Bioinformatics, 36(4), 2020, 1277–1278
doi: 10.1093/bioinformatics/btz689
Advance Access Publication Date: 3 September 2019
Applications Note

Sequence analysis

VisFeature: a stand-alone program for visualizing and analyzing statistical features of biological sequences

Jun Wang[1], Pu-Feng Du[1,*], Xin-Yu Xue[1], Guang-Ping Li[1], Yuan-Ke Zhou[1], Wei Zhao[1], Hao Lin[2] and Wei Chen[3,4,*]

[1]College of Intelligence and Computing, Tianjin University, Tianjin 300350, China, [2]Key Laboratory for Neuro-Information of Ministry of Education, School of Life Science and Technology, Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu 610054, China, [3]Innovative Institute of Chinese Medicine and Pharmacy, Chengdu University of Traditional Chinese Medicine, Chengdu 611137, China and [4]Center for Genomics and Computational Biology, School of Life Sciences, North China University of Science and Technology, Tangshan 063000, China

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on July 5, 2019; revised on August 12, 2019; editorial decision on August 28, 2019; accepted on August 30, 2019

Abstract

Summary: Many efforts have been made in developing bioinformatics algorithms to predict functional attributes of genes and proteins from their primary sequences. One challenge in this process is to intuitively analyze and to understand the statistical features that have been selected by heuristic or iterative methods. In this paper, we developed VisFeature, which aims to be a helpful software tool that allows the users to intuitively visualize and analyze statistical features of all types of biological sequence, including DNA, RNA and proteins. VisFeature also integrates sequence data retrieval, multiple sequence alignments and statistical feature generation functions.

Availability and implementation: VisFeature is a desktop application that is implemented using JavaScript/Electron and R. The source codes of VisFeature are freely accessible from the GitHub repository (https://github.com/wangjun1996/VisFeature). The binary release, which includes an example dataset, can be freely downloaded from the same GitHub repository (https://github.com/wangjun1996/VisFeature/releases).

Contact: pdu@tju.edu.cn or chenweiimu@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Over the last two decades, the number of known biological sequences has been growing exponentially. It is urgent to understand their functional attributes. Many efficient computational methods for generating statistical features from sequences have been developed. Several web servers and stand-alone programs have been released for practical applications, such as PseAAC (Shen and Chou, 2008), PseAAC-General (Du *et al.*, 2014), PseKNC-General (Chen *et al.*, 2015), Pse-in-One (Liu *et al.*, 2015) and UltraPse (Du *et al.*, 2017). These software tools provide efficient and convenient solutions to generate statistical features for biological sequences. However, a helpful software tool for visualizing the statistical features is still lacking. Although existing programs can display nucleotide sequences using dinucleotide property curves in a genome browser style (Friedel *et al.*, 2009), the abilities to visualize protein sequences and to intuitively compare statistical features are still missing.

To this end, we developed VisFeature, which is an open-source stand-alone program that can visualize and analyze various types of statistical features of all types of biological sequences, including DNA, RNA and proteins. To the best of our knowledge, this is the first toolkit that is designed especially for this purpose. VisFeature integrates sequence feature visualization and analysis, statistical feature visualization and comparison, online database querying, multiple sequence alignment and color sequence visualization together. All these functions are useful in the explorative stage of developing predictive algorithms for functional attributes.

2 Implementations

VisFeature is mainly implemented by using JavaScript, with the Electron framework. The remaining part of VisFeature is implemented by R scripts.

The input of VisFeature is a FASTA format file. This file can be chosen directly from the local computer. With an internet connection, VisFeature is capable of querying the UniProt database or NCBI databases using sequence identifiers or query expressions. The sequences in the query results can be saved as a FASTA file. This is the second way to obtain a FASTA file in VisFeature.

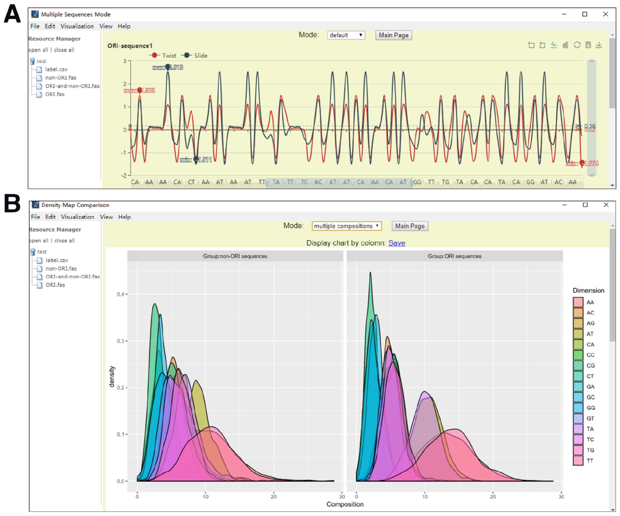

Fig. 1. Visualization modules. (**A**) Physicochemical property values of sequences: the vertical axis is the value of physicochemical properties, while the horizontal axis is the position on the sequence. Since the sequence is zoomed out at a certain scale, the labels on the horizontal axis are not for continuous dinucleotides on the sequence. (**B**) The distribution comparison for di-nucleotide compositions between different groups. Different colors show different types of di-nucleotides

VisFeature offers two modules for feature visualization (Fig. 1).

(1) The module for visualizing physicochemical properties as curves: It is implemented by Echarts (Li *et al.*, 2018). This module provides many interactive visualization functions. For example, it can generate curves or bars for simple visualization, or perform multiple sequence alignment by calling the clustalw program (Larkin *et al.*, 2007). Users are allowed to zoom-in and zoom-out by rolling the mouse wheel. The vertical and horizontal sidebars can be dragged to change the level of details. Many physicochemical property values are integrated within this module, including 566 physiochemical properties for amino acids, 148 for dinucleotides on DNA, 22 for dinucleotides on RNA and 12 for trinucleotides on DNA, which are all collected from literatures (Chen *et al.*, 2015; Kawashima *et al.*, 2008). The differences of physicochemical properties among sequences and their trends along the sequence can be visualized in this module.

(2) The module for visualizing and comparing feature vectors as density maps. After generating statistical features for every sequence using the 'Density Map Comparison' function, a label file should be uploaded to assign group labels to each sequence. Two different modes are implemented in this module. One is called the 'single composition' mode, while the other is called the 'multiple composition' mode. In the 'single composition' mode, when the statistical features are generated, VisFeature computes the density maps of the features on each dimension. In the 'multiple composition' mode, VisFeature generates density maps for each dimension in each group separately. The density maps of different dimensions in the same group are stacked together using transparent figures. This allows the users to compare distributions of features in different groups, which is very useful in exploring informative sequence features to predict functional attributes of biological sequences.

For the users to understand the VisFeature functions quickly, we not only provided a screen recording (Supplementary Video S1) in Supplementary Information, but also included an example dataset in the binary version of VisFeature. This dataset contains 811 sequences, which are obtained from the iORI-PseKNC study (Li *et al.*, 2015). A set of slides is provided as Appendix 1, which is a simple step-by-step guide to experience VisFeature with the example dataset.

3 Conclusion

VisFeature is an open-source stand-alone program for visualizing and analyzing statistical features of all types of biological sequences. It provides an intuitive way to explore the trends of physicochemical properties along the sequences. It can visualize the differences of feature distributions between different sequence groups. These functions make VisFeature a helpful tool in developing predictive algorithms for functional attributes. As far as we know, VisFeature is the first software tool that integrates sequence retrieval, alignments and feature generation, visualization and distribution comparison together for all types of biological sequences.

Funding

This work was supported by National Key R&D Program of China [2018YFC0910405]; National Natural Science Foundation of China [NSFC 61872268 and 31771471]; Natural Science Foundation for Distinguished Young Scholar of Hebei Province [No. C2017209244] and Open Project Funding of CAS Key Lab of Network Data Science and Technology, Institute of Computing Technology, Chinese Academy of Sciences [CASNDST201705].

Conflict of Interest: none declared.

References

Chen,W. *et al.* (2015) PseKNC-General: a cross-platform package for generating various modes of pseudo nucleotide compositions. *Bioinformatics*, **31**, 119–120.

Du,P. *et al.* (2014) PseAAC-General: fast building various modes of general form of Chou's pseudo-amino acid composition for large-scale protein datasets. *Int. J. Mol. Sci.*, **15**, 3495–3506.

Du,P.-F. *et al.* (2017) UltraPse: a universal and extensible software platform for representing biological sequences. *Int. J. Mol. Sci.*, **18**, 2400.

Friedel,M. *et al.* (2009) DiProGB: the dinucleotide properties genome browser. *Bioinformatics*, **25**, 2603–2604.

Kawashima,S. *et al.* (2008) AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res.*, **36**, D202–205.

Larkin,M.A. *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.

Li,D. *et al.* (2018) ECharts: a declarative framework for rapid construction of web-based visualization. *Vis. Inf.*, **2**, 136–146.

Li,W.-C. *et al.* (2015) iORI-PseKNC: a predictor for identifying origin of replication with pseudo k-tuple nucleotide composition. *Chemometr. Intell. Lab. Syst.*, **141**, 100–106.

Liu,B. *et al.* (2015) Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Res.*, **43**, W65–71.

Shen,H.-B. and Chou,K.-C. (2008) PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition. *Anal. Biochem.*, **373**, 386–388.