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

Wei Chen1,2,3,*, Xitong Zhang4, Jordan Brooker5, Hao Lin3,6, Liqing Zhang2,* and Kuo-Chen Chou1,3,7,*

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

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)
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., 6120256 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