protHMM: An R Package for Protein Feature Extraction using Profile Hidden Markov Models

6 July 2023

Summary

Proteins are one of the fundamental building blocks of life, involved in almost every process or reaction. As such, the study of protein structure and function is necessary to help understand more about newly discovered proteins. A number of biological methods are used to determine protein structure and function such as fluorescence microscopy and X-ray crystallography. However, these methods are time consuming; thus, a number of computational methods have been, and continue to be developed to identify protein structure and function. The development of these methods help us understand proteins and their complex biological functions for use in downstream tasks such as novel drug discovery.

Many of these methods rely on a key piece: the features extracted from the proteins used to train the model. protHMM implements 20 different feature extraction methods from HMM representation of proteins for use in such models. Several methods have been ported for use with HMMs from PSSMs, while others were used with HMMs originally. Examples of features implemented in protHMM include the CHMM vector (An et al. 2019), the bigrams and trigrams vectors (Lyons et al. 2015), singular value decomposition (Song et al. 2018) and the separated dimers vector (Saini et al. 2015).

Installation of protHMM can be done through the official CRAN repository using the command `install.packages("protHMM")`.

Statement of need

Lyons et al. (2015) and Xia et al. (2017) have both shown the useful of features derived from HMM representations of proteins in bioinformatics tasks such as protein fold classification. However, there are limited software implementations of the feature extraction methods discussed in the aforementioned papers.

protHMM implements a comprehensive library of feature extraction methods to apply to profile hidden markov model (HMM) representations of proteins. protHMM implements features used for subcellular localization, protein-protein interaction, protein structural class predication, and protein fold classification. This implementation allows profile hidden markov model representations of proteins generated by HHBlits and HMMer to be used smoothly in the bioinformatics workflow.

PSSMCOOL is a similar package to protHMM which extracts feature sets from Position Specific Scoring Matrices (PSSMs). PSSMCOOL currently implements more feature extraction techniques than protHMM; however, PSSMCOOL is only able to extract information from PSSMs and thus cannot take advantage of the alignment benefits that HHBlits and HMMer provide over NCBI-BLAST (**xia-2017?**). Furthermore, `protHMM` implements many different novel feature extraction techniques such as Local Binary Pattern and Lyons et al.'s (2016) alignment-based similarity feature.

Dependencies

protHMM relies on the utils package for the reading of HMM profiles into R. protHMM also relies on the phonTools package, the stats package and the gtools package for the calculation feature vectors such as GSD. LPC and SCSH.

References

An, Ji-Yong, Yong Zhou, Yu-Jun Zhao, and Zi-Ji Yan. 2019. "An Efficient Feature Extraction Technique Based on Local Coding PSSM and Multifeatures Fusion for Predicting Protein-Protein Interactions." *Evolutionary Bioinformatics* 15 (January): 117693431987992.

Lyons, James, Abdollah Dehzangi, Rhys Heffernan, Yuedong Yang, Yaoqi Zhou, Alok Sharma, and Kuldip K. Paliwal. 2015. "Advancing the Accuracy of Protein Fold Recognition by Utilizing Profiles From Hidden Markov Models." *IEEE Transactions on Nanobioscience* 14 (7): 761–72. https://doi.org/10.1109/tnb.2015.2457906.

Saini, Harsh, Gaurav Raicar, Alok Sharma, Sunil K. Lal, Abdollah Dehzangi, James Lyons, Kuldip K. Paliwal, Seiya Imoto, and Satoru Miyano. 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 (September): 291–98. https://doi.org/10.1016/j.jtbi.2015.05.030.

Song, Xiaoyu, Zhan-Heng Chen, Xiangyang Sun, Zhu-Hong You, Liping Li, and Yang Zhao. 2018. "An Ensemble Classifier with Random Projection for Predicting Protein–Protein Interactions Using Sequence and Evolutionary Information." *Applied Sciences* 8 (1): 89. https://doi.org/10.3390/app8010089.