Package 'PSSMCOOL'

October 12, 2022

Type Package

Title Features Extracted from Position Specific Scoring Matrix (PSSM)

Version 0.2.4

Imports utils, gtools, infotheo, phonTools, dtt

Maintainer Alireza mohammadi <alireza691111@gmail.com>

Depends R (>= 3.1.0)

Description Returns almost all features that has been extracted from Position Specific Scoring Matrix (PSSM) so far, which is a matrix of L rows (L is protein length) and 20 columns produced by 'PSI-BLAST' which is a program to produce PSSM Matrix from multiple sequence alignment of proteins see <https://www.ncbi.nlm.nih.gov/books/NBK2590/> for mor details. some of these features are described in Zahiri, J., et al.(2013) <DOI:10.1016/j.ygeno.2013.05.006>, Saini, H., et al.(2016) <DOI:10.17706/jsw.11.8.756-767>, Ding, S., et al.(2014) <DOI:10.1016/j.biochi.2013.09.013>, Cheng, C.W., et al.(2008) <DOI:10.1186/1471-2105-9-S12-S6>, Juan, E.Y., et al.(2009) <DOI:10.1109/CISIS.2009.194>.

License GPL-3

URL https://github.com/BioCool-Lab/PSSMCOOL

BugReports https://github.com/BioCool-Lab/PSSMCOOL/issues

Encoding UTF-8

RoxygenNote 7.1.2

Suggests testthat, spelling, waveslim, knitr, rmarkdown

Language en-US

VignetteBuilder knitr

NeedsCompilation no

Author Javad zahiri [aut],
Alireza mohammadi [aut, cre],
Saber mohammadi [aut]

Repository CRAN

Date/Publication 2022-01-04 13:20:05 UTC

R topics documented:

PSSMCOOL-package *PSSMCOOL:Extracting Various Feature vectors from PSSM Matrix*

Description

This package is a Comprehensive toolkit for generating various numerical features from PSSM Matrix correspond to each protein sequence and covers all features introduced in POSSUM Package and some other new features. see PSSMCOOL_Rpubs for more information.

Author(s)

Maintainer: Alireza Mohammadi <alireza691111@gmail.com>

Authors:

- Alireza Mohammadi <alireza691111@gmail.com>
- Javad Zahiri <zahiri@modares.ac.ir>
- Saber Mohammadi <sabermohammdi.1994@gmail.com>

See Also

All functions in this package:

- AATP_TPC
- AB_PSSM
- Averag_Block
- consunsus_sequence
- CS_PSe_PSSM
- DFMCA_PSSM
- Discrete_Cosine_Transform
- disulfid
- DP_PSSM
- dpc_pssm
- dwt_PSSM
- EDP_EEDP_MEDP
- FPSSM
- FPSSM2
- grey_pssm_pseAAC
- k_mers
- k_separated_bigrams_pssm

- kiderafactor
- LPC_PSSM
- MBMGACPSSM
- pse_pssm
- pssm400
- pssm_ac
- pssm_cc
- PSSM_SD
- pssm_seg
- PSSMBLOCK
- rpssm
- RPM_PSSM
- scsh2
- single_Average
- smoothed_PSSM
- SOMA_PSSM
- SVD_PSSM
- trigrame_pssm

Useful links:

- https://rpubs.com/alireza_69/640294
- https://github.com/BioCool-Lab/PSSMCOOL/
- Report bugs at https://github.com/BioCool-Lab/PSSMCOOL/issues

---

aac_pssm *AAC-PSSM feature vectors*

Description

AAC-PSSM Feature stands for Amino Acid composition which is actually mean of PSSM Matrix columns which its length is 20. combination of this feature vector and DPC-PSSM feature vector would be AADP-PSSM feature vector.

Usage

```
aac_pssm(pssm_name)
```

Arguments

pssm_name	name of PSSM Matrix file

Value

feature vector of length 20

References

Liu, T., Zheng, X. and Wang, J. (2010) Prediction of protein structural class for low-similarity sequences using support vector machine and PSI-BLAST profile, Biochimie, 92, 1330-1334.

Examples

```
X<-aac_pssm(system.file("extdata", "C7GQS7.txt.pssm", package="PSSMCOOL"))
```

aadp_pssm *DPC-PSSM,AAC-PSSM and AADP-PSSM feature vectors*

Description

This feature is combination of amino asid composition and dipeptide composition feature vectors. DPC feature stands for dipeptide composition, to get this feature vector for both different columns the elements in two consecutive rows corresponds to this columns, would be multiplied and this scenario is done for all L-1 consecutive rows (L is protein length) and finally summation of these numbers divides by L-1. since the result depends on two different columns, eventually DPC Feature vector would be of length 400. AAC-PSSM is actually mean of PSSM Matrix columns which its length is 20. eventually AADP-PSSM is combination of these vectors and with length 420.

Usage

```
aadp_pssm(pssm_name)
```

Arguments

pssm_name	name of PSSM Matrix file

Value

feature vector of length 420

References

Liu, T., Zheng, X. and Wang, J. (2010) Prediction of protein structural class for low-similarity sequences using support vector machine and PSI-BLAST profile, Biochimie, 92, 1330-1334.

Examples

```
X<-aadp_pssm(system.file("extdata", "C7GQS7.txt.pssm", package="PSSMCOOL"))
```

AATP_TPC *AATP_TPC feature vector*

Description

For getting this features which have been used to protein structural class prediction, at first mean of every column in PSSM Matrix is computed to achieve a 20-dimensional vector called AAC.then by combining it with other vector of length 400 called TPC, which is similar to `dpc_pssm` AATP feature vector of length 420 is obtained.

Usage

```
AATP_TPC(pssm_name)
```

Arguments

`pssm_name`	is name of PSSM Matrix file

Value

list of two feature vectors with 400 and 420 dimensions

References

Zhang, S., Ye, F. and Yuan, X. (2012) Using principal component analysis and support vector machine to predict protein structural class for low-similarity sequences via PSSM, Journal of Biomolecular Structure & Dynamics, 29, 634-642.

See Also

`dpc_pssm`

Examples

```
X<-AATP_TPC(paste0(system.file("extdata",package="PSSMCOOL"),"/C7GQS7.txt.pssm"))
```

AB_PSSM *AB-PSSM feature vector*

Description

to get This feature at first, each protein sequence is divided into 20 equal parts, each of which is called a block, and in each block the row vectors of the PSSM matrix related to that block are added together and The resulting final vector is divided by the length of that block, which is 5 Finally, by placing these 20 vectors side by side, feature vector of length 400 is obtained.

Usage

```
AB_PSSM(pssm_name)
```

Arguments

`pssm_name`	name of PSSM Matrix file

Value

AB-PSSM feature vector of length 400

References

Jeong, J.C., Lin, X. and Chen, X.W. (2011) On position-specific scoring matrix for protein function prediction , IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM, 8, 308-315.

Examples

```
X<- AB_PSSM(system.file("extdata","C7GRQ3.txt.pssm",package="PSSMCOOL"))
```

`Averag_Block` *Averag Block feature vector*

Description

In this feature at first PSSM Matrix is divided to 20 Blocks. Then for each Block mean of columns is computed to get 20-dimensional vector, eventually by appending these vectors to each other final feature vector is obtained which would be of length 400. this feature vector is similar to `PSSMBLOCK` for N=20.

Usage

```
Averag_Block(pssm_name)
```

Arguments

`pssm_name`	name of PSSM Matrix file

Value

feature vector of length 400

References

L. Nanni, A. Lumini, and S. J. T. S. W. J. Brahnam, "An empirical study of different approaches for protein classification," vol. 2014, 2014.

Examples

```
v<-Averag_Block(system.file("exttdata", "C7GQS7.txt.pssm", package="PSSMCOOL"))
```

consunsus_sequence *consunsus_sequence*

Description

This feature vector is constructed from PSSM Matrix as: $\alpha(i) = argmax(P_{i,j})$ where i varies between 1 and L and j between 1 and 20, L indicates protein length and "arg" represents the argument of the maximum the ith base of the consensus sequence (CS) is then set to be the $\alpha(i)$th amino acid in the amino acid alphabet and a consensus sequence is constructed.

Usage

```
consunsus_sequence(pssm_name)
```

Arguments

pssm_name	is the name of PSSM Matrix file

Value

consunsus sequence wich extracted from PSSM

References

Y. Liang, S. Liu, S. J. C. Zhang, and m. m. i. medicine, "Prediction of protein structural classes for low-similarity sequences based on consensus sequence and segmented PSSM," vol. 2015, 2015.

Examples

```
X<-consunsus_sequence(system.file("extdata", "C7GQS7.txt.pssm", package="PSSMCOOL"))
```

CS_PSe_PSSM *CSP-SegPseP-SegACP feature vector*

Description

This feature vector is constructed by fusing consensus sequence (CS), segmented PsePSSM, and segmented auto-covariance transformation (ACT) based on PSSM. by consensus sequence a 40-dimensional feature vector is obtained, in segmented PsePSSM group, by dividing PSSM Matrix to 2 and 3 segments a 380-dimensional feature vector is obtained and in ACT group, similar to the previous group at first PSSM Matrix is divided to 2 and 3 segments then a feature vector of length 280 is obtained.eventually by fusing these features a 700-dimensional feature vector is obtained.

Usage

```
CS_PSe_PSSM(pssm_name, vec_name)
```

Arguments

`pssm_name`	name of PSSM Matrix file
`vec_name`	a character that user imports to specify kind of feature vector which it can be varied between four values

Details

If vec_name equals to "segmented_psepssm" then a feature vector of length 380 is obtained. if vec_name equals to "segmented_acpssm" then a feature vector of length 280 is obtained, and if vec_name equals to "cspssm" the obtained feature vector would be of length 40 eventually if vec_name equals to "total" then feature vector would be of length 700.

Value

feature vector that its length depends on the vec_name which user imports. vec_name can be one of "cspssm", "segmented_psepssm", "segmented_acpssm", "total".

References

Y. Liang, S. Liu, S. J. C. Zhang, and m. m. i. medicine, "Prediction of protein structural classes for low-similarity sequences based on consensus sequence and segmented PSSM," vol. 2015, 2015.

Examples

```
X<-CS_PSe_PSSM(system.file("extdata", "C7GSI6.txt.pssm", package="PSSMCOOL"),"total")
```

`DFMCA_PSSM` *DMACA-PSSM feature*

Description

In this feature each column of PSSM Matrix, can be regarded as a time series. Each PSSM contains 20 columns Hence, each PSSM can be considered as 20 time series.The detrended moving-average cross-correlation analysis (DMCA) is developed to measure the level of cross-correlation between two non-stationary time series by fusing the detrended cross-correlation analysis (DCCA) and the detrended moving average(DMA).this function utilizes this algorithm for each column and each pair of columns to produce a feature vector of length 290.

Usage

```
DFMCA_PSSM(pssm_name, n = 7)
```

Arguments

`pssm_name`	name of PSSM Matrix file
`n`	A parameter called the window size that must be smaller than the length of the sequence

Value

feature vector of length 210

Note

parameter n must be equal or greater than 3 and equal or less then L which L is length of protein

References

Y. Liang, S. Zhang, S. J. S. Ding, and Q. i. E. Research, "Accurate prediction of Gram-negative bacterial secreted protein types by fusing multiple statistical features from PSI-BLAST profile," vol. 29, no. 6, pp. 469-481, 2018.

Y. Liang and S. J. A. b. Zhang, "Prediction of apoptosis protein's subcellular localization by fusing two different descriptors based on evolutionary information," vol. 66, no. 1, pp. 61-78, 2018.

Examples

```
X<-DFMCA_PSSM(system.file("extdata", "C7GQS7.txt.pssm", package="PSSMCOOL"),7)
```

`Discrete_Cosine_Transform`

Discrete Cosin Transform Feature

Description

To construct this feature vector, Two-Dimensional DCT algorithm has been used by applying `dct` function from dtt package which DCT stands for Discrete Cosin Transform.

Usage

```
Discrete_Cosine_Transform(pssm_name)
```

Arguments

`pssm_name`	name of PSSM Matrix file

Value

feature vector of length 400

References

Wang, L., et al., Advancing the prediction accuracy of protein-protein interactions by utilizing evolutionary information from position-specific scoring matrix and ensemble classifier. 2017. 418: p. 105-110.

Y. Wang, Y. Ding, F. Guo, L. Wei, and J. J. P. o. Tang, "Improved detection of DNA-binding proteins via compression technology on PSSM information," vol. 12, no. 9, 2017.

Examples

```
X<-Discrete_Cosine_Transform(system.file("extdata", "C7GQS7.txt.pssm", package="PSSMCOOL"))
```

`disulfid` *Disulfide connectivity feature*

Description

This feature is used to predict the disulfide bond within a protein.

Usage

```
disulfid(pssm_name)
```

Arguments

`pssm_name`	name of PSSM Matrix file

Details

For the purpose of predicting disulfide bond in protein at first, the total number of cysteine amino acids in the protein sequence is counted and their position in the protein sequence is identified. Then, using a sliding window with length of 13, moved on the PSSM matrix from top to bottom so that the middle of the window is on the amino acid cysteine, then the rows below the matrix obtained from the PSSM matrix with dimension of 13 x 20 are placed next to each other to get a feature vector with a length of 260 = 20 * 13 per cysteine, and if the position of the first and last cysteine in the protein sequence is such that the middle of sliding window is not on cysteine residue when moving on PSSM Matrix, then the required number of zero rows from top and bottom is added to the PSSM matrix to achieve this goal.Thus, for every cysteine amino-acid presented in protein sequence, a feature vector with a length of 260 is formed.Then all the pairwise combinations of these cysteines is wrote in the first column of a table, and in front of each of these pairwise combinations, the corresponding feature vectors are glued together to get a feature vector of length 520 for each of these compounds.Finally, the table obtained in this way will have the number of rows equal to the number of all pairwise combinations of these cysteines and the number of columns will be equal to 521 (the first column includes the name of these pair combinations). And it is easy to divide this table into training and testing data and predict the desired disulfide bonds between cysteines.

Value

a table with number of all cysteine pairs in rows and 521 columns correspond to feature vector length.

References

D.-J. Yu et al., "Disulfide connectivity prediction based on modelled protein 3D structural information and random forest regression," vol. 12, no. 3, pp. 611-621, 2014.

N. J. Mapes Jr, C. Rodriguez, P. Chowriappa, S. J. C. Dua, and s. b. journal, "Residue adjacency matrix based feature engineering for predicting cysteine reactivity in proteins," vol. 17, pp. 90-100, 2019.

Examples

```
X<-disulfid(system.file("extdata", "C7GQS7.txt.pssm", package="PSSMCOOL"))
```

`dpc_pssm` *DPC_PSSM feature vector*

Description

This Feature stands for dipeptide composition, to get this feature vector for both different columns the elements in two consecutive rows corresponds to this columns, would be multiplied and this scenario is done for all L-1 consecutive rows (L is protein length) and finally summation of these numbers divides by L-1. since the result depends on two different columns, eventually a Feature vector of length 400 will be obtained.

Usage

```
dpc_pssm(pssm_name)
```

Arguments

`pssm_name`	name of PSSM Matrix file

Value

feature vector of length 400

References

Liu, T., Zheng, X. and Wang, J. (2010) Prediction of protein structural class for low-similarity sequences using support vector machine and PSI-BLAST profile, Biochimie, 92, 1330-1334.

Examples

```
X<-dpc_pssm(system.file("extdata", "C7GQS7.txt.pssm", package="PSSMCOOL"))
```

DP_PSSM *DP_PSSM feature vector*

Description

This feature results from the connection of two vectors. The vector is the first feature of a vector with a length of 40, which calculates the average of positive and negative values for each column separately and puts them together. in the second feature vector, correspond to each column the difference between the numbers in the rows that have distance of k is calculated, and then the square average for the differences that are positive is calculated, and the same action for the differences that are negative is performed. since k varies between 1 and α, and because the value of α in this function is equal to 2, the length of the second feature vector will be 80, which by merging with the first feature vector, the total feature vector of length 120 will be obtained.

Usage

```
DP_PSSM(pssm_name, a = 5)
```

Arguments

pssm_name	name of PSSM matrix file
a	fixed parameter that user chooses which usually equals to 2

Value

feature vector of length 240

References

Juan, E.Y., et al. (2009) Predicting Protein Subcellular Localizations for Gram-Negative Bacteria using DP-PSSM and Support Vector Machines. Complex, Intelligent and Software Intensive Systems, 2009. CISIS'09. International Conference on. IEEE, pp. 836-841.

Examples

```
X<-DP_PSSM(system.file("extdata", "C7GQS7.txt.pssm", package="PSSMCOOL"))
```

dwt_PSSM *discrete wavelet transform feature vector*

Description

In construction of this feature vector, the dwt.nondyadic function is used from "waveslim", package to calculate the discrete wavelet transform for each column of the PSSM matrix, which considers it as a discrete signal. At last, 4 levels DWT is used to analysis of these discrete signals of PSSM (each column) and extracted the PSSM-DWT feature from PSSM of protein.

Usage

```
dwt_PSSM(pssm_name)
```

Arguments

pssm_name	name of pssm Matrix file

Value

feature vector of length 80

References

Y. Wang, Y. Ding, F. Guo, L. Wei, and J. J. P. o. Tang, "Improved detection of DNA-binding proteins via compression technology on PSSM information," vol. 12, no. 9, 2017.

Y. Wang, Y. Ding, J. Tang, Y. Dai, F. J. I. A. t. o. c. b. Guo, and bioinformatics, "CrystalM: a multi-view fusion approach for protein crystallization prediction," 2019.

Examples

```
X<-dwt_PSSM(system.file("extdata", "C7GQS7.txt.pssm", package="PSSMCOOL"))
```

EDP_EEDP_MEDP *EDP_EEDP_MEDP feature vector*

Description

these are three feature vectors (EDP, EEDP, MEDP) which are used for prediction of protein structural class for low-similarity sequences.at first ED-PSSM Matrix with 20*20 dimensions is constructed from PSSM Matrix then by using this Matrix, EDP and EEDP vectors are obtained eventually MEDP feature vector is obtained by fusing these vectors.

Usage

```
EDP_EEDP_MEDP(pssm_name)
```

Arguments

`pssm_name`	is name of PSSM Matrix file

Value

a list of three feature vectors (EDP, EEDP, MEDP)

References

Zhang, L., Zhao, X. and Kong, L. (2014) Predict protein structural class for low-similarity sequences by evolutionary difference information into the general form of Chou's pseudo amino acid composition, Journal of Theoretical Biology, 355, 105-110.

Examples

```
X<-EDP_EEDP_MEDP(paste0(system.file("extdata",package="PSSMCOOL"),"/C7GS61.txt.pssm"))
```

FPSSM *D-FPSSM and SF-PSSM feature vectors*

Description

This function produces list of two feature vectors named D-FPSSM and S-FPSSM which then used by FPSSM2 function to construct feature vector of length 100 for each pair of proteins which then used for protein-protein interaction prediction in each dataset.

Usage

```
FPSSM(pssm_name, hk = 20)
```

Arguments

`pssm_name`	name of PSSM Matrix file
`hk`	a parameter that indicates which amino acid alphabet must be used

Value

two feature vectors of different length which is used in later steps.

References

Zahiri, J., et al. (2013) PPIevo: protein-protein interaction prediction from PSSM based evolutionary information, Genomics, 102, 237-242.

Examples

```
X<-FPSSM(system.file("extdata","C7GQS7.txt.pssm",package="PSSMCOOL"),20)
```

FPSSM2 *Mixture of Two FPSSM Features*

Description

This function takes two PSSM files as argument and uses FPSSM function for making feature vector of length 100 correspond to this pair of proteins.

Usage

```
FPSSM2(pssm_name1, pssm_name2, hk)
```

Arguments

pssm_name1	The name of first PSSM Matrix file
pssm_name2	The name of second PSSM Matrix file
hk	a parameter that indicates which amino acid alphabet must be used

Value

Feature vector of length 100

References

Zahiri, J., et al. (2013) PPIevo: protein-protein interaction prediction from PSSM based evolutionary information, Genomics, 102, 237-242.

See Also

entropy

mutinformation

Examples

```
s1<-system.file("extdata","C7GQS7.txt.pssm",package="PSSMCOOL")
s2<-system.file("extdata","C7GRQ3.txt.pssm",package="PSSMCOOL")
s<-FPSSM2(s1,s2,8)
```

`grey_pssm_pseAAC` *grey pssm feature vector*

Description

This function produces a feature vector of length 100 which the first 20 components of this vector is the normalized occurrence frequency of the native amino acids in the protein. the next 20 components are mean of 20 PSSM columns and grey system model approach as elaborated in (Min et al. 2013) is used to define the next 60 components.

Usage

```
grey_pssm_pseAAC(pssm_name)
```

Arguments

`pssm_name`	name of PSSM matrix file

Value

feature vector of length 100

References

J.-L. Min, X. Xiao, and K.-C. J. B. r. i. Chou, "iEzy-Drug: A web server for identifying the interaction between enzymes and drugs in cellular networking," vol. 2013, 2013.

X. Xiao, M. Hui, and Z. J. T. J. o. m. b. Liu, "IAFP-Ense: an ensemble classifier for identifying antifreeze protein by incorporating grey model and PSSM into PseAAC," vol. 249, no. 6, pp. 845-854, 2016.

M. Kabir et al., "Improving prediction of extracellular matrix proteins using evolutionary information via a grey system model and asymmetric under-sampling technique," vol. 174, pp. 22-32, 2018.

Examples

```
X<-grey_pssm_pseAAC(system.file("extdata", "C7GQS7.txt.pssm", package="PSSMCOOL"))
```

`kiderafactor` *kiderafactor feature*

Description

For product of this feature vector similar to smoothed_PSSM feature, firstly PSSM Matrix is smoothed by appending zero vectors to its head and tail and utilizing sliding window of size odd, then this smoothed PSSM Matrix is condensed by the Kidera factors to produce feature vector for each residue.

Usage

```
kiderafactor(pssm_name, v = NULL)
```

Arguments

`pssm_name` name of PSSM Matrix file

`v` vector of amino acids positions which we want to produce feature vector for them.

Value

matrix of feature vectors

References

C. Fang, T. Noguchi, H. J. I. j. o. d. m. Yamana, and bioinformatics, "Condensing position-specific scoring matrixs by the Kidera factors for ligand-binding site prediction," vol. 12, no. 1, pp. 70-84, 2015.

See Also

`smoothed_PSSM`

Examples

```
X<-kiderafactor(system.file("extdata", "C7GQS7.txt.pssm", package="PSSMCOOL"),c(2,3,8,9))
```

k_mers *3-mer and 2-mer in dataframe*

Description

This function produces all possible 2-mers or 3-mers by counting paths of length 2 or 3 in a dataframe which is thought as a graph

Usage

```
k_mers(s, h)
```

Arguments

s	a dataframe with 2 columns
h	is length of k-mer

Value

all k-mers by counting paths of length h in dataframe which is considered as a graph

Examples

```
s1<-LETTERS[1:4]
s2<-LETTERS[3:6]
s<-data.frame(s1,s2)
dc<-k_mers(s,3)
```

k_separated_bigrams_pssm *k_separated_bigrams_pssm feature vector*

Description

This feature is almost identical to the dpc_pssm feature, and in fact the DPC feature is part of this feature (for k=1) and for two different columns, considers rows that differ by the size of the unit k.

Usage

```
k_separated_bigrams_pssm(pssm_name, k = 1)
```

Arguments

pssm_name	is name of PSSM Matrix file
k	a parameter that specifies separated length between amino acids

Value

a feature vector of length 400

References

Saini, H., et al.(2016) Protein Fold Recognition Using Genetic Algorithm Optimized Voting Scheme and Profile Bigram.

Examples

```
 X<-k_separated_bigrams_pssm(system.file("extdata", "C7GQS7.txt.pssm", package="PSSMCOOL"),1)
```

LPC_PSSM *Linear predictive coding feature*

Description

This function uses Linear predictive coding algorithm for each column of PSSM Matrix . so in this script lpc function is used which produces a 14-dimensional vector for each column, since PSSM has 20 column eventually it will be obtained a 20*14=280 dimensional feature vector for each PSSM Matrix by this function.

Usage

```
LPC_PSSM(pssm_name)
```

Arguments

pssm_name	name of PSSM Matrix file

Value

feature vector of length 280

References

L. Li et al., "PSSP-RFE: accurate prediction of protein structural class by recursive feature extraction from PSI-BLAST profile, physical-chemical property and functional annotations," vol. 9, no. 3, 2014.

Examples

```
X<-LPC_PSSM(system.file("extdata", "C7GQS7.txt.pssm", package="PSSMCOOL"))
```

MBMGACPSSM *MBMGACPSSM feature*

Description

In this function three different autocorrelation descriptors based on PSSM are adopted, which include: normalized Moreau-Broto autocorrelation, Moran autocorrelation and Geary autocorrelation descriptors.Autocorrelation descriptor is a powerful statistical tool and defined based on the distribution of amino acid properties along the sequence, which measures the correlation between two residues separated by a distance of d in terms of their evolution scores.

Usage

```
MBMGACPSSM(pssm_name)
```

Arguments

`pssm_name`	name of PSSM Matrix file

Value

feature vector of length 560

References

Y. Liang, S. Liu, S. J. M. C. i. M. Zhang, and i. C. Chemistry, "Prediction of protein structural class based on different autocorrelation descriptors of position-specific scoring matrix," vol. 73, no. 3, pp. 765-784, 2015.

Examples

```
X<-MBMGACPSSM(system.file("extdata", "C7GQS7.txt.pssm", package="PSSMCOOL"))
```

pse_pssm *pseudo position-specific scoring matrix feature*

Description

This feature vector is combination of F_{PSSM} feature vector and vector of correlation factors correspond to 20 columns in PSSM Matrix. F_{PSSM} actually is mean of PSSM Matrix columns of length 20.

Usage

```
pse_pssm(pssm_name, g = 1)
```

Arguments

`pssm_name` is the name of PSSM matrix file

`g` a parameter Which its size corresponds to the database used.

Value

feature vector of length 20+20*g

References

D.-J. Yu et al., "Learning protein multi-view features in complex space," vol. 44, no. 5, pp. 1365-1379, 2013.

Chou, K.C. and Shen, H.B. (2007) MemType-2L: a web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM, Biochemical and Biophysical Research Communications, 360, 339-345.

Examples

```
X<-pse_pssm(system.file("extdata", "C7GQS7.txt.pssm", package="PSSMCOOL"))
```

`pssm400` *PSSM400 feature*

Description

This function firstly normalizes PSSM Matrix by formula: $P - min(P)/max(P) - min(P)$ then for any standard amino acid specifies its position in protein sequence whereby a sub-matrix from PSSM corresponding to these positions will be extracted, then for this sub-matrix computes `colSums` of its columns to create a vector of length 20, eventually a feature vector of length 400 will be obtained.

Usage

```
pssm400(pssm_name)
```

Arguments

`pssm_name` name of PSSM Matrix file

Value

feature vector of length 400

Note

if a specific amino acid did not exist in protein then `colSums` of whole PSSM is computed.

Examples

```
X<-pssm400(system.file("extdata","C7GQS7.txt.pssm",package="PSSMCOOL"))
```

PSSMBLOCK *PSSM BLOCK feature vector*

Description

In this feature at first PSSM Matrix is divided to Blocks based on Number N which user imports. Then for each Block mean of columns is computed to get 20-dimensional vector, eventually by appending these vectors to each other final feature vector is obtained.

Usage

```
PSSMBLOCK(pssm_name, N = 5)
```

Arguments

pssm_name	neme of PSSM Matrix file
N	number of blocks

Value

feature vector that it’s length depends on parameter N

References

J.-Y. An, L. Zhang, Y. Zhou, Y.-J. Zhao, and D.-F. J. J. o. c. Wang, "Computational methods using weighed-extreme learning machine to predict protein self-interactions with protein evolutionary information," vol. 9, no. 1, p. 47, 2017.

Examples

```
X<-PSSMBLOCK(system.file("extdata", "C7GQS7.txt.pssm", package="PSSMCOOL"),5)
```

pssm_ac *auto covariance transformation feature vector*

Description

The AC variable measures the correlation of the same property between two residues separated by a distance of lg along the sequence

Usage

```
pssm_ac(pssm_name, lg = 10)
```

Arguments

`pssm_name`	name of the PSSM Matrix file
`lg`	a parameter which indicates distance between two residues

Value

feature vector which its length depends on parameter lg. by default lg is 10 hence feature vector would be of length 200.

Note

in use of this function The lg parameter must be less than the length of the smallest sequence in the database.

References

Dong, Q., Zhou, S. and Guan, J. (2009) A new taxonomy-based protein fold recognition approach based on autocross-covariance transformation, Bioinformatics, 25, 2655-2662.

Examples

```
X<-pssm_ac(system.file("extdata", "C7GQS7.txt.pssm", package="PSSMCOOL"))
```

pssm_cc *Cross covarianse feature vector*

Description

The PSSM-CC variable measures the correlation of two different properties between two residues separated by a distance of lg along the sequence.

Usage

```
pssm_cc(pssm_name, g = 10)
```

Arguments

pssm_name	name of PSSM Matrix file
g	shortest protein length in dataset minus one

Value

feature vector of length 3800

References

Dong, Q., Zhou, S. and Guan, J. (2009) A new taxonomy-based protein fold recognition approach based on autocross-covariance transformation, Bioinformatics, 25, 2655-2662.

Examples

```
X<-pssm_cc(system.file("extdata","C7GQS7.txt.pssm",package="PSSMCOOL"))
```

pssm_composition *PSSM-COMPOSITION feature*

Description

This feature, which stands for auto covariance transformation, for jth column calculates the average of this column, and then subtracts the resulting number from the elements on the i and (i + g)th rows of this column, and finally multiplies them. by changing the variable i from 1 to L-g, it calculates the sum of these, since the variable j changes between 1 and 20, and the variable g between 1 and 20 eventually a feature vector of length 200 will be obtained.

Usage

```
pssm_composition(pssm_name)
```

Arguments

`pssm_name`	name of PSSM Matrix files

Value

feature vector of length 400

References

L. Zou, C. Nan, and F. J. B. Hu, "Accurate prediction of bacterial type IV secreted effectors using amino acid composition and PSSM profiles," vol. 29, no. 24, pp. 3135-3142, 2013.

Examples

```
X<-pssm_composition(system.file("extdata", "C7GQS7.txt.pssm", package="PSSMCOOL"))
```

PSSM_SD *PSSM-SD feature*

Description

In this feature, by considering a specific column, at first sum of all components in this column is denoted by "L", then starting from the first row in this column, the components are added together to reaching a value less than or equal to 25 is calculated and stored . In the next step, the same work is done starting from the first row to reaching a value less than or equal to 50 is done started from the last row, To reaching 25 By appending these saved numbers together for each column, a vector of length 4 is obtained. If this is done for all the columns and the obtained vectors are connected to each other, for each protein, a feature vector of length 80 is obtained which its name is PSSM-SD.

Usage

```
PSSM_SD(pssm_name)
```

Arguments

`pssm_name`	name of PSSM Matrix file

Value

feature vector of length 80

References

A. Dehzangi, K. Paliwal, J. Lyons, A. Sharma, A. J. I. A. T. o. C. B. Sattar, and Bioinformatics, "A segmentation-based method to extract structural and evolutionary features for protein fold recognition," vol. 11, no. 3, pp. 510-519, 2014.

Examples

```
X<-PSSM_SD(system.file("extdata", "C7GQS7.txt.pssm", package="PSSMCOOL"))
```

---

pssm_seg *PSSM-Seg feature vector*

Description

This feature vector uses PSSM-SD to produce Segmented Auto Covariance Features.

Usage

```
pssm_seg(pssm_name, m = 4)
```

Arguments

pssm_name	name of PSSM Matrix file
m	a parameter between 1 and 11

Value

feature vector of length 100

References

A. Dehzangi, K. Paliwal, J. Lyons, A. Sharma, A. J. I. A. T. o. C. B. Sattar, and Bioinformatics, "A segmentation-based method to extract structural and evolutionary features for protein fold recognition," vol. 11, no. 3, pp. 510-519, 2014.

See Also

PSSM_SD

Examples

```
q<-pssm_seg(system.file("extdata", "C7GQS7.txt.pssm", package="PSSMCOOL"),3)
```

RPM_PSSM *RPM-PSSM feature vector*

Description

In this feature The idea is similar to the probe concept used in microarray technologies, where probes are used to identify genes. For the convenience, we call it residue probing method. In our application, each probe is an amino acid, which corresponds to a particular column in the PSSM profiles. For each probe, we average the PSSM scores of all the amino acids in the associated column with a PSSM value greater than zero in the sequence, which leads to a 1 20 feature vector. Once again, for the 20 probes, the final feature for each protein sequence is a 1 400 vector.

Usage

```
RPM_PSSM(pssm_name)
```

Arguments

pssm_name	name of PSSM Matrix file

Value

RPM-PSSM feature vector of length 400

References

Jeong, J.C., Lin, X. and Chen, X.W. (2011) On position-specific scoring matrix for protein function prediction , IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM, 8, 308-315.

Examples

```
X<- RPM_PSSM(system.file("extdata","C7GRQ3.txt.pssm",package="PSSMCOOL"))
```

rpssm *RPSSM feature*

Description

To obtain this feature, first the columns of the PSSM matrix are merged to obtain an L*10 matrix. Then, with a relationship similar to the auto covariance transformation feature, this feature with a length of 110 is obtained from this matrix.

Usage

```
rpssm(pssm_name)
```

Arguments

pssm_name	name of PSSM Matrix file

Value

feature vector of length 110

References

Ding, S., et al. (2014) A protein structural classes prediction method based on predicted secondary structure and PSI-BLAST profile, Biochimie, 97, 60-65.

Examples

```
X<-rpssm(system.file("extdata", "C7GQS7.txt.pssm", package="PSSMCOOL"))
```

---

scsh2 *SCSH Feature vector*

Description

This function gets a PSSM Matrix as input and extracts corresponding protein and consunsus sequence from it. By placing these two vectors next to each other, a dataframe is created. In each row of this dataframe, each component is connected to the next row components by an arrow. so a directed graph is produced. next by use of previous functions a feature vector of length 400 or 8000 is created.

Usage

```
scsh2(pssm_name, k = 2)
```

Arguments

pssm_name	is name of PSSM Matrix file
k	a parameter indicates length of k-mer

Value

feature vector of length 400 or 8000

References

Zahiri, J., et al., LocFuse: human protein–protein interaction prediction via classifier fusion using protein localization information. Genomics, 2014. 104(6): p. 496-503.

Examples

```
X<- scsh2(system.file("extdata","C7GRQ3.txt.pssm",package="PSSMCOOL"),2)
```

single_Average *single Average feature*

Description

This descriptor is a variant of the Average Block descriptor and is designed to group together rows related to the same amino acid, thus considering domains of a sequence with similar conservation rates.

Usage

```
single_Average(pssm_name)
```

Arguments

pssm_name	name of PSSM Matrix file

Value

feature vector of length 400

References

L. Nanni, A. Lumini, and S. J. T. S. W. J. Brahnam, "An empirical study of different approaches for protein classification," vol. 2014, 2014.

See Also

Averag_Block

Examples

```
X<-single_Average(system.file("extdata", "C7GQS7.txt.pssm", package="PSSMCOOL"))
```

smoothed_PSSM *smoothed PSSM feature*

Description

In this function at first a Matrix called smoothed-PSSM is constructed from PSSM Matrix by applying "ws" parameter which called sliding window size and taken from user and usually is equals to 7. Then using other window size parameter "w" which usually equals to 11 at each position smoothed feature vector is constructed.

Usage

```
smoothed_PSSM(pssm_name, ws = 7, w = 50, v = NULL)
```

Arguments

`pssm_name`	name of PSSM Matrix file
`ws`	window size for smoothing PSSM Matrix
`w`	window size for extracting feature vector
`v`	vector of desired positions to extract their features

Details

In the construction of a smoothed PSSM, each row vector of a residue α_i is represented and smoothed by the summation of ws surrounding row vectors $(V_{smoothed_i} = V_{i-(ws-1)/2} + ... + V_i + ... + V_{i+(ws+1)/2})$ For the N-terminal and C-terminal of a protein, (w-1)/2 ZERO vectors, are appended to the head or tail of a smoothed PSSM profile. Using the smoothed PSSM encoding scheme the feature vector of a residue α_i is represented by $(V_{smoothed_i-(ws-1)/2}, ..., V_{smoothed_i}, ..., V_{smoothed_i+(ws+1)/2})$ The feature values in each vector are normalized to a range between -1 and 1.

Value

a matrix of feature vectors

References

Cheng, C.W., et al. (2008) Predicting RNA-binding sites of proteins using support vector machines and evolutionary information, BMC Bioinformatics, 9 Suppl 12, S6.

See Also

kiderafactor

Examples

```
X<-smoothed_PSSM(system.file("extdata", "C7GQS7.txt.pssm", package="PSSMCOOL"),7,11,c(2,3,8,9))
```

`SOMA_PSSM` *SOMA PSSM Feature*

Description

In this function each column can be viewed as a stochastic time series, and each PSSM contains 20 columns, in other words, each PSSM contains 20 stochastic time series and Second-order moving average (SOMA) algorithm is applied to these columns to extract SOMA PSSM feature vector.

Usage

```
SOMA_PSSM(pssm_name)
```

Arguments

`pssm_name`	name of PSSM file

Value

feature vector of length 160

References

Y. Liang, S. J. J. o. M. G. Zhang, and Modelling, "Predict protein structural class by incorporating two different modes of evolutionary information into Chou's general pseudo amino acid composition," vol. 78, pp. 110-117, 2017.

Examples

```
X<-SOMA_PSSM(system.file("extdata", "C7GQS7.txt.pssm", package="PSSMCOOL"))
```

`SVD_PSSM` *Singular Value Decomposition (SVD)*

Description

Singular value decomposition is a general purpose matrix factorization approach that has many useful applications in signal processing and statistics. In this function SVD is applied to a matrix representation of a protein with the aim of reducing its dimensionality Given an input matrix Mat with dimensions N*M SVD is used to calculate its factorization of the form: $Mat = U\Sigma V$, where Σ is a diagonal matrix whose diagonal entries are known as the singular values of Mat. The resulting descriptor is the ordered set of singular values: $SVD \in \mathcal{R}^L$, where L=min(M,N). and here `svd` function is used for this purpose.

Usage

```
SVD_PSSM(pssm_name)
```

Arguments

`pssm_name`	name of PSSM Matrix file

Value

feature vector of length 20

References

L. Nanni, A. Lumini, and S. J. T. S. W. J. Brahnam, "An empirical study of different approaches for protein classification," vol. 2014, 2014.

Examples

```
X<-SVD_PSSM(system.file("extdata", "C7GQS7.txt.pssm", package="PSSMCOOL"))
```

three_mer *3-Mer and 2-Mer*

Description

This function produces all possible k-mers from 20 amino acids for use in other functions.

Usage

```
three_mer(k)
```

Arguments

k	is length of k-mer which user imports

Value

a matrix which its first row includes all k-mers

Examples

```
ax<-three_mer(3)
```

trigrame_pssm *trigrame feature vector*

Description

This feature vector is 8000-dimentional feature vector wich is computed from tri-gram probability matrix T obtained from PSSM Matrix.to achieve this purpose elements in three successive rows and arbitrary columns are multiplied together then these results are added together by changing variable i from 1 to L-1, which i is counter of row and L indicates protein length. since there are 20 columns thus final feature vector would be of length 8000.

Usage

```
trigrame_pssm(pssm_name)
```

Arguments

pssm_name	name of PSSM Matrix file

Value

feature vector of lenght 8000

References

Paliwal, K.K., et al. (2014) A tri-gram based feature extraction technique using linear probabilities of position specific scoring matrix for protein fold recognition, IEEE transactions on nanobioscience, 13, 44-50

Examples

```
X<-trigrame_pssm(paste0(system.file("extdata",package="PSSMCOOL"),"/C7GSI6.txt.pssm"))
```

Index