Bioinformatics, 34(14), 2018, 2499–2502
doi: 10.1093/bioinformatics/bty140
Advance Access Publication Date: 8 March 2018
Applications Note

OXFORD

Sequence analysis

iFeature: a Python package and web server for features extraction and selection from protein and peptide sequences

Zhen Chen[1,†], Pei Zhao[2,†], Fuyi Li[3], André Leier[4,5], Tatiana T. Marquez-Lago[4,5], Yanan Wang[6], Geoffrey I. Webb[7], A. Ian Smith[3], Roger J. Daly[3,*], Kuo-Chen Chou[8,9,*] and Jiangning Song[3,7,*]

[1]School of Basic Medical Science, Qingdao University, 38 Dengzhou Road, Qingdao, 266021, China, [2]State Key Laboratory of Cotton Biology, Institute of Cotton Research of Chinese Academy of Agricultural Sciences (CAAS), Anyang, 455000, China, [3]Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, Melbourne, VIC 3800, Australia, [4]Department of Genetics and [5]Department of Cell, Developmental and Integrative Biology, School of Medicine, University of Alabama at Birmingham, AL, USA, [6]Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China, [7]Monash Centre for Data Science, Faculty of Information Technology, Monash University, Melbourne, VIC 3800, Australia, [8]Gordon Life Science Institute, Boston, MA 02478, USA and [9]Center for Informational Biology, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, China

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Alfonso Valencia

Received on December 2, 2017; revised on February 15, 2018; editorial decision on March 4, 2018; accepted on March 6, 2018

Abstract

Summary: Structural and physiochemical descriptors extracted from sequence data have been widely used to represent sequences and predict structural, functional, expression and interaction profiles of proteins and peptides as well as DNAs/RNAs. Here, we present *iFeature*, a versatile Python-based toolkit for generating various numerical feature representation schemes for both protein and peptide sequences. *iFeature* is capable of calculating and extracting a comprehensive spectrum of 18 major sequence encoding schemes that encompass 53 different types of feature descriptors. It also allows users to extract specific amino acid properties from the AAindex database. Furthermore, *iFeature* integrates 12 different types of commonly used feature clustering, selection and dimensionality reduction algorithms, greatly facilitating training, analysis and benchmarking of machine-learning models. The functionality of *iFeature* is made freely available via an online web server and a stand-alone toolkit.

Availability and implementation: http://iFeature.erc.monash.edu/; https://github.com/Superzchen/iFeature/.

Contact: jiangning.song@monash.edu or kcchou@gordonlifescience.org or roger.daly@monash.edu

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

1 Introduction

In recent years, machine learning techniques have been increasingly used as a powerful means to predict structural and functional properties of proteins and to assist in the annotation of genomic and proteomic data (Larranaga *et al.*, 2006; Libbrecht and Noble, 2015). In this regard, it has proven crucial to transform protein and peptide sequences into effective mathematical expressions that describe their intrinsic correlation with the corresponding structural and functional attributes (Chou, 2011). Over the past decades, an increasing number of diverse feature encoding methods or descriptors extracted from protein and peptide sequence information have been proposed for improving various predictions. Applications include predicting protein structural and function classes (Chou and Fasman, 1978), protein-protein interactions, protein–ligand interactions (Cao *et al.*, 2015; Shen *et al.*, 2007), subcellular locations (Chou and Shen, 2008), enzyme substrates (Barkan *et al.*, 2010; Rottig *et al.*, 2010; Song *et al.*, 2010), among others.

Several web servers and stand-alone software packages have been developed to calculate a variety of structural and physicochemical features, including PROFEAT (Li *et al.*, 2006; Rao *et al.*, 2011), PseAAC (Shen and Chou, 2008), PseAAC-Builder (Du *et al.*, 2012), propy (Cao *et al.*, 2013), PseAAC-General (Du *et al.*, 2014), protr/ProtrWeb (Xiao *et al.*, 2015), Rcpi (Cao *et al.*, 2015) and PseKRAAC (Zuo *et al.*, 2017). However, in addition to feature extraction, feature selection and ranking analysis is an equally crucial step in machine learning of protein structures and functions. To the best of our knowledge, there is no universal toolkit or web server currently available that integrates both functions of feature extraction and feature selection analysis. It is in this spirit that we developed *iFeature*, a versatile open-source Python toolkit that bridges this gap. *iFeature* can be used not only to extract a great variety of numerical feature encoding schemes from protein or peptide sequences, but also for feature clustering, ranking, selection and dimensionality reduction, all of which will greatly facilitate users' subsequent efforts to identify relevant features and construct effective machine learning-based models. In order to facilitate users' interpretability of outcomes, the clustering and dimensionality reduction results can be visualized in form of scatter diagrams. *iFeature* also supports the integration of different feature types, making it more convenient to train models by combining different feature groups. Lastly, we developed a user-friendly web server for *iFeature*.

2 Implementation

An important advantage of *iFeature* is that it integrates the multifaceted functionality of feature calculation, extraction, clustering, selection and dimensionality reduction analysis. A complete list of the 18 major encoding schemes is summarized in Table 1. We briefly discuss below.

The first group includes six feature sets, i.e. amino acid composition, composition of *k*-spaced amino acid pairs (Chen *et al.*, 2013; Liu *et al.*, 2017), enhanced amino acid composition, dipeptide composition, dipeptide deviation from expected mean (Saravanan and Gautham, 2015) and tripeptide composition (Bhasin and Raghava, 2004). The secondary group is labeled 'grouped amino acid composition', which also consists of five descriptors (Table 1). For this group, 20 amino acid types are first categorized according to their physicochemical properties, and then the composition of each category is calculated. The third group is the binary encoding scheme in which each amino acid is represented by a 20-dimensional binary vector. The fourth group includes three types of autocorrelation feature sets: normalized Moreau–Broto autocorrelation, Moran autocorrelation and Geary autocorrelation (Sokal and Thomson, 2006). This feature group allows users to select properties from the AAindex database (Kawashima *et al.*, 2008). The fifth group consists of three feature sets: composition, transition and distribution (Dubchak *et al.*, 1995, 1999). The sixth group is the conjoint triad (Shen *et al.*, 2007). The seventh group contains two sequence-order feature sets, sequence-order-coupling number and quasi-sequence-order (Chou, 2000; Chou and Cai, 2004; Schneider and Wrede, 1994). The eighth group includes the pseudo-amino acid composition and the amphiphilic pseudo-amino acid composition (Chou, 2001, 2005). The ninth group includes two K-nearest neighbor features: KNNprotein and KNNpeptide (Chen *et al.*, 2013). The tenth group is the PSSM encoding scheme, which extracts features from the position-specific scoring matrix (PSSM; Altschul, 1997) generated by PSI-BLAST. The eleventh group is the AAindex encoding scheme where each amino acid is represented by a 531-dimensional vector (Tung and Ho, 2008). The twelfth group is the BLOSUM matrix-derived descriptor (Lee *et al.*, 2011). The thirteenth group is the Z-scale encoding where each amino acid is represented by five physicochemical descriptor variables. Feature groups 14 to 17 are derived from information about the predicted protein secondary structure, disorder, accessible surface area and torsional angles, respectively. The last group includes 16 types of pseudo K-tuple reduced amino acid compositions (Zuo *et al.*, 2017).

Moreover, as high-dimensional features can potentially cause over fitting or high-dimensional disaster (Bellman and Bellman, 1961) and increase of redundant information, machine learning models trained using such high-dimensional initial features often perform poorly in practice. To solve this problem, *iFeature* further integrates several commonly used feature clustering, selection and dimensionality reduction algorithms to filter out redundant features and retain the useful and relevant ones. All implemented feature analysis algorithms are listed in Table 2. All clustering methods support sample and feature clustering procedures. In cases where users are not familiar with computer programming using Python, we also implemented an online web server of *iFeature*. It is configured on the extensible cloud computing facility supported by the e-Research Centre at Monash University, equipped with 16 cores, 64 GB memory and a 2 TB hard disk. This configuration can be easily upgraded in line with increasing user demands in the future.

3 Results

In this work, we have developed *iFeature*, a comprehensive, flexible and open-source Python toolkit for generating various sequences, structural and physiochemical features derived from protein/peptide sequences. *iFeature* also allows users to integrate various feature clustering, selection and dimensionality reduction algorithms that facilitate feature importance analysis, model training and benchmarking of machine learning-based models. *iFeature* has been extensively tested to guarantee correctness of computations, and was purposely designed to ensure workflow efficiency. To the best of our knowledge, this is the first universal toolkit for integrated feature calculation, clustering and selection analysis. In the future, we will integrate more analysis and clustering algorithms to enable interactive analysis and machine learning-based modeling. *iFeature* is expected to be widely used as a powerful tool in bioinformatics, computational biology and proteome research.

Table 1. List of various descriptors calculated by *iFeature*

Descriptor groups	Descriptor	Dimension
Amino acid composition	Amino acid composition (AAC)	20
	Enhanced amino acid composition (EAAC)	—
	Composition of *k*-spaced amino acid pairs (CKSAAP)	2400
	Dipeptide composition (DPC)	400
	Dipeptide deviation from expected mean (DDE)	400
	Tripeptide composition (TPC)	8000
Grouped amino acid composition	Grouped amino acid composition (GAAC)	5
	Enhanced grouped amino acid composition (GEAAC)	—
	Composition of *k*-spaced amino acid group pairs (CKSAAGP)	150
	Grouped dipeptide composition (GDPC)	25
	Grouped tripeptide composition (GTPC)	125
Binary	Binary (BINARY)	—
Autocorrelation	Moran (Moran)	240
	Geary (Geary)	240
	Normalized Moreau-Broto (NMBroto)	240
C/T/D	Composition (CTDC)	39
	Transition (CTDT)	39
	Distribution (CTDD)	195
Conjoint triad	Conjoint triad (CTriad)	343
	Conjoint *k*-spaced triad (KSCTriad)	343x(k+1)
Quasi-sequence-order	Sequence-order-coupling number (SOCNumber)	60
	Quasi-sequence-order descriptors (QSOrder)	100
Pseudo-amino acid composition	Pseudo-amino acid composition (PAAC)	50
	Amphiphilic PAAC (APAAC)	80
K-nearest neighbor	K-nearest neighbor for proteins (KNNprotein)	60
	K-nearest neighbor for peptide (KNNpeptide)	60
PSSM	Position-specific scoring matrix (PSSM) profile	—
AAindex	AAindex (AAINDEX)	—
BLOSUM62	BLOSUM62 matrix	—
Z-scale	Z-scale (ZSCALE)	—
Predicted secondary structure	Secondary structure elements content (SSEC)	3
	Secondary structure elements binary (SSEB)	—
Predicted protein disorder	Disorder (Disorder)	—
	Disorder content (DisorderC)	2
	Disorder binary (DicorderB)	—
Predicted accessible surface area	Accessible surface area (ASA)	—
Predicted main-chain torsional angles	Torsional angles (TS)	—
Pseudo K-tuple reduced amino acids composition	PseKRAAC (type1 to type16)	—

Table 2. A list of various feature clustering, selection and dimensionality reduction algorithms available in *iFeature*

Type of functionality	Algorithm
Feature clustering	*K*-means (*k*means)
	Hierarchical clustering (hcluster)
	Mean shift (meanshift)
	DBSCAN (dbscan)
	Affinity propagation (apc)
Feature selection	Chi-square test (CHI2)
	Information gain (IG)
	Mutual information (MIC)
	Pearson's correlation coefficient (pearsonr)
Dimensionality reduction	Principal component analysis (PCA)
	Latent Dirichlet allocation (LDA)
	t-Distributed Stochastic Neighbor Embedding (t-SNE)

Funding

This work was supported by grants from the Australian Research Council [ARC; LP110200333 and DP120104460], National Natural Science Foundation of China [NSFC; 31701142], National Health and Medical Research Council of Australia [NHMRC; APP1058540], the National Institute of Allergy and Infectious Diseases of the National Institutes of Health [R01 AI111965] and a Major Inter-Disciplinary Research (IDR) Grant Awarded by Monash University. A.L. and T.T.M.-L. were supported by Informatics startup packages through the UAB School of Medicine.

Conflict of Interest: none declared.

References

Altschul,S.F. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.

Barkan,D.T. *et al*. (2010) Prediction of protease substrates using sequence and structure features. *Bioinformatics*, **26**, 1714–1722.

Bellman,R.E. (1961) *Adaptive Control Processes: A Guided Tour*. Princeton University Press, Princeton, NJ.

Bhasin,M. and Raghava,G.P. (2004) Classification of nuclear receptors based on amino acid composition and dipeptide composition. *J. Biol. Chem.*, **279**, 23262–23266.

Cao,D.S. *et al.* (2013) 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.

Chen,X. *et al.* (2013) Incorporating key position and amino acid residue features to identify general and species-specific Ubiquitin conjugation sites. *Bioinformatics*, **29**, 1614–1622.

Chen,Z. *et al.* (2013) hCKSAAP_UbSite: improved prediction of human ubiquitination sites by exploiting amino acid pattern and properties. *Biochim. Biophys. Acta*, **1834**, 1461–1467.

Chou,K.C. (2000) Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. *Biochem. Biophys. Res. Commun.*, **278**, 477–483.

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

Chou,K.C. (2005) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics*, **21**, 10–19.

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 Cai,Y.D. (2004) Prediction of protein subcellular locations by GO-FunD-PseAA predictor. *Biochem. Biophys. Res. Commun.*, **320**, 1236–1239.

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. (1978) Prediction of the secondary structure of proteins from their amino acid sequence. *Adv. Enzymol. Relat. Areas Mol. Biol.*, **47**, 45–148.

Du,P. *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.

Du,P. *et al.* (2014) PseAAC-General: fast building various modes of general form of Chou's pseudo-amino acid composition for large-scale protein datasets. *Int. J. Mol. Sci.*, **15**, 3495–3506.

Dubchak,I. *et al.* (1995) Prediction of protein folding class using global description of amino acid sequence. *Proc. Natl. Acad. Sci. USA*, **92**, 8700–8704.

Dubchak,I. *et al.* (1999) Recognition of a protein fold in the context of the Structural Classification of Proteins (SCOP) classification. *Proteins*, **35**, 401–407.

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

Larranaga,P. *et al.* (2006) Machine learning in bioinformatics. *Brief. Bioinform.*, **7**, 86–112.

Lee,T.Y. *et al.* (2011) Incorporating distant sequence features and radial basis function networks to identify ubiquitin conjugation sites. *PLoS One*, **6**, e17331.

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.

Libbrecht,M.W. and Noble,W.S. (2015) Machine learning applications in genetics and genomics. *Nat. Rev. Genet.*, **16**, 321–332.

Liu,L.M. *et al.* (2017) iPGK-PseAAC: identify lysine phosphoglycerylation sites in proteins by incorporating four different tiers of amino acid pairwise coupling information into the general PseAAC. *Med. Chem.*, **13**, 552–559.

Rao,H.B. *et al.* (2011) Update of PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Res.*, **39**, W385–W390.

Rottig,M. *et al.* (2010) Combining structure and sequence information allows automated prediction of substrate specificities within enzyme families. *PLoS Comput. Biol.*, **6**, e1000636.

Saravanan,V. and Gautham,N. (2015) Harnessing computational biology for exact linear B-cell epitope prediction: a novel amino acid composition-based feature descriptor. *Omics*, **19**, 648–658.

Schneider,G. and Wrede,P. (1994) The rational design of amino acid sequences by artificial neural networks and simulated molecular evolution: de novo design of an idealized leader peptidase cleavage site. *Biophys. J.*, **66**, 335–344.

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

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.

Sokal,R.R. and Thomson,B.A. (2006) Population structure inferred by local spatial autocorrelation: an example from an Amerindian tribal population. *Am. J. Phys. Anthropol.*, **129**, 121–131.

Song,J. *et al.* (2010) Cascleave: towards more accurate prediction of caspase substrate cleavage sites. *Bioinformatics*, **26**, 752–760.

Tung,C.W. and Ho,S.Y. (2008) Computational identification of ubiquitylation sites from protein sequences. *BMC Bioinformatics*, **9**, 310.

Xiao,N. *et al.* (2015) protr/ProtrWeb: r package and web server for generating various numerical representation schemes of protein sequences. *Bioinformatics*, **31**, 1857–1859.

Zuo,Y. *et al.* (2017) PseKRAAC: a flexible web server for generating pseudo K-tuple reduced amino acids composition. *Bioinformatics*, **33**, 122–124.