Package: protHMM (via r-universe)

September 5, 2024

Type Package

Title Protein Feature Extraction from Profile Hidden Markov Models

Version 0.1.1

Maintainer Shayaan Emran <shayaan.emran@gmail.com>

Description Calculates a comprehensive list of features from profile hidden Markov models (HMMs) of proteins. Adapts and ports features for use with HMMs instead of Position Specific Scoring Matrices, in order to take advantage of more accurate multiple sequence alignment by programs such as 'HHBlits' <DOI:10.1038/nmeth.1818> and 'HMMer' <http://hmmer.org>. Features calculated by this package can be used for protein fold classification, protein structural class prediction, sub-cellular localization and protein-protein interaction, among other tasks. Some examples of features extracted are found in Song et al. (2018) <DOI:10.3390/app8010089>, Jin & Zhu (2021) <DOI:10.1155/2021/8629776>, Lyons et al. (2015) <DOI:10.1109/tnb.2015.2457906> and Saini et al. (2015) <DOI:10.1016/j.jtbi.2015.05.030>.

License GPL (>= 3)

Encoding UTF-8

LazyData true

RoxygenNote 7.2.3

Imports gtools, utils, stats, phonTools

Suggests covr, knitr, rmarkdown, testthat (>= 3.0.0)

Config/testthat/edition 3

VignetteBuilder knitr

URL https://github.com/semran9/protHMM/, https://semran9.github.io/protHMM/

BugReports https://github.com/semran9/protHMM/issues

Repository https://semran9.r-universe.dev

RemoteUrl https://github.com/semran9/prothmm

RemoteRef HEAD

RemoteSha c87243606574287e068b584d4eca5ec86f6df25c

Contents

chmm . . . 2
fp_hmm . . . 3
hmm_ac . . . 4
hmm_bigrams . . . 5
hmm_cc . . . 5
hmm_distance . . . 6
hmm_GA . . . 7
hmm_GSD . . . 8
hmm_LBP . . . 8
hmm_LPC . . . 9
hmm_MA . . . 10
hmm_MB . . . 10
hmm_read . . . 11
hmm_SCSH . . . 12
hmm_SepDim . . . 13
hmm_Single_Average . . . 13
hmm_smooth . . . 14
hmm_svd . . . 15
hmm_trigrams . . . 16
IM_psehmm . . . 16
pse_hmm . . . 17

Index **19**

---

chmm *chmm*

Description

This feature begins by creating a CHMM, which is created by constructing 4 matrices, A, B, C, D from the original HMM H. A contains the first 75 percent of the original matrix H row-wise, B the last 75 percent, C the middle 75 percent and D the entire original matrix. These are then merged to create the new CHMM Z. From there, the Bigrams feature is calculated with a flattened 20 x 20 matrix B, in which $B[i, j] = \sum_{a=1}^{L-1} Z_{a,i} \times Z_{a+1,j}$. H corresponds to the original HMM matrix, and L is the number of rows in Z. Local Average Group, or LAG is then calculated by splitting up the CHMM into 20 groups along the length of the protein sequence and calculating the sums of each of the columns of each group, making a 1 x 20 vector per group, and a length 20 x 20 vector for all groups. These features are then fused.

Usage

```
chmm(hmm)
```

Arguments

hmm	The name of a profile hidden markov model file.

Value

A fusion vector of length 800.

A LAG vector of length 400.

A Bigrams vector of length 400.

References

An, J., Zhou, Y., Zhao, Y., & Yan, Z. (2019). An Efficient Feature Extraction Technique Based on Local Coding PSSM and Multifeatures Fusion for Predicting Protein-Protein Interactions. Evolutionary Bioinformatics, 15, 117693431987992.

Examples

```
h<- chmm(system.file("extdata", "1DLHA2-7", package="protHMM"))
```

---

fp_hmm *fp_hmm*

Description

This feature consists of two vectors, d, s. Vector d corresponds to the sums across the sequence for each of the 20 amino acid columns. Vector s corresponds to a flattened matrix $S[i,j] = \sum_{k=1}^{L} H[k,j] \times \delta[k,i]$ in which $\delta[k,i] = 1$ when $A_i = H[k,j]$. A refers to a list of all possible amino acids, i, j span from $1:20$.

Usage

```
fp_hmm(hmm)
```

Arguments

hmm	The name of a profile hidden markov model file.

Value

A vector of length 20.

A vector of length 400.

References

Zahiri, J., Yaghoubi, O., Mohammad-Noori, M., Ebrahimpour, R., & Masoudi-Nejad, A. (2013). PPIevo: Protein–protein interaction prediction from PSSM based evolutionary information. Genomics, 102(4), 237–242.

Examples

```
h<- fp_hmm(system.file("extdata", "1DLHA2-7", package="protHMM"))
```

---

hmm_ac *hmm_ac*

Description

This feature calculates the covariance between two residues separated by a lag value within the same amino acid emission frequency column along the protein sequence.

Usage

```
hmm_ac(hmm, lg = 4)
```

Arguments

hmm	The name of a profile hidden markov model file.
lg	The lag value, which indicates the distance between residues.

Value

A vector of length 20 $\times$ the lag value; by default this is a vector of length 80.

Note

The lag value must be less than the length of the protein sequence

References

Dong, Q., Zhou, S., & Guan, J. (2009). A new taxonomy-based protein fold recognition approach based on autocross-covariance transformation. Bioinformatics, 25(20), 2655–2662.

Examples

```
h<- hmm_ac(system.file("extdata", "1DLHA2-7", package="protHMM"))
```

hmm_bigrams *hmm_bigrams*

Description

This feature is calculated with a 20 x 20 matrix B, in which $B[i,j] = \sum_{a=1}^{L-1} H_{a,i} H_{a+1,j}$. H corresponds to the original HMM matrix, and L is the number of rows in H. Matrix B is then flattened to a feature vector of length 400, and returned.

Usage

```
hmm_bigrams(hmm)
```

Arguments

hmm	The name of a profile hidden markov model file.

Value

A vector of length 400

References

Lyons, J., Dehzangi, A., Heffernan, R., Yang, Y., Zhou, Y., Sharma, A., & Paliwal, K. K. (2015). Advancing the Accuracy of Protein Fold Recognition by Utilizing Profiles From Hidden Markov Models. IEEE Transactions on Nanobioscience, 14(7), 761–772.

Examples

```
h<- hmm_bigrams(system.file("extdata", "1DLHA2-7", package="protHMM"))
```

hmm_cc *hmm_cc*

Description

The feature calculates the covariance between different residues separated along the protein sequences by a lag value across different amino acid emission frequency columns.

Usage

```
hmm_cc(hmm, lg = 4)
```

Arguments

hmm	The name of a profile hidden markov model file.
lg	The lag value, which indicates the distance between residues.

Value

A vector of length 20 x 19 x the lag value; by default this is a vector of length 1520.

Note

The lag value must less than the length of the amino acid sequence.

References

Dong, Q., Zhou, S., & Guan, J. (2009). A new taxonomy-based protein fold recognition approach based on autocross-covariance transformation. Bioinformatics, 25(20), 2655–2662.

Examples

```
h<- hmm_cc(system.file("extdata", "1DLHA2-7", package="protHMM"))
```

`hmm_distance` *hmm_distance*

Description

This feature calculates the cosine distance matrix between two HMMs A and B before dynamic time warp is applied to the distance matrix calculate the cumulative distance between the HMMs, which acts as a measure of similarity, The cosine distance matrix D is found to be $D[a_i, b_j] = 1 - \frac{a_i b_j^T}{a_i a_i^T b_j b_j^T}$, in which a_i and a_i refer to row vectors of A and B respectively. This in turn means that D is of dimensions $nrow(A), nrow(b)$. Dynamic time warp then calculates the cumulative distance by calculating matrix $C[i, j] = min(C[i-1, j], C[i, j-1], C[i-1, j-1]) + D[i, j]$, where $C_{i,j}$ is 0 when i or j are less than 1. The lower rightmost point of the matrix C is then returned as the cumulative distance between proteins.

Usage

```
hmm_distance(hmm_1, hmm_2)
```

Arguments

`hmm_1`	The name of a profile hidden markov model file.
`hmm_2`	The name of another profile hidden markov model file.

Value

A double that indicates distance between the two proteins.

References

Lyons, J., Paliwal, K. K., Dehzangi, A., Heffernan, R., Tsunoda, T., & Sharma, A. (2016). Protein fold recognition using HMM–HMM alignment and dynamic programming. Journal of Theoretical Biology, 393, 67–74.

Examples

```
h<- hmm_distance(system.file("extdata", "1DLHA2-7", package="protHMM"),
system.file("extdata", "1TEN-7", package="protHMM"))
```

---

hmm_GA *hmm_GA*

Description

This feature calculates the Geary autocorrelation of each amino acid type for each distance d less than or equal to the lag value and greater than or equal to 1.

Usage

```
hmm_GA(hmm, lg = 9)
```

Arguments

hmm	The name of a profile hidden markov model file.
lg	The lag value, which indicates the distance between residues.

Value

A vector of length lg $\times$ 20, by default this is 180.

Note

The lag value must be less than the length of the protein sequence

References

Liang, Y., Liu, S., & Zhang. (2015). Prediction of Protein Structural Class Based on Different Autocorrelation Descriptors of Position–Specific Scoring Matrix. MATCH: Communications in Mathematical and in Computer Chemistry, 73(3), 765–784.

Examples

```
h<- hmm_GA(system.file("extdata", "1DLHA2-7", package="protHMM"))
```

hmm_GSD *hmm_GSD*

Description

This feature initially creates a grouping matrix G by assigning each position a number $1:3$ based on the value at each position of HMM matrix H; 1 represents the low probability group, 2 the medium and 3 the high probability group. The number of total points in each group for each column is then calculated, and the sequence is then split based upon the the positions of the 1st, 25th, 50th, 75th and 100th percentile (last) points for each of the three groups, in each of the 20 columns of the grouping matrix. Thus for column j, $S(k,j,z) = \sum_{i=1}^{(z)*.25*N} |G[i,j] = k|$, where k is the group number, $z = 1:4$ and N corresponds to number of rows in matrix G.

Usage

```
hmm_GSD(hmm)
```

Arguments

hmm	The name of a profile hidden markov model file.

Value

A vector of length 300.

References

Jin, D., & Zhu, P. (2021). Protein Subcellular Localization Based on Evolutionary Information and Segmented Distribution. Mathematical Problems in Engineering, 2021, 1–14.

Examples

```
h<- hmm_GSD(system.file("extdata", "1DLHA2-7", package="protHMM"))
```

hmm_LBP *hmm_LBP*

Description

This feature uses local binary pattern with a neighborhood of radius 1 and 8 sample points to extract features from the HMM. A 256 bin histogram is extracted as a 256 length feature vector.

Usage

```
hmm_LBP(hmm)
```

Arguments

hmm	The name of a profile hidden markov model file.

Value

A vector of length 256.

References

Li, Y., Li, L., Wang, L., Yu, C., Wang, Z., & You, Z. (2019). An Ensemble Classifier to Predict Protein–Protein Interactions by Combining PSSM-based Evolutionary Information with Local Binary Pattern Model. International Journal of Molecular Sciences, 20(14), 3511.

Examples

```
h<- hmm_LBP(system.file("extdata", "1DLHA2-7", package="protHMM"))
```

---

hmm_LPC *hmm_LPC*

Description

This feature uses linear predictive coding (LPC) to map each HMM to a $20 \times 14 = 280$ dimensional vector, where for each of the 20 columns of the HMM, LPC is used to extract a 14 dimensional vector D_n

Usage

```
hmm_LPC(hmm)
```

Arguments

hmm	The name of a profile hidden markov model file.

Value

A vector of length 280.

References

Qin, Y., Zheng, X., Wang, J., Chen, M., & Zhou, C. (2015). Prediction of protein structural class based on Linear Predictive Coding of PSI-BLAST profiles. Central European Journal of Biology, 10(1).

Examples

```
h<- hmm_LPC(system.file("extdata", "1DLHA2-7", package="protHMM"))
```

hmm_MA *hmm_MA*

Description

This feature calculates the normalized Moran autocorrelation of each amino acid type, for each distance d less than or equal to the lag value and greater than or equal to 1.

Usage

```
hmm_MA(hmm, lg = 9)
```

Arguments

hmm	The name of a profile hidden markov model file.
lg	The lag value, which indicates the distance between residues.

Value

A vector of length lg $\times$ 20, by default this is 180.

Note

The lag value must be less than the length of the protein sequence

References

Liang, Y., Liu, S., & Zhang. (2015). Prediction of Protein Structural Class Based on Di fferent Autocorrelation Descriptors of Position–Specific Scoring Matrix. MATCH: Communications in Mathematical and in Computer Chemistry, 73(3), 765–784.

Examples

```
h<- hmm_MA(system.file("extdata", "1DLHA2-7", package="protHMM"))
```

hmm_MB *hmm_MB*

Description

This feature calculates the normalized Moreau-Broto autocorrelation of each amino acid type, for each distance d less than or equal to the lag value and greater than or equal to 1.

Usage

```
hmm_MB(hmm, lg = 9)
```

Arguments

hmm	The name of a profile hidden markov model file.
lg	The lag value, which indicates the distance between residues.

Value

A vector of length lg $\times$ 20, by default this is 180.

Note

The lag value must be less than the length of the protein sequence

References

Liang, Y., Liu, S., & Zhang. (2015). Prediction of Protein Structural Class Based on Different Autocorrelation Descriptors of Position–Specific Scoring Matrix. MATCH: Communications in Mathematical and in Computer Chemistry, 73(3), 765–784.

Examples

```
h<- hmm_MB(system.file("extdata", "1DLHA2-7", package="protHMM"))
```

---

hmm_read *hmm_read*

Description

Reads in the amino acid emission frequency columns of a profile hidden markov model matrix and converts each position to frequencies.

Usage

```
hmm_read(hmm)
```

Arguments

hmm	The name of a profile hidden markov model file.

Value

A 20 x L matrix, in which L is the sequence length.

Examples

```
h<- hmm_read(system.file("extdata", "1DLHA2-7", package="protHMM"))
```

hmm_SCSH *hmm_SCSH*

Description

This feature returns the 2 and 3-mer compositions of the protein sequence. This is done by first finding all possible 2 and 3-mers for any protein (20^2 and 20^3 permutations for 2 and 3-mers respectively). With those permutations, vectors of length 400 and 8000 are created, each point corresponding to one 2 or 3-mer. Then, the protein sequence that corresponds to the HMM scores is extracted, and put into a bipartite graph with the protein sequence. Each possible path of length 1 or 2 is found, and the corresponding vertices on the graph are noted as 2 and 3-mers. For each 2 or 3-mer found from these paths, 1 is added to the position that responds to that 2/3-mer in the 2-mer and 3-mer vectors , which are the length 400 and 8000 vectors created previously. The vectors are then returned.

Usage

```
hmm_SCSH(hmm)
```

Arguments

hmm	The name of a profile hidden markov model file.

Value

A vector of length 400.

A vector of length 8000.

References

Mohammadi, A. M., Zahiri, J., Mohammadi, S., Khodarahmi, M., & Arab, S. S. (2022). PSSM-COOL: a comprehensive R package for generating evolutionary-based descriptors of protein sequences from PSSM profiles. Biology Methods and Protocols, 7(1).

Examples

```
h_400<- hmm_SCSH(system.file("extdata", "1DLHA2-7", package="protHMM"))[[1]]
h_8000<- hmm_SCSH(system.file("extdata", "1DLHA2-7", package="protHMM"))[[2]]
```

hmm_SepDim *hmm_SepDim*

Description

This feature calculates the probabilistic expression of amino acid dimers that are spatially separated by a distance l. Mathematically, this is done with a 20 x 20 matrix F, in which $F[m,n] = \sum_{i=1}^{L-l} H_{i,m}H_{i+k,n}$. H corresponds to the original HMM matrix, and L is the number of rows in H. Matrix F is then flattened to a feature vector of length 400, and returned.

Usage

```
hmm_SepDim(hmm, l = 7)
```

Arguments

hmm	The name of a profile hidden markov model file.
l	Spatial distance between dimer residues.

Value

A vector of length 400

References

Saini, H., Raicar, G., Sharma, A., Lal, S. K., Dehzangi, A., Lyons, J., Paliwal, K. K., Imoto, S., & Miyano, S. (2015). Probabilistic expression of spatially varied amino acid dimers into general form of Chou's pseudo amino acid composition for protein fold recognition. Journal of Theoretical Biology, 380, 291–298.

Examples

```
h<- hmm_SepDim(system.file("extdata", "1DLHA2-7", package="protHMM"))
```

hmm_Single_Average *hmm_Single_Average*

Description

This feature groups together rows that are related to the same amino acid. This is done using a vector $SA(k)$, in which k spans $1:400$ and $SA(k) = avg_{i=1,2...L}H[i,j] \times \delta(P(i),A(z))$, in which H is the HMM matrix, P in the protein sequence, A is an ordered set of amino acids, the variables $j, z = 1:20$, the variable $k = j+20\times(z-1)$ when creating the vector, and $\delta()$ represents Kronecker's delta.

Usage

```
hmm_Single_Average(hmm)
```

Arguments

hmm	The name of a profile hidden markov model file.

Value

A vector of length 400.

References

Nanni, L., Lumini, A., & Brahnam, S. (2014). An Empirical Study of Different Approaches for Protein Classification. The Scientific World Journal, 2014, 1–17.

Examples

```
h<- hmm_Single_Average(system.file("extdata", "1DLHA2-7", package="protHMM"))
```

---

hmm_smooth *hmm_smooth*

Description

This feature smooths the HMM matrix H by using sliding window of length sw to incorporate information from up and downstream residues into each row of the HMM matrix. Each HMM row r_i is made into the summation of $r_{i-(sw/2)} + ...r_i... + r_{i+(sw/2)}$, for $i = 1 : L$, where L is the number of rows in H. For rows such as the beginning and ending rows, 0 matrices of dimensions $sw/2, 20$ are appended to the original matrix H.

Usage

```
hmm_smooth(hmm, sw = 7)
```

Arguments

hmm	The name of a profile hidden markov model file.
sw	The size of the sliding window.

Value

A matrix of dimensions L $\times$ 20.

References

Fang, C., Noguchi, T., & Yamana, H. (2013). SCPSSMpred: A General Sequence-based Method for Ligand-binding Site Prediction. IPSJ Transactions on Bioinformatics, 6(0), 35–42.

Examples

```
h<- hmm_smooth(system.file("extdata", "1DLHA2-7", package="protHMM"))
```

---

hmm_svd *hmm_svd*

Description

This feature uses singular value decomposition (SVD) to reduce the dimensionality of the inputted hidden markov model matrix. SVD factorizes a matrix C of dimensions i, j to $U[i, r] \times \Sigma[r, r] \times V[r, j]$. The diagonal values of Σ are known as the singular values of matrix C, and are what are returned with this function.

Usage

```
hmm_svd(hmm)
```

Arguments

hmm	The name of a profile hidden markov model file.

Value

A vector of length 20.

References

Song, X., Chen, Z., Sun, X., You, Z., Li, L., & Zhao, Y. (2018). An Ensemble Classifier with Random Projection for Predicting Protein–Protein Interactions Using Sequence and Evolutionary Information. Applied Sciences, 8(1), 89.

Examples

```
h<- hmm_svd(system.file("extdata", "1DLHA2-7", package="protHMM"))
```

hmm_trigrams *hmm_trigrams*

Description

This feature is calculated with a 20 x 20 x 20 block B, in which $B[i,j,k] = \sum_{a=1}^{L-2} H_{a,i} H_{a+1,j} H_{a+2,k}$. H corresponds to the original HMM matrix, and L is the number of rows in H. Matrix B is then flattened to a feature vector of length 8000, and returned.

Usage

```
hmm_trigrams(hmm)
```

Arguments

hmm	The name of a profile hidden markov model file.

Value

A vector of length 8000

References

Lyons, J., Dehzangi, A., Heffernan, R., Yang, Y., Zhou, Y., Sharma, A., & Paliwal, K. K. (2015). Advancing the Accuracy of Protein Fold Recognition by Utilizing Profiles From Hidden Markov Models. IEEE Transactions on Nanobioscience, 14(7), 761–772.

Examples

```
h<- hmm_trigrams(system.file("extdata", "1DLHA2-7", package="protHMM"))
```

IM_psehmm *IM_psehmm*

Description

The first twenty numbers of this feature correspond to the means of each column of the HMM matrix H. The rest of the features in the feature vector are found in matrix $T[i,j]$, where $T[i,j] = \frac{1}{L-i} \sum_{n=1}^{20-i} [H_{m,n} - H_{m,n+i}]^2, m = 1:L, i = 1:dandj = 1:20$.

Usage

```
IM_psehmm(hmm, d = 13)
```

Arguments

hmm	The name of a profile hidden markov model file.
d	The maximum distance between residues column-wise.

Value

A vector of length $20 + 20 \times d - d \times \frac{d+1}{2}$

Note

d must be less than 20.

References

Ruan, X., Zhou, D., Nie, R., & Guo, Y. (2020). Predictions of Apoptosis Proteins by Integrating Different Features Based on Improving Pseudo-Position-Specific Scoring Matrix. BioMed Research International, 2020, 1–13.

Examples

```
h<- IM_psehmm(system.file("extdata", "1DLHA2-7", package="protHMM"))
```

---

pse_hmm *pse_hmm*

Description

The first twenty numbers of this feature correspond to the means of each column of the HMM matrix H. The rest of the features in the feature vector are given by correlation of the ith most contiguous values along the chain per each amino acid column, where $0 < i < g + 1$. This creates a vector of $20 \times$ g, and this combines with the first 20 features to create the final feature vector.

Usage

```
pse_hmm(hmm, g = 15)
```

Arguments

hmm	The name of a profile hidden markov model file.
g	The contiguous distance between residues.

Value

A vector of length $20 + g \times 20$, by default this is 320.

Note

g must be less than the length of the protein sequence

References

Chou, K., & Shen, H. (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(2), 339–345.

Examples

```
h<- pse_hmm(system.file("extdata", "1DLHA2-7", package="protHMM"))
```

Index

chmm, 2

fp_hmm, 3

hmm_ac, 4
hmm_bigrams, 5
hmm_cc, 5
hmm_distance, 6
hmm_GA, 7
hmm_GSD, 8
hmm_LBP, 8
hmm_LPC, 9
hmm_MA, 10
hmm_MB, 10
hmm_read, 11
hmm_SCSH, 12
hmm_SepDim, 13
hmm_Single_Average, 13
hmm_smooth, 14
hmm_svd, 15
hmm_trigrams, 16

IM_psehmm, 16

pse_hmm, 17