PyFeat: A Python-based Effective Feature Generation Tool for DNA, RNA, and Protein Sequences

Rafsanjani Muhammod[1, †], Sajid Ahmed[1, †], Dewan Md Farid[1], Swakkhar Shatabda[1,*], Alok Sharma[2,3,4,*], Abdollah Dehzangi[5,*]

[1]Department of Computer Science and Engineering, United International University, Dhaka, Bangladesh, [2]School of Engineering and Physics, University of the South Pacific, Private Mail Bag, Laucala Campus, Suva, Fiji, [3]RIKEN Center for Integrative Medical Sciences, Yokohama, Kanagawa, Japan, [4]Institite for Integrated and Intelligent Systems, Griffith University, Brisbane, Queensland, Australia, [5]Department of Computer Science, Morgan State University, Baltimore, Maryland, USA

[†]: These authors contributed equally to this work. *: Corresponding authors

Abstract:

Motivation: Extracting useful feature set which contains significant discriminatory information is a critical step in effectively presenting sequence data to predict structural, functional, interaction and expression of proteins, DNAs, and RNAs. Also, being able to filter features with significant information and avoid sparsity in the extracted features require the employment of efficient feature selection techniques. Here we present PyFeat as a practical and easy to use toolkit implemented in Python for extracting various features from proteins, DNAs, and RNAs. To build PyFeat we mainly focused on extracting features that capture information about the interaction of neighboring residues to be able to provide more local information. We then employ AdaBoost technique to select features with maximum discriminatory information. In this way, we can significantly reduce the number of extracted features and enable PyFeat to represent the combination of effective features from large neighboring residues. As a result, PyFeat is able to extract features from 13 different techniques and represent context free combination of effective features. The source code for PyFeat standalone toolkit and employed benchmarks with a comprehensive user manual explaining its system and workflow in a step by step manner are publicly available.

Results: https://github.com/mrzResearchArena/PyFeat/blob/master/RESULTS.md

Availability: Toolkit, source code, and manual to use PyFeat: https://github.com/mrzResearchArena/PyFeat/

1 INTRODUCTION

Extracting effective features from sequence data which contains significant discriminatory information is considered the most important step in developing accurate computational methods to predict structural, functional, interaction or expression levels. Most of the methods that have been previously proposed produce a large number of features that in most of the cases do not provide sufficient discriminatory information and at the same time introduce noise and case sparsity in the feature set. Therefore, in addition to feature, extracting informative ones is also a challenging task in computational biology. Over the past decade, several studies have proposed such toolkits that extract different feature sets for different sequence data such as protein and other peptides, RNAs, or DNAs (Cao et al., 2013; Chen et al., 2018; Liu, 2017; Liu et al, 2015; Liu et al., 2017). Despite all the efforts have been made so far, still accessibility to the available methods is limited. Especially toolkits that provide features for prediction task involving proteins, DNAs, and RNAs are crucial to develop due to involvement of data from various platform. Here we propose PyFeat as a comprehensive toolkit implemented in Python for generating various numerical feature presentation schemes from DNA, RNA and protein sequences. As a result, it can be widely used in different applications in bioinformatics and biological science. This tool is also able to select the best features among a large number of features that are generated mainly based on the Gap based feature extraction to provide useful local discriminatory information. Feature generation methods implemented in PyFeat aimed at capturing the frequency distributions of various permutations of the base nucleotides/amino acids in the sequences which in turn are able to represent the sequences in the model training process, efficiently. Here we also demonstrate the effectiveness of features generated by PyFeat to tackle challenging problems using different sequence data (DNAs, RNAs, and Proteins) which for all three cases outperforms previously proposed models found in the literature (Liu, 2017).

2 IMPLEMENTATIONS

K-mer frequency is considered as an important method to extract local features. However, as the K (length of sub-sequences) increases, the number of features dramatical increases as well. Therefore, a small K can only be used to avoid sparsity (Ghandi et al., 2014). For example, for DNA or RNA sequences K less than 5 is used. The case for proteins is even more dramatical as the number of produced features are by far more extensive due to the number of alphabets. To deal with this limitation, we focused on the concept of kGap to implement PyFeat (Cao at al., 2013; Liu et al., 2017a). We have considered gaps in the nucleotide/amino acid sub-sequences and frequency of these sub-sequences are regarded as features for the prediction models. In PyFeat, the number of gaps is an argument that can be adjusted by users. When K is too large in kGap, it can produce large number of features. Therefore, implementing a feature reduction method is necessary to filter out informative and discriminatory features, and discard redundancy and noisy data to avoid sparsity and the curse of dimensionality. At the same time, it should be considered

that discarding informative features may cause in loss of long-range sequence-order information resulting in underfitting or high bias.

Table 1: Features Description

Feature Name	Number of Features	Applicable for
zCurve	3 features for DNA/RNA	DNA/RNA
gcContent	1 feature for DNA/RNA	DNA/RNA
ATGC ratio	1 feature for DNA/RNA	DNA/RNA
Cumulative Skew	2 features for DNA/RNA	DNA/RNA
Chou's Pseudo composition	when --kGap=1,, 84 features for DNA/RNA and 8,420 for protein	DNA/RNA/Protein
monoMonoKGap	when --kGap=1, 16 features for DNA/RNA and 400 for protein	DNA/RNA/Protein
monoDiKGap	when --kGap=1, 64 features for DNA/RNA and 8,000 for protein	DNA/RNA/Protein
monoTriKGap	when --kGap=1, 256 features for DNA/RNA and 160,000 for protein	DNA/RNA/Protein
diMonoKGap	when --kGap=1, 64 features for DNA/RNA and 8,000 for protein	DNA/RNA/Protein
diDiKGap	when --kGap=1, 256 features for DNA/RNA and 160,000 for protein	DNA/RNA/Protein
diTriKGap	when --kGap=1, 1024 features for DNA/RNA and 3,200,000 for protein	DNA/RNA/Protein
triMonoKGap	when --kGap=1, 256 features for DNA/RNA and 160,000 for protein	DNA/RNA/Protein
triDiKGap	when --kGap=1, 1024 features for DNA/RNA and 3,200,000 for protein	DNA/RNA/Protein

In our experiments, we have incorporated features with kGap values ranging from 1 to 5 for DNA and RNA sequences, and values ranging from 1 to 10 for protein sequences. When the kGap value is small, length of the generated feature set is also small, and the occurrence frequency of these features retain local or short-range sequence order information while features with moderately large kGap values preserve global or long-range sequence-order information. Based on this analysis, we extract features from 13 different methods which are listed in Table 1. A full description of these feature extraction methods and the number of produced features is provided in a user-friendly manual that is available online as a supplementary material.

For feature selection and to reduce the impact of the curse of dimensionality and at the same time maintain informative features (Keogh and Mueen, 2017) we have employed AdaBoost classification model to calculate the average impurity-curtailment achieved by splitting upon each of the features in all of the trees trained on different weight distributions of the instances. We then select n features with the maximum score for model training. It is to be noted that this selection mechanism is much more cost-effective compared to the wrapper-based methods since only one run of the AdaBoost model is sufficient for the selection process. Moreover, it is more effective compared to the other methods as different trees incorporate different instance weight distributions into the impurity measure which in turn adds diversity to the way features are selected for node splitting in different trees, thus making the selection process less likely to be adversely affected by the presence correlated features having equally high predictive capability. It makes the choice of features diverse and robust in the presence of high feature multi-collinearity (Wang, 2012). As a result, although the number of extracted features using each method can be very large, PyFeat is able to reduce dimensionality quite dramatically. The explanation of how to use feature extraction, adjusting *K* in KGap technique, and model selection are available online.

3 RESULTS AND DISCUSSIONS

To demonstrate the effectiveness of features extracted using PyFeat, we used three different case studies for DNA, RNA, and protein-based problems. We employed extracted features from PyFeat to predict three different tasks. For DNA, to predict Sigma70 promoter region (Lin et al., 2017); for RNA, to predict adenosine (A) to inosine (I) site which direct the transcription of majority of the genes (Chen et al., 2016); and for proteins, to predict DNA binding proteins (Chowdhury et al., 2017). Extracted features by PyFeat used for different classifiers to build the models. As a result, for all three cases, models built based on features extracted from PyFeat outperformed previous studies found in the literature. For Sigma70 promoter prediction task we achieved 92.88% prediction accuracy, for prediction of Adenosine (A) to Inosine (I) site task we achieved 88.50% prediction accuracy, and for prediction of DNA binding proteins we achieved 83.33% prediction accuracy which for all cases significantly outperform previous results found in the literature (Chen et al., 2016; Liu et al., 2017; Lin et al., 2017; Chowdhury et al., 2017). The full table and description of the problems are provided in supplementary material as well as PyFeat repository in GitHub.

These results demonstrate the effectiveness of extracted features as well as dimensionality reduction scheme provided in PyFeat. To the best of our knowledge, the collection of these feature extractions and dimensionality reduction methods have not been presented in such an integrated toolkit, before. In addition, the conducted experimentation for all three DNA, RNA, and protein-based problems lead to results better than those reported in previous studies (Jani et al., 2018).

In the future, we will integrate more feature extraction, feature reduction, and analysis techniques to expand PyFeat to enable interactive analysis and machine learning-based modeling. PyFeat is expected to be widely used as a powerful tool in bioinformatics, computational biology, and proteome research. PyFeat and its source code and manual are publicly available at: https://github.com/mrzResearchArena/PyFeat/

REFERENCES

Cao, D. S., Xu, Q. S., and Liang. Y. Z. ProPy: a tool to generate various modes of chou's pseaac. *Bioinformatics*, 29(7):960-962, 2013.

Chowdhury, S. Y., Shatabda, S., & Dehzangi, A. Idnaprot-es: Identification of DNA-binding proteins using evolutionary and structural features. *Scientific Reports*, 7(1), 14938, 2017.

Chen, Z., Zhao, P., Li, F., Leier, A., Marquez-Lago, T. T., Wang, Y., et al. iFeature: a python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics*, 1:4, 2018.

Chen, W., Feng, P., Ding, H., and Lin. H., Pai: Predicting adenosine to inosine editing sites by using pseudo nucleotide compositions. *Scientific reports*, 6:35123, 2016.

Ghandi, M., Lee, D., Mohammad-Noori, M., and Beer. M. A. Enhanced regulatory sequence prediction using gapped k-mer features. *PLoS computational biology*, 10(7): e1003711, 2014.

Jani, M.R., Mozlish, M.T.K., Ahmed, S., Tahniat, N.S., Farid, D.M. and Shatabda, S. iRecSpot-EF: Effective sequence based features for recombination hotspot prediction. *Computers in biology and medicine*, 103, pp.17-23, 2018.

Keogh, E, and Mueen, A. Curse of dimensionality. *Encyclopedia of Machine Learning and Data Mining*, 314-315. Springer, 2017.

Liu, B., BioSeq-analysis: a platform for DNA, RNA and Protein sequence analysis based on machine learning approaches. *Briefings in bioinformatics*, 2017.

Liu, B., Liu, F., Wang, X., Chen, J., Fang, L., and Chou, K. C. Pse-in-one: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic acids research*, 43(W1): W65:W71, 2015.

Liu, B., Wu H., Zhang, D., Wang, X., and Chou. K. C. Pse-analysis: a python package for DNA/RNA and protein/peptide sequence analysis based on pseudo components and kernel methods. *Oncotarget*, 8(8):13338, 2017.

Lin, H., Liang, Z. Y., Tang, H., and Chen, W. Identifying sigma70 promoters with novel pseudo nucleotide composition. *IEEE/ACM transactions on computational biology and bioinformatics*, 2017.

Lin, H., Deng, E. Z., Ding, H., Chen, W., and Chou. K. C. ipro54-pseknc: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. *Nucleic acids research*, 42(21):12961-12972, 2014.

Wang, R. AdaBoost for feature selection, classification and its relation with SVM, a review. Physics Procedia, 25, 800-807, 2012.