PROFEAT 2016

User Guide (Parameter)

Table of Contents

Introduction to Input Parameters in PROFEAT

This document contains the parameters controlling the calculation in PROFEAT. It contains 12 sections and each section is composed of a notice line which does not influence the calculation and one or more parameters lines. A detailed description is given as follows.

Section 1: parameter controls the purpose of calculation (ipp)

A parameter controls the purpose of the calculation.

- ipp=1: Calculation is for protein sequences. Input file "input-protein.dat" is required.
- ipp=2: Calculation is for ligands. Input file "input-ligand.dat" is required.
- ipp=3: Calculation is for protein-protein interaction. Input files "input-protein.dat"and "input-ppi.dat" are required.
- ipp=4: Calculation is for protein-ligand interaction. Input files "input-protein.dat", "input-ligand.dat", "input-pli.dat" are required.

Section 2: parameter for amino acid composition (iaac)

A parameter controls if calculating "Amino acid composition (AAC)" descriptors for a protein sequence or a peptide sequence.

- iaac=0: do not calculate AAC descriptors .
- iaac=1: calculate AAC descriptors for the whole proteins sequence.
- iaac>1: the protein sequence is divided equally into iaac segments and AAC descriptors are calculated for each segment.

Section 3: parameter for dipeptide composition (idpc)

A parameter controls if calculating "dipeptide composition (DPC)" descriptors for a protein sequence or a peptide sequence.

- idpc=0: do not calculate DPC descriptors.
- idpc=1: calculate DPC descriptors for the whole proteins sequence.
- idpc>1: the protein sequence is divided equally into idpc segments and DPC descriptors are calculated for each segment.

Section 4: parameter for autocorrelation descriptors (imb, imoran, igeary)

Three lines of parameters are required to control the calculation of autocorrelation descriptor for protein or peptide sequence. The three lines are:

iatd, nlag

imb, imoran, igeary

nid, (iad(i), i=1, nid)

Where iatd=1 or 0: do or not do calculation for these autocorrelation descriptors (ATS) respectively.

nlag: the maximum lag of the autocorrelation descriptors (recommended values: 25-30).

- imb=0: do not calculate Moreau-Broto ATS descriptors.
- imb=1: calculate Moreau-Broto ATS descriptors for the whole proteins sequence.
- imb>1: the protein sequence is divided equally into imb segments and Moreau-Broto ATS descriptors are calculated for each segment.\

- imoran=0: do not calculate Moran ATS descriptors.
- imoran=1: calculate Moran ATS descriptors for the whole proteins sequence.
- imoran>1: the protein sequence is divided equally into imoran segments and Moran ATS descriptors are calculated for each segment.

- igeary=0: do not calculate Geary ATS descriptors.
- igeary=1: calculate Geary ATS descriptors for the whole proteins sequence.
- igeary>1: the protein sequence is divided equally into igeary segments and Geary ATS descriptors are calculated for each segment.

nid: number of amino acid indexes will be read and used as physicochemical properties of the amino acid in the ATS descriptors.

(iad(i),i=1,nid): the corresponding nid serial numbers in the input file "input-aaindexdb.dat" for amino acid index database.

Section 5: parameter for composition, transition, distribution (ictd)

A parameter controls if calculating "composition, transition, distribution (CTD)" descriptors for a protein sequence or a peptide sequence.

- ictd=0: do not calculate CTD descriptors.
- ictd=1: calculate CTD descriptors for the whole proteins sequence.
- ictd>1: the protein sequence is divided equally into ictd segments and CTD descriptors are calculated for each segment.

Section 6: parameter for quasi-sequence-order descriptors (iqso)

A parameter controls if calculating "Quasi-sequence-order descriptors (QSO)" descriptors f for a protein sequence or a peptide sequence.

- iqso=0: do not calculate QSO descriptors.
- iqso=1: calculate QSO descriptors for the whole proteins sequence.
- iqso>1: the protein sequence is divided equally into iqso segments and QSO descriptors are calculated for each segment.

Section 7: parameter for pseudo-amino acid composition (ipaac)

Parameters for "pseudo-amino acid composition (PAAC)" descriptors for a protein sequence or a peptide sequence.

ipaac

w, lamda

nset

nst

nacp(1), (idacp(1,j),j=1,nacp(1))

nacp(2), (idacp(2,j),j=1,nacp(2))

…

nacp(nst), (idacp(nst,j),j=1,nacp(nst))

where

- ipaac=0: do not calculate PAAC descriptors.
- ipaac=1: calculate PAAC descriptors for the whole proteins sequence.
- ipaac>1: the protein sequence is divided equally into ipaac segments and PAAC descriptors are calculated for each segment.

w: weighting factor, lamda: maximum number of the tier correlation factor(< length of protein or length of the sequence segment).

nst: choice for the amino acid properties for the correlation function. If nst=0, default amino acid properties are used for correlation function in PAAC descriptor calculation. These default amino acid properties are the 3 amino acid properties with [AAindex] 115, 485, 486 in input-aaindexdb.dat. if nst>0, nst set of amino acid properties are used for correlation function in PAAC descriptor calculation and additional nst input lines for specification of amino acid properties are needed.

nacp(k): number of set of amino acid properties that will be used to get the averaged correlation and (idacp(1,j),j=1,nacp(1)) are the corresponding amino acid index serial number in input-aaindexdb.dat

Section 8: parameter for amphiphilic pseudo-amino acid composition (iapaac)

A parameter controls if calculating "Amphiphilic pseudo-amino acid composition (APAAC)" descriptors for a protein sequence or a peptide sequence.

- iapaac=0: do not calculate APAAC descriptors.
- iapaac=1: calculate APAAC descriptors for the whole proteins sequence.
- iapaac>1: the protein sequence is divided equally into iapaac segments and APAAC descriptors are calculated for each segment.

Note: the weighting factor W, and the maximum number of the tier correlation factor lamda are set same values as in PAAC descriptors.

Section 9: parameter for topological descriptors (itop, ibcut)

itop is the parameter controlling if calculating "topological (TOP)" descriptors for protein or peptide sequence.

- itop=0: do not calculate TOP descriptors.
- itop=1: calculate TOP descriptors for the whole proteins sequence.
- itop>1: the protein sequence is divided equally into itop segments and TOP descriptors are calculated for each segment.
- ibcut= 1 or 0 (useful only when itop is not equal 0): do or not do BCUT descriptor respectively.

Note: the topological descriptors are calculated at atomic level. It is recommended not to calculate topological descriptors for a whole protein sequence because the number of atoms will be too large for most common protein sequence and hence the computation will be time-consuming (i.e., itop=1 is not recommended).

Moreover, among the TOP descriptors, BCUT descriptors are eigenvalue-derived descriptors and they are time-consuming and hence it is not recommended to calculate BCUT descriptors for a large peptides or a whole protein.

Section 10: parameter for total amino acid properties (iaap, naap)

Two lines of parameters are required to control here and they are:

iaap

naap, (iappn(i), i=1,naap)

Where iapp is the parameter to control if calculating "total amino acid properties (AAP)" descriptors for a protein sequence or a peptide sequence.

- iapp=0: do not calculate AAP descriptors.
- iapp=1: calculate AAP descriptors for the whole proteins sequence.
- iapp>1: the protein sequence is divided equally into iapp segments and AAP descriptors are calculated for each segment.

naap is the number of amino acid properties to be used and (iappn(i),i=1,naap) are the corresponding serial numbers of the naap amin acid indexes in input file "input-aaindexdb.dat".

Section 11: parameter for protein-protein interaction descriptors

Methodpp is the parameter choice of the method in construction of descriptor vector V for protein-protein interaction from two protein descriptor vectors Va (Va(i), i=1,n) and Vb (Vb(i), i=1,n). It is useful only when ipp=3.

- Methodpp=1: two vectors Vab and Vba with dimension of 2n are constructed: Vab=(Va,Vb) for interaction between protein A and protein B and Vba=(Vb,Va) for interaction between protein B and protein A.
- Methodpp=2: one vector V with dimension of 2n is constructed: V={Va(i)+Vb(i), Va(i) ×Vb(i), i=1,2 … n}.
- Methodpp=3: one vector V with dimension of n2 is constructed: V={Va(i) ×Vb(j), i=1,2, …, n, j=1,2… n}.

Section 12: parameter for protein-ligand interaction descriptors

Methodpl is the parameter choice of the method in construction of descriptor vector **V** for protein-ligand interaction from the protein descriptor vector $\mathbf{V_p}$ ($V_P(i)$, i=1,n_P) and ligand descriptor $\mathbf{V_L}$ ($V_l(i)$, i=1,n_L). It is useful only when ipp=4.

- Methodpl=1: one vector V with dimension of nP+nL are constructed: V=(VP,VL) for interaction between Protein P and ligand L.
- Methodpl=2: One vector V with dimension of nP × nL is constructed by the tensor product: V={V(k)=Vp(i) ×Vl(j), i=1,2, …, nP, j=1,2,…, nL, k=(i-1) × nP +j }.

Amino Acids Indexes for 20 Natural Amino Acids

This file is necessary for all calculations and contains the amino acids indexes for the 20 natural amino acids. The first 484 amino acid indexes are taken and processed from the amino acid index database AAINDEX (http://www.genome.jp/aaindex/). The format for one amino acid index is as follows:

Table 1 Example of amino acid index in input-aaindexdb.dat

```
[AAINDEX]    1
H ANDN920101
D alpha-CH chemical shifts (Andersen et al., 1992)
R LIT:1810048b PMID:1575719
A Andersen, N.H., Cao, B. and Chen, C.
T Peptide/protein structure analysis using the chemical shift index method:
upfield alpha-CH values reveal dynamic helices and aL sites
J Biochem. and Biophys. Res. Comm. 184, 1008-1014 (1992)
C BUNA790102    0.949
I   A/L    R/K    N/M    D/F    C/P    Q/S    E/T    G/W    H/Y    I/V
   4.35   4.38   4.75   4.76   4.65   4.37   4.29   3.97   4.63   3.95
   4.17   4.36   4.52   4.66   4.44   4.50   4.35   4.70   4.60   3.95
//
```

For each amino acid index, the first line is "[AAINDEX]" followed by a serial number and the last line begins with "//". Between these two lines, the values in the two lines followed the line that begins with "I" are the properties of the natural 20 amino acids in order "ARNDCQEGHILKMFPSTWYV" in free format. In the above table, these lines are in bold characters and they are necessary for each amino acid index. The rest lines are only for description of the source of the data and must not begin with "I" and may be omitted in a simplified format as in the following table.

Table 2 Simplified example of amino acid index in "input-aaindexdb.dat"

```
[AAINDEX]    1
I
   4.35   4.38   4.75   4.76   4.65   4.37   4.29   3.97   4.63   3.95
   4.17   4.36   4.52   4.66   4.44   4.50   4.35   4.70   4.60   3.95
//
```

In file "input-aaindexdb.dat", the first 484 amino acid indexes are from the amino acid index database AAINDEX. The user can add amino acid index to end of the file "input-aaindexdb.dat", however the serial number must be greater than the maximum serial number of the previous amino acid indexes (i.e., the serial number of the last amino acid index) in "input-aaindexdb.dat". For example:

Table 3 Example of amino acid index provided by users

```
[AAINDEX] 485
T     Hydrophobicity index xh10 from:
T     http://www.sjtu.edu.cn/bioinf/PseAAC/PseAAreadme.htm
I
 0.62   -2.53   -0.78   -0.90   0.29   -0.85   -0.74   0.48   -0.40  1.38
 1.06   -1.50    0.64    1.19   0.12   -0.18   -0.05   0.81   0.26   1.08
//
```

The allowed maximum number of amino acid index in the updated version of PROFEAT is 1000.

Reference: *Kawashima, S., Kanehisa, M. "AAINDEX: amino acid index database", Nucleic Acids Research, 2000, 28, 374.*