BIOINFORMATICS APPLICATIONS NOTE Vol. 31 no. 1 2015, pages 119–120
doi:10.1093/bioinformatics/btu602

Sequence analysis

Advance Access publication September 16, 2014

PseKNC-General: a cross-platform package for generating various modes of pseudo nucleotide compositions

Wei Chen[1,2,3,*], Xitong Zhang[4], Jordan Brooker[5], Hao Lin[3,6], Liqing Zhang[2,*] and Kuo-Chen Chou[1,3,7,*]

[1]Department of Physics, School of Sciences, Center for Genomics and Computational Biology, Hebei United University, Tangshan 063009, China, [2]Department of Computer Science, Virginia Tech, Blacksburg, VA 24060, [3]School of Life Science and Technology, Bioinformatics and Computer-Aided Drug Discovery, Gordon Life Science Institute, Boston, MA 02478, [4]Department of Electrical and Computer Engineering, University of Virginia, Charlottesville, VA 22904, [5]Department of Computer Science, Vassar College, Poughkeepsie, NY 12604, USA, [6]Excellence in Genomic Medicine Research, Key Laboratory for Neuro-Information of Ministry of Education, Center of Bioinformatics, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, China and [7]Excellence in Genomic Medicine Research, Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah 21589, Saudi Arabia

Associate Editor: John Hancock

ABSTRACT

Summary: The avalanche of genomic sequences generated in the post-genomic age requires efficient computational methods for rapidly and accurately identifying biological features from sequence information. Towards this goal, we developed a freely available and open-source package, called PseKNC-General (the general form of pseudo k-tuple nucleotide composition), that allows for fast and accurate computation of all the widely used nucleotide structural and physicochemical properties of both DNA and RNA sequences. PseKNC-General can generate several modes of pseudo nucleotide compositions, including conventional k-tuple nucleotide compositions, Moreau–Broto autocorrelation coefficient, Moran autocorrelation coefficient, Geary autocorrelation coefficient, Type I PseKNC and Type II PseKNC. In every mode, >100 physicochemical properties are available for choosing. Moreover, it is flexible enough to allow the users to calculate PseKNC with user-defined properties. The package can be run on Linux, Mac and Windows systems and also provides a graphical user interface.

Availability and implementation: The package is freely available at: http://lin.uestc.edu.cn/server/pseknc.

Contact: chenweiimu@gmail.com or lqzhang@vt.edu or kcchou@gordonlifescience.org.

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

Received on July 22, 2014; revised on August 19, 2014; accepted on August 31, 2014

1 INTRODUCTION

The pseudo nucleotide composition was proposed in 2011 (Zhou *et al.*, 2011), and its basic algorithm was elaborated in a recent study (Chen *et al.*, 2014a). Compared with the conventional nucleotide composition, pseudo nucleotide compositions have the advantage of converting DNA sequences of various lengths to a fixed-length digital vector to enable sequence comparison, while at the same time keeping the long-range sequence order information. Since its introduction, pseudo nucleotide compositions have been applied in many branches of computational genomics, such as predicting promoters (Zhou *et al.*, 2013), predicting recombination spots (Chen *et al.*, 2013; Guo *et al.*, 2012; Qiu *et al.*, 2014), predicting nucleosome positioning sequences (Guo *et al.*, 2014), predicting DNA methylation status (Zhou *et al.*, 2011), predicting splice sites (Chen *et al.*, 2014b), identifying translation initiation site (Chen *et al.*, 2014c) and so on.

Currently, the only available tool for generating pseudo nucleotide compositions is the recently developed online web server PseKNC (Chen *et al.*, 2014a). However, PseKNC can process at most only 100 sequences in one submission, and also is limited to only DNA sequences and a small number of physicochemical features. No software is available for processing large-scale datasets with the flexibility of adjusting algorithm parameters used in the calculation. This fact decreases the efficiency of the related studies.

In the present work, we provide a cross-platform stand-alone and open-source package, called **PseKNC-General**, which could convert large-scale sequence datasets to pseudo nucleotide compositions with numerous choices of physicochemical property combinations. It can not only deal with DNA but also with RNA sequences. Moreover, a graphical user interface (GUI) shell program is also provided along with the command-line version. To our knowledge, **PseKNC-General** is the first open-source package that can encode a large number of genomic sequences based on user-defined physicochemical properties.

2 PACKAGE DESCRIPTION

The current PseKNC-General package can generate a large number of features as summarized in Table 1. These features can be divided into three groups. The first group includes six features: nucleotide composition (1-tuple), dinucleotide composition (2-tuple), trinucleotide composition (3-tuple), tetranucleotide composition (4-tuple), pentanucleotide composition (5-tuple)

*To whom correspondence should be addressed.

Table 1. List of various PseKNC modes

Feature group	Features
K-tuple nucleotide composition	Nucleotide composition
	Dinucleotide composition
	Trinucleotide composition
	Tetranucleotide composition
	Pentanucleotide composition
	Hexanucleotide composition
Autocorrelation	Normalized Moreau–Broto autocorrelation
	Moran autocorrelation
	Geary autocorrelation
PseKNC	Type 1 PseKNC
	Type 2 PseKNC

and hexanucleotide composition (6-tuple). The second group contains three feature sets: Moreau-Broto autocorrelation coefficient (Feng and Zhang, 2000), Moran autocorrelation coefficient (Horne, 1988) and Geary autocorrelation coefficient (Sokal and Thomson, 2006). These three autocorrelation features describe the level of correlation between two k-tuple nucleotides in terms of their specific structural and/or physicochemical properties. The third group consists of two types of pseudo k-tuple nucleotide compositions: Type I PseKNC and Type II PseKNC.

Compared with the **PseKNC** server, **PseKNC-General** has the following major advantages and improvements.

First, **PseKNC-General** is a stand-alone program and can be run on local computers. It does not require uploading sequences to the server and hence can process large-scale datasets with no data-size limit. Therefore, it is faster and easier to generate the pseudo k-tuple compositions and to adjust the parameters.

Second, **PseKNC-General** can be used for not only DNA sequences but also RNA sequences.

Third, **PseKNC-General** incorporates additional structural and physicochemical properties (Supplementary Material S1), and more conveniently allows users to add or specify their own structural and physicochemical properties. It is thus both more flexible and more expandable.

Fourth, a user-friendly GUI is also provided along with the command-line program. Every parameter in PseKNC-General can be easily configured from the GUI. Hence, users can comfortably choose the mode that they prefer to run the program. All the source code together with the detailed illustration documents of the package can be freely downloaded from http://lin.uestc.edu.cn/server/pseknc.

Finally, results can be saved in three different file formats: the LibSVM format, the CSV format and the tab-delimited format. All these formats are suitable for downstream computational analyses, such as machine learning.

3 CONCLUSIONS

The continued expansion of genomic sequences necessitates the development of computational tools to annotate functional elements from DNA or RNA sequences. One of the important tasks is to formulate the genomic sequences with an effective expression form that can reflect the intrinsic correlation with their structures and functions. PseKNC-General makes this task readily achievable for any expert and/or non-expert users via its highly flexible, configurable and user-friendly design. It allows for fast and accurate computation of a broad range of physicochemical properties of k-tuple nucleotide in DNA/RNA sequences. The effectiveness and usefulness of these properties have been demonstrated by a series of existing works. Hence, we anticipate that **PseKNC-General** will become a useful package in exploring problems concerning computational genomics and genome sequence analysis. In the future, we will collect more experimental physicochemical properties for other k-tuple nucleotides (k = 4, 5 or higher) and integrate them into the current version of PseKNC-General to further enhance its power.

ACKNOWLEDGEMENTS

The authors wish to thank the three anonymous reviewers for their constructive comments, which were helpful for strengthening the presentation of this study.

Funding: This work was supported by the National Nature Science Foundation of China [61100092 to W.C., 61202256 to H.L.], the Nature Scientific Foundation of Hebei Province [C2013209105 to W.C.] and the National Science Foundation [NSF OCI-1124123 to L.Z.].

Conflict of interest: none declared.

REFERENCES

Chen,W. *et al.* (2013) iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. *Nucleic Acids Res.*, **41**, e68.

Chen,W. *et al.* (2014a) iSS-PseDNC: identifying splicing sites using pseudo dinucleotide composition. *BioMed Res. Int.*, **2014**, 623149.

Chen,W. *et al.* (2014b) PseKNC: a flexible web server for generating pseudo K-tuple nucleotide composition. *Anal. Biochem.*, **456**, 53–60.

Chen,W. *et al.* (2014c) iTIS-PseTNC: A sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition. *Anal. Biochem.*, **462**, 76–83.

Feng,Z.P. and Zhang,C.T. (2000) Prediction of membrane protein types based on the hydrophobic index of amino acids. *J. Protein Chem.*, **19**, 269–275.

Guo,S.H. *et al.* (2012) Recombination spots prediction using DNA physical properties in the saccharomyces cerevisiae genome. *AIP Conf. Proc.*, **1479**, 1556.

Guo,S.H. *et al.* (2014) iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. *Bioinformatics*, **30**, 1522–1529.

Horne,D.S. (1988) Prediction of protein helix content from an autocorrelation analysis of sequence hydrophobicities. *Biopolymers*, **27**, 451–477.

Qiu,W.R. (2014) iRSpot-TNCPseAAC: identify recombination spots with trinucleotide composition and pseudo amino acid components. *Int. J. Mol. Sci.*, **15**, 1746–1766.

Sokal,R.R. and Thomson,B.A. (2006) Population structure inferred by local spatial autocorrelation: an example from an Amerindian tribal population. *Am. J. Phys. Anthropol.*, **129**, 121–131.

Zhou,X. *et al.* (2011) Predicting methylation status of human DNA sequences by pseudo-trinucleotide composition. *Talanta*, **85**, 1143–1147.

Zhou,X. *et al.* (2013) Predicting promoters by pseudo-trinucleotide compositions based on discrete wavelets transform. *J. Theor. Biol.*, **319**, 1–7.