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

H. B. Rao¹, F. Zhu¹,², G. B. Yang³, Z. R. Li¹,⁴,* and Y. Z. Chen²,⁴

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.c3s.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 $H_{ij}^1(i)$, hydrophilicity values $H_{ij}^2(i)$ and side chain masses $M^p(i)$ of 20 amino acids ($i = 1, 2, \ldots, 20$) (32):

$$H_{ij}^1(i) = \frac{H_{ij}^1(i) - \frac{20}{20} \sum_{i=1}^{20} H_{ij}^1(i)}{\sqrt{\frac{\sum_{i=1}^{20} (H_{ij}^1(i) - \frac{20}{20} \sum_{i=1}^{20} H_{ij}^1(i))^2}{20}}}$$

$$H_{ij}^2(i) = \frac{H_{ij}^2(i) - \frac{20}{20} \sum_{i=1}^{20} H_{ij}^2(i)}{\sqrt{\frac{\sum_{i=1}^{20} (H_{ij}^2(i) - \frac{20}{20} \sum_{i=1}^{20} H_{ij}^2(i))^2}{20}}}$$

$$M^p(i) = \frac{M^p(i) - \frac{20}{20} \sum_{i=1}^{20} M^p(i)}{\sqrt{\frac{\sum_{i=1}^{20} (M^p(i) - \frac{20}{20} \sum_{i=1}^{20} M^p(i))^2}{20}}}$$

Then, a correlation function can be computed as:

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

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

$$\theta_1 = \frac{1}{N - 1} \sum_{i=1}^{N-1} \Theta(R_i, R_{i+1})$$

$$\theta_2 = \frac{1}{N - 2} \sum_{i=1}^{N-2} \Theta(R_i, R_{i+2})$$

$$\theta_3 = \frac{1}{N - 3} \sum_{i=1}^{N-3} \Theta(R_i, R_{i+3})$$

$$\theta_\lambda = \frac{1}{N - \lambda} \sum_{i=1}^{N-\lambda} \Theta(R_i, R_{i+\lambda}), (\lambda < N)$$

$\lambda (< N)$ 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:

$$X_{u} = \frac{f_u}{\sum_{i=1}^{20} f_i + \sum_{j=1}^{\lambda} \theta_j}$$

$$X_{u} = \frac{w \theta_{20} - \sum_{i=1}^{20+\lambda} f_i + \sum_{j=1}^{\lambda} \theta_j}$$

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:

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

from which sequence order factors can be defined as:

$$\tau_1 = \frac{1}{N - 1} \sum_{i=1}^{N-1} H_{i,i+1}^1, \ \tau_2 = \frac{1}{N - 1} \sum_{i=1}^{N-2} H_{i,i+1}^2, \$$

$$\tau_3 = \frac{1}{N - 2} \sum_{i=1}^{N-2} H_{i,i+2}, \ \tau_4 = \frac{1}{N - 2} \sum_{i=1}^{N-2} H_{i,i+2}^2, \ldots ,$$

$$\tau_{2\lambda-1} = \frac{1}{N - \lambda} \sum_{i=1}^{N-\lambda} H_{i,i+\lambda}, \ \tau_{2\lambda} = \frac{1}{N - \lambda} \sum_{i=1}^{N-\lambda} H_{i,i+\lambda}^2 (\lambda < N)$$

(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

Submit  Reset

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

<table>
<thead>
<tr>
<th>Feature group</th>
<th>Features</th>
<th>No. of descriptors</th>
<th>No. of descriptor values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Composition-1</td>
<td>Amino acid composition</td>
<td>1</td>
<td>20</td>
</tr>
<tr>
<td>Composition-2</td>
<td>Dipeptide composition</td>
<td>1</td>
<td>400</td>
</tr>
<tr>
<td>Autocorrelation 1</td>
<td>Normalized Moreau–Broto autocorrelation</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>Autocorrelation 2</td>
<td>Moran autocorrelation</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>Autocorrelation 3</td>
<td>Geary autocorrelation</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>Composition, Transition, Distribution</td>
<td>Sequence order coupling number</td>
<td>2</td>
<td>90</td>
</tr>
<tr>
<td>Quasi-sequence order descriptors</td>
<td>Sequence order coupling number</td>
<td>2</td>
<td>150</td>
</tr>
<tr>
<td>PAAC</td>
<td>PAAC</td>
<td>b</td>
<td>a</td>
</tr>
<tr>
<td>APAAC</td>
<td>APAAC</td>
<td>c</td>
<td>b</td>
</tr>
<tr>
<td>Topological descriptors</td>
<td>Topological descriptors</td>
<td>d</td>
<td>d</td>
</tr>
<tr>
<td>TAAPs</td>
<td>TAP</td>
<td>d</td>
<td>d</td>
</tr>
</tbody>
</table>

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

The number depends on the choice of the number of the set of amino acid properties and the choice of the

The number depends on the choice of the number of properties of amino acid.

and APAAC are defined as:

\[
p_u = \sum_{i=1}^{20} f_i w \sum_{j=1}^{2} \tau_j  \quad \text{when } 1 \leq u \leq 20
\]

\[
p_u = \sum_{i=1}^{20} f_i w \sum_{j=1}^{2} \tau_j  \quad \text{when } 20+1 \leq u \leq 20+\lambda
\]

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: \( P_{\text{tot}(i)} = \sum_{j=1}^{N} P_{\text{norm}(i)}/N \), where \( P_{\text{norm}(i)} \) represents the property \( i \) of amino acid \( R_j \) that is normalized between 0 and 1 using the following expression, \( P_{\text{norm}(i)} = (P_{ij} - P_{\text{min}ij})/(P_{\text{max}ij} - P_{\text{min}ij}) \), where \( P_{ij} \) is the original amino acid property \( i \) for residue \( j \) and \( P_{\text{max}ij} \) and \( P_{\text{min}ij} \) 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 \( V_a = \{V_{ai}(i), i = 1, 2, \ldots, n\} \) and \( V_b = \{V_{bj}(i), i = 1, 2, \ldots, n\} \) 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 \( V_{ab} = (V_{ai}, V_{bj}) \) for interaction between proteins \( A \) and \( B \) and \( V_{ba} = (V_{bi}, V_{ai}) \) for interaction between proteins \( B \) and \( A \) (8,9). In the second method, one vector \( V \) with dimension of \( 2n \) is constructed: \( V = \{V_{ai}(i) + V_{bj}(i), V_{ai}(i) \times V_{bj}(i), i = 1, 2, \ldots, n\} \) 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: \( V = \{V_{ai}(i) \times V_{bj}(j), i = 1, 2, \ldots, n, j = 1, 2, \ldots, n, k = (i - 1) \times n + j\} \).

**Protein–ligand interaction descriptors**

Protein–ligand interaction descriptor vector \( V \) can be constructed from the protein descriptor vector \( V_p = (V_{pi}(i), i = 1 \ldots, n_p) \) and ligand descriptor vector \( V_l = (V_{li}(i), i = 1 \ldots, n_l) \) 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 \( n_p \times n_l \) is constructed \( V = (V_{pi}(i) \times V_{li}(j), i = 1, 2, \ldots, n_p, j = 1, 2, \ldots, n_l, k = (i - 1) \times n_p + j) \).

**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.

REFERENCES


