

Update of PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence

H. B. Rao¹, F. Zhu^{1,2}, G. B. Yang³, Z. R. Li^{1,4,*} and Y. Z. Chen^{2,4}

¹College of Chemistry, Sichuan University, Chengdu, 610064, P. R. China, ²Department of Pharmacy, Bioinformatics and Drug Design Group, National University of Singapore, Singapore 117543, ³College of Chemical Engineering, Sichuan University, Chengdu 610064 and ⁴State Key Laboratory of Biotherapy, Sichuan University, Chengdu 610041, P. R. China

Received January 23, 2011; Revised March 17, 2011; Accepted April 12, 2011

ABSTRACT

Sequence-derived structural and physicochemical features have been extensively used for analyzing and predicting structural, functional, expression and interaction profiles of proteins and peptides. PROFEAT has been developed as a web server for computing commonly used features of proteins and peptides from amino acid sequence. To facilitate more extensive studies of protein and peptides, numerous improvements and updates have been made to PROFEAT. We added new functions for computing descriptors of protein–protein and protein–small molecule interactions, segment descriptors for local properties of protein sequences, topological descriptors for peptide sequences and small molecule structures. We also added new feature groups for proteins and peptides (pseudo-amino acid composition, amphiphilic pseudo-amino acid composition, total amino acid properties and atomic-level topological descriptors) as well as for small molecules (atomic-level topological descriptors). Overall, PROFEAT computes 11 feature groups of descriptors for proteins and peptides, and a feature group of more than 400 descriptors for small molecules plus the derived features for protein–protein and protein–small molecule interactions. Our computational algorithms have been extensively tested and used in a number of published works for predicting proteins of specific structural or functional classes, protein–protein

interactions, peptides of specific functions and quantitative structure activity relationships of small molecules. PROFEAT is accessible free of charge at <http://bidd.cz3.nus.edu.sg/cgi-bin/prof/protein/profnew.cgi>.

INTRODUCTION

Sequence-derived structural and physicochemical features are highly useful for representing and distinguishing proteins or peptides of different structural, functional and interaction properties, and have been widely used in developing methods and software for predicting protein structural and functional classes (1–7), protein–protein interactions (8–10), protein–ligand interactions (11,12), protein substrates (13,14), molecular binding sites on proteins (15–20), subcellular locations (21), protein crystallization propensity (22–24) and peptides of specific properties (25–30). Web servers, such as PROFEAT (31) and PseAAC (<http://www.csbio.sjtu.edu.cn/bioinf/PseAA/>) (32), have been built to facilitate the computation of protein and peptide features.

Nonetheless, some features important for studying proteins, peptides and molecular interactions have not been provided in these web servers. Examples of these features include atomic-level topological descriptors that are useful for structure–property correlations (33) and descriptors of total amino acid properties (TAAPs) that have been used for modeling protein conformational stability (34), ligand binding site structural features (35) and interaction with small molecules (36). Moreover, the descriptors provided in those available web servers are not suitable for analyzing local properties of sequence

*To whom correspondence should be addressed. Tel: 86-28-85406139; Fax: 86-28-85407797; Email: lizerong@scu.edu.cn

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

© The Author(s) 2011. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0>), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

subsections, and additional works are needed to use descriptors to study protein–protein and protein–ligand interactions. Therefore, it is desirable to provide segment descriptors for local properties of subsections of protein sequences, and descriptors that can be straightforwardly used for exploring protein–protein and protein–small molecule interactions.

We updated PROFEAT by adding new functions for computing descriptors of protein–protein and protein–small molecule interactions, segment descriptors for local properties of subsections of protein sequences, atomic-level topological descriptors for peptide sequences and small molecule structures, and topological polar surface areas of small molecules. Moreover, we added new feature groups such as pseudo-amino acid composition (PAAC), amphiphilic PAAC (APAAC), TAAPs, and atomic-level topological descriptors. The computational algorithms of these newly added feature groups have been extensively tested and used in a number of published works for predicting proteins and peptides of specific properties, protein–protein interactions, and quantitative structure activity relationships of small molecules. A list of publications using features covered by PROFEAT is provided in Supplementary Table S1 and in PROFEAT online server which can be accessed at http://bidd.cz3.nus.edu.sg/prof/part_of_publications.htm. PROFEAT homepage is shown in Figure 1. A list of features for proteins and peptides covered by this version of PROFEAT is summarized in Table 1 and a list of the topological descriptors for peptides and small molecules computed by PROFEAT is summarized in Supplementary Table S2.

METHODS FOR NEWLY ADDED FEATURES AND FUNCTIONS

PAAC descriptors

First, three variables are derived from the original hydrophobicity values $H_1^0(i)$, hydrophilicity values $H_2^0(i)$ and side chain masses $M^0(i)$ of 20 amino acids ($i = 1, 2, \dots, 20$) (32):

$$H_1(i) = \frac{H_1^0(i) - \sum_{i=1}^{20} \frac{H_1^0(i)}{20}}{\sqrt{\frac{\sum_{i=1}^{20} [H_1^0(i) - \sum_{i=1}^{20} \frac{H_1^0(i)}{20}]^2}{20}}} \quad (1)$$

$$H_2(i) = \frac{H_2^0(i) - \sum_{i=1}^{20} \frac{H_2^0(i)}{20}}{\sqrt{\frac{\sum_{i=1}^{20} [H_2^0(i) - \sum_{i=1}^{20} \frac{H_2^0(i)}{20}]^2}{20}}} \quad (2)$$

$$M(i) = \frac{M^0(i) - \sum_{i=1}^{20} \frac{M^0(i)}{20}}{\sqrt{\frac{\sum_{i=1}^{20} [M^0(i) - \sum_{i=1}^{20} \frac{M^0(i)}{20}]^2}{20}}} \quad (3)$$

Then, a correlation function can be computed as:

$$\Theta(R_i, R_j) = \frac{1}{3} \left\{ [H_1(R_i) - H_1(R_j)]^2 + [H_2(R_i) - H_2(R_j)]^2 + [M(R_i) - M(R_j)]^2 \right\} \quad (4)$$

from which, sequence order-correlated factors are defined as:

$$\begin{aligned} \theta_1 &= \frac{1}{N-1} \sum_{i=1}^{N-1} \Theta(R_i, R_{i+1}) \\ \theta_2 &= \frac{1}{N-2} \sum_{i=1}^{N-2} \Theta(R_i, R_{i+2}) \\ \theta_3 &= \frac{1}{N-3} \sum_{i=1}^{N-3} \Theta(R_i, R_{i+3}) \\ \theta_\lambda &= \frac{1}{N-\lambda} \sum_{i=1}^{N-\lambda} \Theta(R_i, R_{i+\lambda}), (\lambda < N) \end{aligned} \quad (5)$$

$\lambda (< N)$ is a parameter. Let f_i be the normalized occurrence frequency of 20 amino acids in the protein sequence, a set of $20+\lambda$ descriptors called the PAAC are defined as:

$$\begin{aligned} X_u &= \frac{f_u}{\sum_{i=1}^{20} f_i + w \sum_{j=1}^{\lambda} \theta_j} \text{ when } 1 \leq u \leq 20 \\ X_u &= \frac{w\theta_{u-20}}{\sum_{i=1}^{20} f_i + w \sum_{j=1}^{\lambda} \theta_j} \text{ when } 20+1 \leq u \leq 20+\lambda \end{aligned} \quad (6)$$

where w is the weighting factor for the sequence-order effect and is set to be $w = 0.05$ as suggested by Shen (32).



APAAC

From $H_1(i)$ and $H_2(i)$ defined in Equation (1) and (2), the hydrophobicity and hydrophilicity correlation functions are defined (32), respectively, as:

$$H_{i,j}^1 = H_1(i)H_1(j), \quad H_{i,j}^2 = H_2(i)H_2(j)$$

from which sequence order factors can be defined as:

$$\begin{aligned} \tau_1 &= \frac{1}{N-1} \sum_{i=1}^{N-1} H_{i,i+1}^1, \quad \tau_2 = \frac{1}{N-1} \sum_{i=1}^{N-2} H_{i,i+1}^2, \\ \tau_3 &= \frac{1}{N-2} \sum_{i=1}^{N-2} H_{i,i+2}^1, \quad \tau_4 = \frac{1}{N-2} \sum_{i=1}^{N-2} H_{i,i+2}^2, \dots, \\ \tau_{2\lambda-1} &= \frac{1}{N-\lambda} \sum_{i=1}^{N-\lambda} H_{i,i+\lambda}^1, \quad \tau_{2\lambda} = \frac{1}{N-\lambda} \sum_{i=1}^{N-\lambda} H_{i,i+\lambda}^2 \quad (\lambda < N) \end{aligned} \quad (7)$$



PROFEAT- Protein Feature Server (2011)

PROFEAT is developed as a web server for computing commonly used features of proteins and peptides from amino acid sequence and of small molecules from molecular structure.

You can choose to calculate feature vector for:

(A) [Protein](#) (B) [Protein-Protein Interaction Pair](#)
(C) [Small Molecule](#) (D) [Protein-Ligand Interaction Pair](#)

For introduction to PROFEAT, please see the [Reference Manual](#)

In this page, PROFEAT is designed for computing physicochemical properties of proteins and peptides from their primary sequences.

Sequence

Sequence **MUST** be provided in [RAW](#) or [FASTA](#) format

Upload Sequences

Batch Query: maximum 1000 sequences in [FASTA](#) format

New features 2011 version:

- (1) Descriptors of protein-protein interactions;
- (2) Descriptors of protein-small molecule interactions;
- (3) Segment descriptors for local properties of subsections of protein sequences;
- (4) Atomic-level topological descriptors for peptide sequences;
- (5) Atomic-level topological descriptors for small molecule structures;
- (6) Topological polar surface areas of small molecules;
- (7) Pseudo amino acid descriptor for protein sequences.

If you find any error or bug in this web service, please kindly report to [Dr. Zhu](#).

46574 visits since November 6, 2005

Figure 1. PROFEAT new web page.

Table 1. List of PROFEAT computed features for proteins, peptides and protein–protein interactions

Feature group	Features	No. of descriptors	No. of descriptor values
Composition-1	Amino acid composition	1	20
Composition-2	Dipeptide composition	1	400
Autocorrelation 1	Normalized Moreau–Broto autocorrelation	^a	^a
Autocorrelation 2	Moran autocorrelation	^a	^a
Autocorrelation 3	Geary autocorrelation	^a	^a
Composition, Transition, Distribution	Composition	7	21
	Transition	7	21
	Distribution	7	105
Quasi-sequence order descriptors	Sequence order coupling number	2	90
	Quasi-sequence order descriptors	2	150
PAAC	PAAC	^b	^b
APAAC	APAAC	^c	^c
Topological descriptors	Topological descriptors		405
TAAPs	TAP	^d	^d

^aThe number depends on the choice of the number of properties of amino acid and the choice of the maximum values of the lag.

^bThe number depends on the choice of the number of the set of amino acid properties and the choice of the λ value.

^cThe number depends on the choice of the λ value.

^dThe numbers depend on the choice of the number of properties of amino acid.

and APAAC are defined as:

$$p_u = \frac{f_u}{\sum_{i=1}^{20} f_i + w \sum_{j=1}^{2\lambda} \tau_j} \text{ when } 1 \leq u \leq 20$$

$$p_u = \frac{w\tau_u}{\sum_{i=1}^{20} f_i + w \sum_{j=1}^{2\lambda} \tau_j} \text{ when } 20+1 \leq u \leq 20+\lambda$$
(8)

where w is the weighting factor and is taken as $w = 0.5$.

Topological descriptors at atomic level

Topological descriptors are based on graph theory and encode information about the types of atoms and bonds in a molecule and the nature of their connections. Examples of topological descriptors include counts of atom and bond types and indexes that encode the size, shape and types of branching in a molecule (37). These descriptors can be calculated from the 2D structure of a peptide automatically generated from its sequence based on the molecular structures of the amino acid residues in the sequence. Supplementary Table S2 gives a list of the topological descriptors computed by PROFEAT.

TAAP

TAAP descriptor for a specific physicochemical property i is defined as: $P_{tot(i)} = \sum_{j=1}^N p_{norm_j}^i / N$, where $p_{norm_j}^i$ represents the property i of amino acid R_j that is normalized between 0 and 1 using the following expression, $p_{norm_j}^i = (p_j^i - p_{min}^i) / (p_{max}^i - p_{min}^i)$, where p_j^i is the original amino acid property i for residue j . p_{max}^i and p_{min}^i are, respectively, the minimum and maximum values of the original amino acid property i , and N is the length of the sequence (38–40).

Protein–protein interaction descriptors

Protein–protein interaction descriptors can be computed from the descriptors $V_a = \{V_a(i), i = 1, 2, \dots, n\}$ and $V_b = \{V_b(i), i = 1, 2, \dots, n\}$ of individual proteins A and

B by three methods. In the first method, two protein-pair vectors V_{ab} and V_{ba} with dimension of $2n$ are constructed with $V_{ab} = (V_a, V_b)$ for interaction between proteins A and B and $V_{ba} = (V_b, V_a)$ for interaction between proteins B and A (8,9). In the second method, one vector V with dimension of $2n$ is constructed: $V = \{V_a(i) + V_b(i), V_a(i) \times V_b(i), i = 1, 2, \dots, n\}$ which has the property that V is unchanged when a and b are exchanged. In the third method, one vector V with dimension of n^2 is constructed by the tensor product: $V = \{V(k) = V_a(i) \times V_b(j), i = 1, 2, \dots, n, j = 1, 2, \dots, n, k = (i - 1) \times n + j\}$.

Protein–ligand interaction descriptors

Protein–ligand interaction descriptor vector V can be constructed from the protein descriptor vector $V_p (V_p(i), i = 1, \dots, n_p)$ and ligand descriptor vector $V_l (V_l(i), i = 1, \dots, n_l)$ by two methods similar to the first and third method for constructing protein pair descriptors. In the first method, one vector V with dimension of $n_p + n_l$ are constructed $V = (V_p, V_l)$ for interaction between protein and ligand. In the second method, one vector V with dimension of $n_p \times n_l$ is constructed by the tensor product: $V = \{v(k) = V_p(i) \times V_l(j), i = 1, 2, \dots, n_p, j = 1, 2, \dots, n_l, k = (i - 1) \times n_p + j\}$.

Segmented sequence descriptors

To characterize the local feature of a protein sequence, a protein sequence can be divided into several segments and descriptors are calculated for each segment.

Topological descriptors for small molecules

For small molecules, topological descriptors are calculated from the input 2D structures of small molecules in mol or sdf format. Names of these descriptors are the same as those for protein segments which are listed in Supplementary Table S2.

REMARKS

Compared with its earlier version, the updated PROFEAT is significantly enhanced in both the number of newly added features useful for representing various protein properties, and newly added functions for computing features for local properties of protein segments, protein-protein interactions, protein-small molecule interactions and small molecules. These enhancements are intended to provide more comprehensive features for facilitating the analysis and prediction of proteins, peptides, small molecules of different properties and molecular interactions involving proteins, peptides and small molecules. With continued interest in using molecular and interaction features and developing new algorithms for representing these features, new descriptors and functions such as those involving DNA, RNA and other nucleotides can be integrated into PROFEAT in the near future to better facilitate the study of molecular and bio-molecular functions and interactions.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Funding for open access charge: National Natural Science Foundation of China (grant number 20973118).

Conflict of interest statement. None declared.

REFERENCES

- Karchin,R., Karplus,K. and Haussler,D. (2002) Classifying G-protein coupled receptors with support vector machines. *Bioinformatics*, **18**, 147–159.
- Cai,C.Z., Han,L.Y., Ji,Z.L., Chen,X. and Chen,Y.Z. (2003) SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic Acids Res.*, **31**, 3692–3697.
- Dubchak,I., Muchnik,I., Mayor,C., Dralyuk,I. and Kim,S.H. (1999) Recognition of a protein fold in the context of the Structural Classification of Proteins (SCOP) classification. *Proteins*, **35**, 401–407.
- Han,L., Cui,J., Lin,H., Ji,Z., Cao,Z., Li,Y. and Chen,Y. (2006) Recent progresses in the application of machine learning approach for predicting protein functional class independent of sequence similarity. *Proteomics*, **6**, 4023–4037.
- Langlois,R.E. and Lu,H. (2010) Boosting the prediction and understanding of DNA-binding domains from sequence. *Nucleic Acids Res.*, **38**, 3149–3158.
- Yan,R.X., Si,J.N., Wang,C. and Zhang,Z. (2009) DescFold: a web server for protein fold recognition. *BMC Bioinformatics*, **10**, 416.
- Zhu,F., Han,L., Zheng,C., Xie,B., Tammi,M.T., Yang,S., Wei,Y. and Chen,Y. (2009) What are next generation innovative therapeutic targets? Clues from genetic, structural, physicochemical, and systems profiles of successful targets. *J. Pharmacol. Exp. Ther.*, **330**, 304–315.
- Bock,J.R. and Gough,D.A. (2001) Predicting protein-protein interactions from primary structure. *Bioinformatics*, **17**, 455–460.
- Lo,S.L., Cai,C.Z., Chen,Y.Z. and Chung,M.C. (2005) Effect of training datasets on support vector machine prediction of protein-protein interactions. *Proteomics*, **5**, 876–884.
- Qiu,J. and Noble,W.S. (2008) Predicting co-complexed protein pairs from heterogeneous data. *PLoS Comput. Biol.*, **4**, e1000054.
- Yamanishi,Y., Araki,M., Gutteridge,A., Honda,W. and Kanehisa,M. (2008) Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*, **24**, i232–i240.
- Xia,Z., Wu,L.Y., Zhou,X. and Wong,S.T. (2010) Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces. *BMC Syst. Biol.*, **4**(Suppl. 2), S6.
- Barkan,D.T., Hostetter,D.R., Mahrus,S., Pieper,U., Wells,J.A., Craik,C.S. and Sali,A. (2010) Prediction of protease substrates using sequence and structure features. *Bioinformatics*, **26**, 1714–1722.
- Rottig,M., Rausch,C. and Kohlbacher,O. (2010) Combining structure and sequence information allows automated prediction of substrate specificities within enzyme families. *PLoS Comput. Biol.*, **6**, e1000636.
- Wang,L. and Brown,S.J. (2006) BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. *Nucleic Acids Res.*, **34**, W243–W248.
- Terribilini,M., Lee,J.H., Yan,C., Jernigan,R.L., Honavar,V. and Dobbs,D. (2006) Prediction of RNA binding sites in proteins from amino acid sequence. *RNA*, **12**, 1450–1462.
- Liu,Z.P., Wu,L.Y., Wang,Y., Zhang,X.S. and Chen,L. (2010) Prediction of protein-RNA binding sites by a random forest method with combined features. *Bioinformatics*, **26**, 1616–1622.
- Carson,M.B., Langlois,R. and Lu,H. (2010) NAPS: a residue-level nucleic acid-binding prediction server. *Nucleic Acids Res.*, **38**, W431–W435.
- Murakami,Y., Spriggs,R.V., Nakamura,H. and Jones,S. (2010) PiRaNha: a server for the computational prediction of RNA-binding residues in protein sequences. *Nucleic Acids Res.*, **38**, W412–W416.
- Chen,C.T., Yang,E.W., Hsu,H.J., Sun,Y.K., Hsu,W.L. and Yang,A.S. (2008) Protease substrate site predictors derived from machine learning on multilevel substrate phage display data. *Bioinformatics*, **24**, 2691–2697.
- Rastogi,S. and Rost,B. (2010) Bioinformatics predictions of localization and targeting. *Methods Mol. Biol.*, **619**, 285–305.
- Overton,I.M., Padovani,G., Girolami,M.A. and Barton,G.J. (2008) ParCrys: a Parzen window density estimation approach to protein crystallization propensity prediction. *Bioinformatics*, **24**, 901–907.
- Kurgan,L., Razib,A.A., Aghakhani,S., Dick,S., Mizianty,M. and Jahandideh,S. (2009) CRYSTALP2: sequence-based protein crystallization propensity prediction. *BMC Struct. Biol.*, **9**, 50.
- Kandaswamy,K.K., Pugalenti,G., Suganthan,P.N. and Gangal,R. (2010) SVMCRY: an SVM approach for the prediction of protein crystallization propensity from protein sequence. *Protein Pept. Lett.*, **17**, 423–430.
- 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.
- Cui,J., Han,L.Y., Lin,H.H., Zhang,H.L., Tang,Z.Q., Zheng,C.J., Cao,Z.W. and Chen,Y.Z. (2007) Prediction of MHC-binding peptides of flexible lengths from sequence-derived structural and physicochemical properties. *Mol. Immunol.*, **44**, 866–877.
- Fjell,C.D., Jenssen,H., Hilpert,K., Cheung,W.A., Pante,N., Hancock,R.E. and Cherkasov,A. (2009) Identification of novel antibacterial peptides by cheminformatics and machine learning. *J. Med. Chem.*, **52**, 2006–2015.
- Khatun,J., Hamlett,E. and Giddings,M.C. (2008) Incorporating sequence information into the scoring function: a hidden Markov model for improved peptide identification. *Bioinformatics*, **24**, 674–681.
- Shah,A.R., Agarwal,K., Baker,E.S., Singhal,M., Mayampurath,A.M., Ibrahim,Y.M., Kangas,L.J., Monroe,M.E., Zhao,R., Belov,M.E. et al. (2010) Machine learning based prediction for peptide drift times in ion mobility spectrometry. *Bioinformatics*, **26**, 1601–1607.
- Jacob,L. and Vert,J.P. (2008) Efficient peptide-MHC-I binding prediction for alleles with few known binders. *Bioinformatics*, **24**, 358–366.
- Li,Z.R., Lin,H.H., Han,L.Y., Jiang,L., Chen,X. and Chen,Y.Z. (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.
32. 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.
 33. Ren, B. (2003) Atomic-level-based AI topological descriptors for structure-property correlations. *J. Chem. Inf. Comput. Sci.*, **43**, 161–169.
 34. Fernandez, L., Caballero, J., Abreu, J.I. and Fernandez, M. (2007) Amino acid sequence autocorrelation vectors and Bayesian-regularized genetic neural networks for modeling protein conformational stability: gene V protein mutants. *Proteins*, **67**, 834–852.
 35. Niwa, T. (2006) Elucidation of characteristic structural features of ligand binding sites of protein kinases: a neural network approach. *J. Chem. Inf. Model*, **46**, 2158–2166.
 36. Niu, B., Jin, Y., Lu, L., Fen, K., Gu, L., He, Z., Lu, W., Li, Y. and Cai, Y. (2009) Prediction of interaction between small molecule and enzyme using AdaBoost. *Mol. Divers*, **13**, 313–320.
 37. Todeschini, R. and Consonni, V. (2000) *Handbook of Molecular Descriptors*. Wiley-VCH, Weinheim.
 38. Gromiha, M.M. and Suwa, M. (2006) Influence of amino acid properties for discriminating outer membrane proteins at better accuracy. *Biochim. Biophys. Acta*, **1764**, 1493–1497.
 39. Huang, L.T. and Gromiha, M.M. (2008) Analysis and prediction of protein folding rates using quadratic response surface models. *J. Comput. Chem.*, **29**, 1675–1683.
 40. Gromiha, M.M. (2003) Importance of native-state topology for determining the folding rate of two-state proteins. *J. Chem. Inf. Comput. Sci.*, **43**, 1481–1485.