Bioinformatics, 31(11), 2015, 1857–1859
doi: 10.1093/bioinformatics/btv042
Advance Access Publication Date: 24 January 2015
Applications Note

Systems biology

protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences

Nan Xiao[1], Dong-Sheng Cao[2,]*, Min-Feng Zhu[1] and Qing-Song Xu[1,]*

[1]School of Mathematics and Statistics and [2]School of Pharmaceutical Sciences, Central South University, Changsha 410083, People's Republic of China

*To whom correspondence should be addressed.

Associate Editor: Ziv Bar-Joseph

Received on October 9, 2014; revised on December 23, 2014; accepted on January 18, 2015

Abstract

Summary: Amino acid sequence-derived structural and physiochemical descriptors are extensively utilized for the research of structural, functional, expression and interaction profiles of proteins and peptides. We developed protr, a comprehensive R package for generating various numerical representation schemes of proteins and peptides from amino acid sequence. The package calculates eight descriptor groups composed of 22 types of commonly used descriptors that include about 22 700 descriptor values. It allows users to select amino acid properties from the AAindex database, and use self-defined properties to construct customized descriptors. For proteochemometric modeling, it calculates six types of scales-based descriptors derived by various dimensionality reduction methods. The protr package also integrates the functionality of similarity score computation derived by protein sequence alignment and Gene Ontology semantic similarity measures within a list of proteins, and calculates profile-based protein features based on position-specific scoring matrix. We also developed ProtrWeb, a user-friendly web server for calculating descriptors presented in the protr package.

Availability and implementation: The protr package is freely available from CRAN: http://cran.r-project.org/package=protr, ProtrWeb, is freely available at http://protrweb.scbdd.com/.

Contact: oriental-cds@163.com or dasongxu@gmail.com

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

1 Introduction

Protein sequence is the ultimate resource for functional protein research. In order to apply various machine learning approaches on protein sequence data, it is common practice to encode sequence information as numerical features. The type of encoding, however, can significantly affect analyses, and choosing a precise and effective encoding is a critical step (Chou, 2011).

In bioinformatics, sequence-derived structural and physicochemical features have been widely applied in the research of protein structure and functions during the past two decades, such as, protein structural and functional classes (Chou and Fasman, 2006), protein–protein interactions (Cao *et al.*, 2014; Shen *et al.*, 2007), subcellular locations and peptides of specific properties (Chou and Shen, 2008), and post-translational modifications (Xu *et al.*, 2013). In chemogenomics, these structural and physicochemical descriptors are also routinely used to characterize target proteins in drug–target pairs for potential drug–target interaction discovery (Cao *et al.*, 2012, 2013a, b). In proteochemometric (PCM) modeling (Wikberg *et al.*, 2004), continuous descriptors derived by various molecular descriptor sets and dimensionality reduction methods (van Westen *et al.*, 2013a, b) are successfully employed in several applications, such as DNA binding pattern analysis, compound selection in lead optimization (van Westen *et al.*, 2011), novel ligand discovery (van Westen *et al.*, 2012), proteases ligand selectivity modeling

(Ain *et al.*, 2014) and predicting resistance to pesticides for agrochemicals (van Westen *et al.*, 2014).

Moreover, for protein and peptides, amino acid sequence and annotation-based similarity scores derived from sequence alignments and Gene Ontology (GO) annotation comparison are also useful representation schemes, which are widely used in modeling, such as genome-wide inference of protein–protein interactions (Zhang *et al.*, 2012).

Several web servers and stand-alone programs, such as PROFEAT (Li *et al.*, 2006), PseAAC (Shen and Chou, 2008), propy (Cao *et al.*, 2013c) have been established to calculate such structural and physicochemical descriptors. However, currently available solutions are often limited to certain types of descriptors, lack flexibility and usually difficult to seamlessly integrate into the predictive modeling pipeline. We still urgently need a comprehensive and flexible toolkit to calculate and customize these descriptors. Here, we introduce protr and ProtrWeb, the R package and web server for calculating various numerical representation schemes of protein and peptides from amino acid sequence. We recommend using protr to represent the proteins or peptides under investigation. Besides, in the context of systems biology, we hope that protr will be useful for exploring the biological questions about structures, functions and interactions of proteins and peptides.

2 Package description

The protr package calculates various commonly used structural and physicochemical descriptors and PCMs modeling descriptors for amino acid sequences. A list of descriptors for proteins covered by protr is summarized in Table 1. These descriptors can be generally divided into eight groups. The first group includes the amino acid composition, dipeptide composition, tripeptide composition. The second group consists of three types of autocorrelation descriptors: normalized Moreau–Broto autocorrelation, Moran autocorrelation and Geary autocorrelation. The third group contains the CTD (composition, transition, distribution) descriptors. The fourth group consists of the conjoint triad descriptor. The fifth group contains two sequence-order descriptor sets: sequence order coupling number and quasi-sequence order descriptor. The sixth group includes two types of pseudo-amino acid compositions (PseAAC): pseudo-amino acid composition (Type I PseAAC) and amphiphilic pseudo-amino acid composition (Type II PseAAC). The seventh group contains seven types of descriptors used for PCM modeling: including the scales-based descriptors derived by principal components analysis, factor analysis and multidimensional scaling, in combination with amino acid properties and 2D and 3D molecular descriptor sets, and BLOSUM/PAM matrix-derived descriptors. The eighth group calculates profile-based protein features based on position-specific scoring matrix (PSSM) (Su *et al.*, 2006). For constructing customized descriptors of certain types, protr supports defining user-specified properties and selecting properties from the AAindex database (Kawashima *et al.*, 2008). See the package vignette in the Supplementary data for computational details of the descriptors, datasets and the full workflow demonstration.

Similarity scores are another useful type of representation encoding the relational information between two proteins. In protr, we incorporated protein sequence alignment (Pages *et al.*, 2014) and GO semantic similarity measures computation (Yu *et al.*, 2010) to derive similarity scores. The parallelized version of functions for computing the pairwise similarity scores are provided to accelerate the computation speed.

Furthermore, protr provided several useful auxiliary functions, such as functions for loading sequences from FASTA/PDB files, batch downloading protein sequences from UniProt, amino acid type sanity checking, partitioning sequences to create sliding windows, etc. These functions make the tasks of protein sequence data retrieval, pre-processing and manipulation easier in R. For users without recourse to R scripting and requiring *ad hoc* analysis of

Table 1. List of various descriptors calculated by protr

Descriptor groups	Descriptor	Number
Amino acid composition	Amino acid composition	20
	Dipeptide composition	400
	Tripeptide composition	8000
Autocorrelation	Normalized Moreau-Broto	240[a]
	Moran	240[a]
	Geary	240[a]
CTD	Composition	21
	Transition	21
	Distribution	105
Conjoint Triad	Conjoint Triad	343
Quasi-sequence-order	Sequence-order-coupling number	60[a]
	Quasi-sequence-order descriptors	100[a]
Pseudo-amino acid composition	Type I	50[a]
	Type II	80[a]
Proteochemometric descriptors	Principal components analysis (amino acid properties based)	175[b]
	Principal components analysis (2D and 3D molecular descriptors based)	4025[b]
	Factor analysis (amino acid properties based)	175[b]
	Factor analysis (2D and 3D molecular descriptors based)	4025[b]
	Multidimensional scaling (amino acid properties based)	175[b]
	Multidimensional scaling (2D and 3D molecular descriptors based)	4025[b]
	BLOSUM and PAM matrix-derived descriptors	175[b]
PSSM	PSSM profile	–

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

[b] The number of descriptor values depends on the choice of the number of components and the choice of the lag parameter

protein sequences, we offered ProtrWeb, an easy-to-use web server for calculating the commonly used descriptors presented in protr.

3 Results

To the best of our knowledge, protr is currently the most comprehensive, flexible and integrated open-source toolkit for protein sequence-derived structural and physiochemical descriptor computation. Users can select appropriate descriptors calculated by protr or ProtrWeb according to their needs, and conveniently apply various statistical analysis and machine learning methods in R to solve various biological questions concerning the structures, functions and interactions of proteins and peptides.

Users of the protr package need to intelligently evaluate the underlying details of the descriptors provided, instead of using protr with their data blindly. It would be wise to use some negative and positive control comparisons where relevant to help guide interpretation of the results.

The protr package has been intensively tested to guarantee the computation correctness and speed. To ensure that our calculation is accurate, the calculated descriptor values were compared with the known values for these sequences.

In future development of protr, it is a potential direction to incorporate 3D structural information of proteins (Grant *et al.*, 2006), which would be beneficial in several analysis and modeling scenerios.

Funding

This study was supported by the National Key Basic Research Program [2015CB910700] and the National Natural Science Foundation of China [Grants Nos. 81402853 and 11271374] and the Postdoctoral Science Foundation of Central South University. The studies meet with the approval of the university's review board.

Conflict of Interest: none declared.

References

Ain,Q.U. *et al.* (2014) Modelling ligand selectivity of serine proteases using integrative proteochemometric approaches improves model performance and allows the multi-target dependent interpretation of features. *Integr. Biol.*, **6**, 1023–1033.

Cao,D.-S. *et al.* (2012) Large-scale prediction of drug-target interactions using protein sequences and drug topological structures. *Anal. Chim. Acta.*, **752**, 1–10.

Cao,D.-S. *et al.* (2013a) Genome-scale screening of drug-target associations relevant to ki using a chemogenomics approach. *PLoS ONE*, **8**, e57680.

Cao,D.-S. *et al.* (2013b) PyDPI: freely available Python package for chemoinformatics, bioinformatics, and chemogenomics studies. *J. Chem. Inf. Model.*, **53**, 3086–3096.

Cao,D.-S. *et al.* (2013c) propy: a tool to generate various modes of Chou's PseAAC. *Bioinformatics*, **29**, 960–962.

Cao,D.-S. *et al.* (2015) Rcpi: R/Bioconductor package to generate various descriptors of proteins, compounds and their interactions. *Bioinformatics*, **31**, 279–281.

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

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,P.Y. and Fasman,G.D. (2006) *Prediction of the Secondary Structure of Proteins From Their Amino Acid Sequence*. Wiley Online Library.

Grant,B.J. *et al.* (2006) Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics*, **22**, 2695–2696.

Kawashima,S. *et al.* (2008) AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res.,* **36**(Suppl. 1), D202–D205.

Li,Z. *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**(Suppl. 2), W32–W37.

Pages,H. *et al.* (2014) Biostrings: string objects representing biological sequences, and matching algorithms. R package version 2.30.1.

Shen,H.-B. and Chou,K.C. (2008) PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition. *Anal. Biochem.*, **373**, 386–388.

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

Su,C.T. *et al.* (2006) Protein disorder prediction by condensed PSSM considering propensity for order or disorder. *BMC Bioinformatics*, **7**, 319.

van Westen,G.J. *et al.* (2013a) Benchmarking of protein descriptor sets in proteochemometric modeling (part 1): comparative study of 13 amino acid descriptor sets. *J. Cheminform.*, **5**, 41.

van Westen,G.J. *et al.* (2013b) Benchmarking of protein descriptor sets in proteochemometric modeling (part 2): modeling performance of 13 amino acid descriptor sets. *J. Cheminform.*, **5**, 42.

van Westen,G.J. *et al.* (2011) Which compound to select in lead optimization? Prospectively validated proteochemometric models guide preclinical development. *PLoS ONE*, **6**, e27518.

van Westen,G.J. *et al.* (2012) Identifying novel adenosine receptor ligands by simultaneous proteochemometric modeling of rat and human bioactivity data. *J. Med. Chem.*, **55**, 7010–7020.

van Westen,G.J. *et al.* (2014) Towards predictive resistance models for agrochemicals by combining chemical and protein similarity via proteochemometric modelling. *J. Chem. Biol.*, **7**, 119–123.

Wikberg,J.E. *et al.* (2004) Proteochemometrics: a tool for modeling the molecular interaction space. In: Kubinyi,H. and Müller,G. (eds) *Chemogenomics in Drug Discovery*. Wiley Online Library, pp 289–309.

Xu,Y. *et al.* (2013) iSNO-PseAAC: predict cysteine s-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition. *PLoS ONE*, **8**, e55844.

Yu,G. *et al.* (2010) GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics*, **26**, 976–978.

Zhang,Q.C. *et al.* (2012) Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature*, **490**, 556–560.