PSSMFG: a bioinformatics tool for generating descriptors based on PSSM profiles

Mozhgan Mozaffari legha, Javad Zahiri*, S. Shahriar arab

Department of Biophysics, Faculty of Biological Science, Tarbiat Modares University (TMU), Tehran, Iran
* zahiri@modares.ac.ir

Abstract: One of fundamental step in the construction of machine learning based models is feature extraction. This step is important to determine the effectiveness of trained models in bioinformatics [1]. In the previous two decades, the variety of feature encoding schemes have been proposed in order to exploit suitable patterns from protein sequences. Most of the schemes are based on sequence information or representation of physicochemical properties. Although direct features derived from sequences themselves are regarded as essential for training models, an increasing number of studies have shown that evolutionary information in the form of PSSM profiles is much more informative than sequence information alone [2].

PSSM-based feature descriptors have been widely used to improve the performance of many types of research such as prediction of protein function, membrane proteins types, proteins structural class [3], protein binding site protein [4], secondary structure [5], disordered regions as well as sub-cellular localization of proteins [6], nevertheless, there is no universal tool for generating this wide variety of descriptors. In this study, PSSMFG (Position-Specific Scoring matrix-based feature generator), a tool that can generate a number of types of PSSM-based feature descriptors, is proposed. PSSMFG performances numerous algorithms available in the literature and provides an easy-to-use interface.

Keywords: Feature encoding; Evolutionary information; PSSM.

References

1. Chou, K.-C., Some remarks on protein attribute prediction and pseudo amino acid composition. Journal of theoretical biology, 2011. 273(1): p. 236-247.
2. An, Y., et al., Comprehensive assessment and performance improvement of effector protein predictors for bacterial secretion systems III, IV and VI. Briefings in bioinformatics, 2016: p. bbw100.
3. Liang, Y., S. Liu, and S. Zhang, Prediction of protein structural classes for low-similarity sequences based on consensus sequence and segmented PSSM. Computational and mathematical methods in medicine, 2015. 2015.
4. Yang, X., et al., SNBRFinder: a sequence-based hybrid algorithm for enhanced prediction of nucleic acid-binding residues. PloS one, 2015. 10(7): p. e0133260.
5. Akcesme, F.B. and M. Can, Protein Secondary Structure Prediction by Using PSSM Pseudo Digital Image of Proteins. Southeast Europe Journal of Soft Computing, 2016. 4(2).
6. Mundra, P., et al., Using pseudo amino acid composition to predict protein subnuclear localization: Approached with PSSM. Pattern Recognition Letters, 2007. 28(13): p. 1610-1615.