USER GUIDE FOR RAAMLAB

1 What is this?

RaaMLab is a free and open-source MATLAB toolbox that is to generate amino acid groupings using amino acids' properties and classification methods, and further extract the structural and physicochemical features of reduced amino acids. It offers four kinds of databases on physico-chemical properties of amino acids and amino acid groupings, 49 amino acid classification methods, and five kinds of biophyscio-chemical features of reduced amino acids including content-based features, correlation-based features, order-based features, position-based features and pseudo-reduced amino acid compositions, which can be easily computed based on the amino acid groupings with the user-defined alphabet size and amino acids' properties.

This document is intended to provide an overview of how one can use the RaaMLab. It's not comprehensive and it's not a manual. If you find mistakes, or have suggestions for improvements, please either fix them yourselves in the source document or send them to the mailing list: daiailiu04@yahoo.com

2 Install RaaMLab

RaaMLab has been successfully tested on Linux and Windows systems. The author could download the RaaMLab from http://bioinfo.zstu.edu.cn/RaaMLab.htm. The RAAM Toolbox requires at least MATLAB Release 7. No compilation is required and the use of MATLAB as a basis allows the toolbox to be used on Windows, Linux, Unix and MAC OS machines. The install process of RaaMLab is very easy:

On Windows:

(1): download the RaaMLab (.zip)

(2): extract or uncompress the .zip file

(3): run startup.m

On Linux:

(1): download the propy package (.tar.gz)

(2): tar -zxf propy-1.0.tar.gz

(3): ./matlab

(4): run startup.m

3 Read protein datasets

You can get a protein datasets protein sequence from the file with sequences that must be in FASTA format:

>ref1

seq1

...

>refn

Seqn

Example: clfa.txt

>>[maxLen,nSeq,Seq,SeqName,SeqLength]=readfasta('testseq.txt');

Reading testseq.txt

P0C6L1

>>maxLen

329

>>Seq

MSAAILLAPGDVIKRSSEELKQRQIQINLVDWMESEGGKEDKTEPKEESKAEGSKDGEGT
QSESGQKEEGGKETKDADVDRRIHTAVGSGSGTKGSGERANENANRGDGKVGGGGGDA
DAGVGATGTNGGRWVVLTEEIARAIESKYGTKIDVYRDDVPAQIIEVERSLQKELGISREG
VAEQTERLRDLRRKEKNGTHAKAVERGGRKQRKKAHGDAQREGVEEEKTSEEPARIGITI
EGVMSQKKLLSMIGGVERKMAPIGARESAVMLVSNSIKDVVRATAYFTAPTGDPHWKEV
AREASKKKNILAYTSTGGDVKTEFLHLIDHL

a) Show the databases in RaaMLab

RaaMLab consists of four databases. The first database is the amino acid index of 20 numerical values of their physicochemical and biochemical properties. The second database is for the amino acid mutation matrix describing the physicochemical and biochemical properties of amino acids' pairs. The third database represents the statistical protein contact potentials. The final database contains amino acid groupings that have been published in the published literatures.

You can obtain all the official information of the databases can be accessed from MATLAB by using ShowIndex command on the command line. Just like

>>ShowIndex(database, index)

```
% database -> kinds of the databases in the RaaMlab Choices are:
%           '1'   - AAindex1 for amino acid index of 20 numerical values
%           '2'   - AAindex2 for the amino acid mutation matrix
%           '3'   - AAindex3 for the statistical protein contact potential
%           '4'   - Published amino acid groupings
% index       -> the index of databases in RaaMlab toolbox
```

Example:

>>ShowIndex(1, 'BHAR880101')

```
//
H BHAR880101
D Average flexibility indices (Bhaskaran-Ponnuswamy, 1988)
R LIT:1414112
A Bhaskaran, R. and Ponnuswamy, P.K.
T Positional flexibilities of amino acid residues in globular proteins
J Int. J. Peptide Protein Res. 32, 241-255 (1988)
C VINM940103       0.869    KARP850102       0.806    WERD780101      -0.803
   RICJ880111      -0.813
I       A/L         R/K         N/M         D/F         C/P         Q/S         E/T         G/W         H/Y
I/V
        0.357     0.529     0.463     0.511     0.346     0.493     0.497     0.544     0.323     0.462
        0.365     0.466     0.295     0.314     0.509     0.507     0.444     0.305     0.420     0.386
//
```

If you don't know the index of databases, you can use the ShowIndex(database) to browse all the index in this database

Example:

>>ShowIndex(4)

No.0 AAindex of the database 4 is H AlphabetID=1

No.1 AAindex of the database 4 is H AlphabetID=2

No.2 AAindex of the database 4 is H AlphabetID=5

No.3 AAindex of the database 4 is H AlphabetID=6

No.4 AAindex of the database 4 is H AlphabetID=25

No.5 AAindex of the database 4 is H AlphabetID=4

No.6 AAindex of the database 4 is H AlphabetID=16

No.7 AAindex of the database 4 is H AlphabetID=20

No.8 AAindex of the database 4 is H AlphabetID=21

No.9 AAindex of the database 4 is H AlphabetID=22

No.10 AAindex of the database 4 is H AlphabetID=23a

No.11 AAindex of the database 4 is H AlphabetID=23b

No.12 AAindex of the database 4 is H AlphabetID=24

No.13 AAindex of the database 4 is H AlphabetID=26

No.14 AAindex of the database 4 is H AlphabetID=28

No.15 AAindex of the database 4 is H AlphabetID=7

No.16 AAindex of the database 4 is H AlphabetID=11

No.17 AAindex of the database 4 is H AlphabetID=12

No.18 AAindex of the database 4 is H AlphabetID=14a

No.19 AAindex of the database 4 is H AlphabetID=14b

No.20 AAindex of the database 4 is H AlphabetID=27

No.21 AAindex of the database 4 is H AlphabetID=8

No.22 AAindex of the database 4 is H AlphabetID=9

No.23 AAindex of the database 4 is H AlphabetID=10

No.24 AAindex of the database 4 is H AlphabetID=13

No.25 AAindex of the database 4 is H AlphabetID=18

No.26 AAindex of the database 4 is H AlphabetID=19

No.27 AAindex of the database 4 is H AlphabetID=3

No.28 AAindex of the database 4 is H AlphabetID=15a

No.29 AAindex of the database 4 is H AlphabetID=15b

No.30 AAindex of the database 4 is H AlphabetID=17

No.31 AAindex of the database 4 is H AlphabetID=29

No.32 AAindex of the database 4 is H AlphabetID=30

No.33 AAindex of the database 4 is H AlphabetID=31

4 Reduced amino acids using the methods in RaaMLab

RaaMLab provides you with 49 classification methods to reduce amino acids into specific groups with size R according to the user's setting. The classification methods consist of 7 distance measures (euclidean, seuclidean, cityblock, mahalanobis, minkowski, hamming and jaccard) and 7 cluster methods (single, complete, average, weighted, centroid, median and ward);

>>[G] = AAreduce(Index,para_dis,para_meth,maxgroup)

```
% Index-> the index of databases in RaaMlab toolbox
%    para_dis-> computes the distance of amino acids based on the index of databases
%
%            'euclidean'      - Euclidean distance
%            'seuclidean'     - Standardized Euclidean distance, each coordinate
%                                    in the sum of squares is inverse weighted by the
%                                    sample variance of that coordinate
%            'cityblock'      - City Block distance
%            'mahalanobis' - Mahalanobis distance
%            'minkowski'      - Minkowski distance with exponent 2
%            'hamming'         - Hamming distance, percentage of coordinates
%                                    that differ
%            'jaccard'         - One minus the Jaccard coefficient, the
%                                    percentage of nonzero coordinates that differ
% para_meth-> creates a hierarchical cluster tree, using the single
%      linkage algorithm of Matlab
%           'single'        --- nearest distance
%           'complete'    --- furthest distance
%           'average'      --- unweighted average distance (UPGMA) (also known as
%                                    group average)
%           'weighted'    --- weighted average distance (WPGMA)
%           'centroid'    --- unweighted center of mass distance (UPGMC) (*)
%           'median'       --- weighted center of mass distance (WPGMC) (*)
%           'ward'          --- inner squared distance (min variance algorithm) (*)
% maxgroup is the size of reduced amino acid set. Maxgroup can be from 1 to 20.
```

Example:

```
>>  reduce=AAreduce('BROC820101','euclidean','single',8)

    Columns 1 through 20

    4     5     1     5     6     5     2     5     5     4     8     5     4

    8     4     5     5     7     4     3
```

If you want to use published amino acid groupings in the database, in which we have collected 34 kinds of published amino acid groupings. You can also used this function

```
>>  reduce=AAreduce('index')
```

```
% Index-> the index of published amino acids in the database of RaaMLab
```

Example:

```
>>  reduce=AAreduce('AlphabetID=30');
  Columns 1 through 20
  1     1     1     1     4     1     1     2     4     4     4     1     4     4
  3     1     1     4     4     4
```

If you want to select no reduced method and then skip this step. You can also used this function

>>reduce=AAreduce('FullAlphabet')

Example:

```
>>  reduce=AAreduce('FullAlphabet')
  Columns 1 through 20
  1     2     3     4     5     6     7     8     9    10    11    12    13
  14    15    16    17    18    19    20
```

5 Calculating the reduced amino acids' features

RaaMLab can compute a large number of structural and physicochemical features of the reduced amino acids. It summarizes four kinds of the features that have been widely used in predicting protein- and peptide-related problems.

5.1 Content-based features of reduced amino acids

Content-based features not only consist of reduced amino acid composition, transition and distribution, but also contain two types of pseudo-reduced amino acid compositions (PRseAAC): type I PRseAAC and type II PRseAAC.

1) K-mer

```
>>[SeqAAC,SeqDAAC,SeqTAAC,SeqFuAAC,SeqFiAAC] = RAAC(Seq,reduce)
   % Calculate the content distribution of the reduced amino acids
   % It contains reduced amino acid composition
   %           direduced-peptide composition,
   %           trireduced-peptide,
   %           tetrareduced-peptide composition
   %           pentareduced-peptide compositionthe databases
   % Seq -> the loaded protein sequence
   % reduce-> the reduced amino acids
```

Example:

>>[maxLen,nSeq,Seq,SeqName,SeqLength]=readfasta('testseq.txt');

>>reduce=AAreduce('BROC820101','euclidean','single',8);

>> [aac,daac,taac,fuaac,fiaac]=RAAC(Seq,reduce)

>>aac

0.0243 0.1185 0.0669 0.1976 0.5289 0 0.0091 0.0547

>>dacc

Columns 1 through 64

0 0.0030 0 0.0061 0.0122 0 0 0.0030

0.0030 0.0213 0.0061 0.0122 0.0671 0 0 0.0091

0 0.0122 0.0061 0.0213 0.0244 0 0 0.0030 0.0091

0.0183 0.0091 0.0457 0.1037 0 0 0.0122 0.0122

0.0640 0.0366 0.0976 0.2927 0 0.0091 0.0183 0

0 0 0 0 0 0 0 0

0 0.0030 0.0030 0.0030 0 0 0 0

0 0.0061 0.0091 0.0274 0 0 0.0091

…..

2) Composition, transition and distribution

>> [Composition,Transition,Distribution] = RCTD(Seq,reduce)

```
% Calculate the composition,transition,distribution reduced amino acids
%          composition: reduced amino acid composition
%          transition: A transition from a reduced amino acid  to another
%          is the percent frequency with which the reduced amino acid
is %followed by the reduced amino acid  or A is followed by I in the reduced %
sequence.
%          distribution: There are five ¡°distribution¡± descriptors for
%          each attribute and they are the position percents in the
whole %sequence for the first residue, 25% residues, 50% residues,
%         75% residues and 100% residues , respectively, for a specified
%         encoded class.
% Seq -> the loaded protein sequence
% reduce-> the reduced amino acids
% RaaMLab: a MATLAB toolbox for generating amino acid group-ings and
RedAA %modes
% Qi Dai, 20 Apri 2014,
```

Example:

>>[maxLen,nSeq,Seq,SeqName,SeqLength]=readfasta('testseq.txt');

>>reduce=AAreduce('BROC820101','euclidean','single',8);

>>[co,tr,di]=RCTD(Seq,reduce);

>>co

0.0243 0.1185 0.0669 0.1976 0.5289 0 0.0091 0.0547

>>tr

0.0061 0 0.0152 0.0244 0 0 0.0030 0.0183 0.0305 0.1311 0 0 0.0091 0.0305 0.0610 0 0.0030 0.0091 0.2012 0 0.0030 0.0213 0 0.0122 0.0457 0 0 0

>>di

0.0851 0.3070 0.3191 0.5957 0.9331 0.0547 0.1763 0.4985 0.6809 0.9757 0.0365 0.3708 0.5015 0.8176 0.9666 0.0030 0.3040 0.5502 0.7964 0.9909 0.0061 0.2432 0.4620 0.7112 0.9970 0 0 0 0 0 0.0973 0.0973 0.4012 0.4012 0.8967 0.0182 0.4103 0.5775 0.9392 1.0000

3) pseudo-reduced amino acid composition (PRseAAC)

>>[Psedo_AAC] = RPseudoAAC1(Index1,Index2,Index3,Seq,reduce,lambda,w)

% Calculate the type I pseudo-reduced amino acids of the protein sequences

% Index1 -> a kind of chosen physico-chemical properties of database1

% Index2 -> a kind of chosen physico-chemical properties of database1

% Index3 -> a kind of chosen physico-chemical properties of database1

% Seq -> the loaded protein sequence

% reduce-> the reduced amino acids

% lambda-> the lag of the RPseudoAAC;

% w-> the weighting factor

Example:

>>[maxLen,nSeq,Seq,SeqName,SeqLength]=readfasta('testseq.txt');

>>reduce=AAreduce('BROC820101','euclidean','single',8);

>>Index1='ARGP820103';

>>Index2='BEGF750102';

>>Index3='BHAR880101';

>>lambda=10;

>>w=0.03;

>>pseudoaac1=RPseudoAAC1(Index1,Index2,Index3,Seq,reduce,lambda,w)

0.0061 0.0299 0.0169 0.0499 0.1335 0 0.0023 0.0138 0.0734 0.0753 0.0736 0.0717 0.0742 0.0777 0.0773 0.0751

0.0752 0.0741

```
>> pseudoaac2=RPseudoAAC2(Index1,Index2,Seq,reduce,lambda,w);
% Calculate the type II pseudo-reduced amino acids of the protein sequences
% Index1 -> a kind of chosen physico-chemical properties of database1
% Index2 -> a kind of chosen physico-chemical properties of database1
% Seq -> the loaded protein sequence
% reduce-> the reduced amino acids
% lambda-> the lag of the RPseudoAAC;
% w-> the weighting factor
```

Example:

```
>>[maxLen,nSeq,Seq,SeqName,SeqLength]=readfasta('testseq.txt');
>>reduce=AAreduce('BROC820101','euclidean','single',8);
>>Index1='ARGP820103';
>>Index2='BEGF750102';
>>lambda=10;
>>w=0.03;
>>pseudoaac2=RPseudoAAC2(Index1,Index2,Index3,Seq,reduce,lambda,w)
```

0.0072 0.0353 0.0199 0.0588 0.1573 0 0.0027 0.0163
0.0305 0.0307 0.0310 0.0354 0.0309 0.0298 0.0292 0.0322
0.0277 0.0307

5.2 Correlation-based features of reduced amino acids

The second RedAA modes are correlation-based features, describing correlation relationships among the distributions of reduced amino acids. There are three different autocorrelation features, normalized Moreau–Broto autocorrelation, Moran autocorrelation and Geary autocorrelation, in the proposed RedAA mode. Each of these features has Z descriptor values, where Z is a given parameter.

```
>> [MBAC1,MBAC2,MBAC3]=RACF('ARGP820103',Seq,reduce,50)
% Compute the correlation-based features of the reduced amino acids
% The output of the function are MBAC1, MBAC2 and MBAC3 which are
% corresponding to the normalized Moreau¨CBroto autocorrelation, Moran
% autocorrelation and Geary autocorrelation, respectively.
% Parameters of the function
% Index -> a kind of chosen physico-chemical properties of database1
```

```
% Seq -> the loaded protein sequence
% reduce-> the reduced amino acids
% d-> the lag of the correlation
```

Example:

```
>>[maxLen,nSeq,Seq,SeqName,SeqLength]=readfasta('testseq.txt');
>>reduce=AAreduce('BROC820101','euclidean','single',8);
>>  [MBAC1,MBAC2,MBAC3]=RACF('ARGP820103',Seq,reduce,10);
MBAC1
4.3348    3.3816    2.3133    1.3540   -0.0926   -1.0424   -2.2279   -2.9674   -4.1426
 -5.1307
MBAC2
 -0.0011   -0.0022   -0.0035   -0.0046   -0.0063   -0.0074   -0.0087   -0.0096   -
 0.0110   -0.0121
MBAC3
6.3266    6.2112    6.3436    6.1355    7.2139    7.0364    7.4299    6.6358
7.0031    6.9807
```

5.3 Order-based features of reduced amino acids

The third RedAA modes are order-based features, reflecting the physico-chemical interaction of each pair reduced amino acids based on user-defined properties. It includes two kinds of order-based features, one is sequence-order-coupling number, and the other is quasi-sequence-order. Schneider–Wrede physicochemical distance matrix and Grantham chemical distance matrix are used to compute them.

```
>> [Order_d,Quasi_order] = RSeqOrder(Seq,reduce,d,w)
 % Compute two kinds of order-based features, one is sequence-order-
 coupling %number,
 % and the other is quasi-sequence-order. They are computed based on the
 % Schneider¨CWrede physicochemical distance matrix and Grantham
 chemical %distance matrix.
 % Order_d are results of sequence-order-coupling number, and
 % Quasi_order shows results of quasi-sequence-order based on
 % Schneider¨CWrede physicochemical distance matrix and Grantham
 chemical %distance matrix.
 % Parameters of function
 % Seq -> the loaded protein sequence
 % reduce-> the reduced amino acids
 % d-> the lag of the correlation
 % w-> the weighting factor
```

Example:

```
>>[maxLen,nSeq,Seq,SeqName,SeqLength]=readfasta('testseq.txt');
>>reduce=AAreduce('BROC820101','euclidean','single',8);
>>  [Order_d,Quasi_order] = RSeqOrder(Seq,reduce,10,0.01)
Order_d =
1.0e+003 *
6.8813    7.0859    7.0148    6.7821    7.4389    7.0820    6.6632    7.1156
6.7700    6.6936    6.8813    7.0859    7.0148    6.7821    7.4389    7.0820
6.6632    7.1156    6.7700    6.6936
Quasi_order =
 0.0000    0.0002    0.0001    0.0003    0.0008         0    0.0000    0.0001
 0.0972    0.0961    0.0505    0.2463    0.1389    0.4104    1.0987         0
 0.0189    0.1137   -0.1804   -0.2016
```

5.4 Position-based features of reduced amino acids

The final RedAA modes are position-based features of reduced amino acids. They describe the dispersion of the probability position distribution of each reduced amino acid along the protein sequences using the coefficient of variation.

```
>>  [AAP] = RAAP(Seq,reduce)
% Compute position-based features of reduced amino acids.
% Parameters of function
% Seq -> the loaded protein sequence
% reduce-> the reduced amino acids
```

Example:

```
>>[maxLen,nSeq,Seq,SeqName,SeqLength]=readfasta('testseq.txt');
>>reduce=AAreduce('BROC820101','euclidean','single',8);
>>  [AAP] = RAAP(Seq,reduce)
AAP
         0    0.4677    0.3162    0.5477    1.1094         0         0    0.4472
```

6 Output all of the features

This function computes all the features of the reduced amino acids and save them in the same directory where the Results file is stored.

>> function [] = integration(filename,Index,para_dis,para_meth,maxgroup)

```
%%
% Calculate the content distribution of the reduced amino acids
% It contains reduced amino acid composition
%           direduced-peptide composition,
%           trireduced-peptide,
%           tetrareduced-peptide composition
%           pentareduced-peptide compositionthe databases
% 'filename' is a name of a multiple fasta file
% Index-> the index of databases in RaaMlab toolbox
 % para_dis-> computes the distance of amino acids based on the index of
  databases
% para_meth-> creates a hierarchical cluster tree
 % maxgroup_> is the size of reduced amino acid set. Maxgroup can be from 1 to
   20.
```

Example:

>> integration('testseq.txt', 'BROC820101','euclidean','single',8)