation	Normalized Moreau–Broto autocorrelation	240 [a]
	Moran autocorrelation	240 [a]
	Geary autocorrelation	240 [a]
Composition, transition and distribution	Composition	21
	Transition	21
	Distribution	105
Quasi-sequence order	Sequence-order-coupling number	60
	Quasi-sequence-order descriptors	100
Pseudo-amino acid composition	Type I pseudo-amino acid composition	50 [a]
	Type II pseudo-amino acid composition	50 [a]

[a]The number depends on the choice of the number of properties of amino acid and the choice of the parameter values in algorithms.

(PseAAC): type I PseAAC with 50 descriptor values and type II PseAAC (i.e. amphiphilic PseAAC) with 50 descriptor values. Apart from these descriptors, it can also compute previous descriptors based on user-defined properties, which are easily accessible from the AAindex database (Kawashima and Kanehisa, 2000). In fact, the aforementioned features can be regarded as different Chou's PseAAC modes. For example, amino acid, dipeptide, tripeptide or n-mer peptide ($n = 4, 5, \ldots$) compositions are just different modes of Chou's PseAAC. Moreover, the higher-level features, such as GO (Gene Ontology) information, FunD (Functional Domain) information and sequential evolution information, are also skilfully fused into the Chou's PseAAC descriptors to characterize different protein information, which is widely used for solving various biological problems. An excellent review by Chou (2011) has pointed out their relevancy.

The propy package contains several functions and modules manipulating proteins and peptides. To obtain protein sequences easily, propy provides a download module, by which the user could easily get protein sequences from the Uniprot website by providing Uniprot IDs or a file containing Uniprot IDs. A check module is also provided to ensure that our input for subsequent calculation is reliable. To facilitate the accessibility of the property or distance matrix of amino acids, propy provides an AAIndex module, which helps the user automatically download the needed property from the AAindex database. There are two means to compute these structural and physicochemical features from protein or peptide sequences. One is to use the built-in modules in the propy package. There exist five modules responding to the calculation of descriptors from five feature groups. The instruction for each module is provided in the form of HTML in propy. We could import related functions to compute these features as needed. The other is to call the GetProDes class by importing the PyPro module, which encapsulates commonly used descriptor calculation methods. We could construct a GetProDes object with a protein sequence input, and then call corresponding methods to calculate these features. A user guide for the use of propy is included in propy to guide how the user uses it to calculate the needed features (Supplementary Material). Additionally, the main advantage of propy is that the users themselves could specify some sets of amino acid properties in the form of dictionary (a data structure in python). More conveniently, the output from the AAIndex module could be directly used as the user-defined property to calculate the aforementioned descriptors, greatly enlarging the applications to our calculated features.

propy is written by the pure python language. We chose to use python because it is open source, and there already exist packages to handle proteins [e.g. Biopython (Cock *et al.*, 2009), PyMol and Pythonscape]. It is convenient for propy to analyse proteins and peptides processed by Biopython. Moreover, it only needs the support of some built-in modules in python. This greatly facilitates the transplantation and applications of the propy package. The use of the dictionary data structure in the propy output makes the users clearly understand the meaning of each feature.

3 DISCUSSION

Sequence analysis of proteins and peptides has become more and more important in various bioinformatics fields. Apart from the prediction of structural and functional classes of proteins or peptides, there exist a few stand-alone applications to calculate protein/peptide descriptors, which are designed to work with drug descriptors in the chemogenomics framework.

propy contains a selection of various Chou's PseAAC descriptors to analyse, classify and compare complex proteins and peptides. They facilitate to exploit machine-learning techniques to drive hypothesis from complex protein or peptide datasets. The usefulness of the features covered by propy for computing the structural and physicochemical features of proteins and peptides has been validated by a number of published studies (Chou, 2009, 2011). The propy implementation of each of these algorithms was extensively tested by using a number of test sequences. The computed descriptor values were compared with the known values for these sequences to ensure that our computation is accurate.

propy is a powerful open source package for the extraction of features of proteins and peptides. In our future work, we plan to

apply the integrated features on various biological research questions and extend the range of functions with new promising descriptors for the coming versions of propy.

ACKNOWLEDGEMENT

The authors thank two anonymous referees for their constructive comments, which greatly helped improve on the original version of the manuscript.

Funding: National Natural Science Foundation of China (21075138, 21275164 and 11271374). The studies meet with the approval of the university's review board.

Conflict of Interest: none declared.

REFERENCES

Chou,K.C. and Shen,H.B. (2008) Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms. *Nat. Protoc.*, **3**, 153–162.

Chou,K.C. (2001) Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins*, **43**, 246–255.

Chou,K.C. (2009) Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. *Curr. Proteomics*, **6**, 262–274.

Chou,K.C. (2011) Some remarks on protein attribute prediction and pseudo amino acid composition (50th Anniversary Year Review). *J. Theor. Biol.*, **273**, 236–247.

Cock,P.J.A. *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.

Du,P.F. *et al.* (2012) PseAAC-builder: a cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions. *Anal. Biochem.*, **425**, 117–119.

He,Z. *et al.* (2010) Predicting drug-target interaction networks based on functional groups and biological features. *PLoS ONE*, **5**, e9603.

Holland,R.C.G. *et al.* (2008) BioJava: an open-source framework for bioinformatics. *Bioinformatics*, **24**, 2096–2097.

Kawashima,S. and Kanehisa,M. (2000) AAindex: amino acid index database. *Nucleic Acids Res.*, **28**, 374.

Li,Z.R. *et al.* (2006) PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Res.*, **34**, W32–W37.

Shen,J. *et al.* (2007) Predicting protein-protein interactions based only on sequences information. *Proc. Natl Acad. Sci. U S A*, **104**, 4337–4341.

Yu,H. *et al.* (2012) A systematic prediction of multiple drug-target interactions from chemical, genomic, and pharmacological data. *PLoS ONE*, **7**, e37608.