Bioinformatics Advance Access published August 26, 2016

Sequence analysis

PseKRAAC: A flexible web server for generating pseudo K-tuple reduced amino acids composition

Yongchun Zuo[1, *†], Yuan Li[1,2†], Yingli Chen[3], Guangpeng Li[1], Zhenhe Yan[1,2*], and Lei Yang[4,*]

[1]The Key Laboratory of Mammalian Reproductive Biology and Biotechnology of the Ministry of Education, College of life sciences, Inner Mongolia University, Hohhot, 010021, China, [2] Department of Mechanical Engineering, Columbia University, New York, 10027, USA, [3]School of Physical Science and Technology, Inner Mongolia University, Hohhot, 010021, China, [4] College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, 150081, China.

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Associate Editor: Dr. John Hancock

Abstract

Summary: The reduced amino acids perform powerful ability for both simplifying protein complexity and identifying functional conserved regions. However, dealing with different protein problems may need different kinds of cluster methods. Encouraged by the success of pseudo-amino acid composition (PseAAC) algorithm, we developed a freely available web server, called PseKRAAC (The pseudo K-tuple reduced amino acids composition). By implementing reduced amino acid alphabets, the protein complexity can be significantly simplified, which leads to decrease chance of overfitting, lower computational handicap, and reduce information redundancy. PseKRAAC delivers more capability for protein research by incorporating three crucial parameters that describes protein composition. Users can easily generate many different modes of PseKRAAC tailored to their needs by selecting various reduced amino acids alphabets and other characteristic parameters. It is anticipated that the PseKRAAC web server will become a very useful tool in computational proteomics and protein sequence analysis.

Availability and Implementation: Freely available on the web at http://bigdata.imu.edu.cn/psekraac

Contact: yczuo@imu.edu.cn or imu.hema@foxmail.com or yanglei_hmu@163.com.

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

1 Introduction

With the emergence of big data in the post-genomic age, an enormous amount of data had been generated, which requires efficient computational methods for rapid and effective identification of biological features contained in sequences(Chen, et al., 2015) . Even more so in the study of proteomics, because of the structure of protein exhibits more complexity when compared to nucleotide, due to the possible 20 amino acid peptides to the 4 nucleic bases. Therefore, the complexity and information content are expanded exponentially when polypeptides are formed. Also, protein sequence vary widely in length, which poses additional difficulty for incorporating the sequence-order information consistently in both dataset construction and algorithm formulation.

To overcome these obstacles, the pseudo-amino acid composition (PseAAC) algorithm was proposed in the year 2001(Chou, 2001). This concept has been widely utilized in nearly all areas of computational proteomics, and was selected as one of the key topics in "Molecular Science for Drug Development and Biomedicine"(Zhong and Zhou, 2014). Encouraged by the success of this idea, various approaches similar to PseAAC had been simulated to deal with problems in protein and protein-related systems, including three powerful software programs for generating different modes of PseAAC: PseAAC-Builder(Du, et al., 2012), propy(Cao, et al., 2013), PseAAC-General(Du, et al., 2014**)** and Pse-in-One(Liu, et al., 2015).

When dealing with extremely large dimensions can potentially cause overfitting or high-dimension disaster (Wang, et al., 2016), restrict by computation handicap, and increase information redundancy, which results in bad prediction accuracy. To solve this problem, we present a convenient approaches based on the idea of pseudo reduced amino acid composition (PseRAAC), and provide a flexible and user-friendly web server for pseudo K-tuple reduced amino acids composition (PseKRAAC) (http://bigdata.imu.edu.cn/ psekraac), where users can easily generate many different modes of PseKRAAC tailored to their needs by selecting various reduced amino acids alphabets and other characteristic parameters.

2 Reduced amino acids alphabets

Based on physicochemical features or evolutionary relationships, amino acids residues can be clustered into groups because they serve similar structural or functional roles in proteins(Wang and Wang, 1999). The reduced amino acids not only simplify the complexity of the protein system, but also improve the ability in finding structurally conserved regions and the structural similarity of entire proteins(Peterson, et al., 2009). In recent years, the alphabet reduction techniques play high potential roles for enhancing the power in dealing with protein sequence analysis(Supplementary Data)(Feng, et al., 2013; Liu, et al., 2015; Liu, et al., 2014). Therefore, it is reasonable to use the reduced amino acids alphabets to formulate PseKRAAC for protein sequences. Here, 16

types of reduced amino acid alphabets were proposed to generate various different modes of PseRAAC(Table 1)(Liu, et al., 2015).

Table 1. List of 16 types of reduced amino acid alphabets in protein.

Type	Method description	Clusters	Dimension
1	RedPSSM	2-19	$RAAC^K$
2	BLOSUM 62 matrix	2-6,8,15	$RAAC^K$
3A	PAM matrix	2-19	$RAAC^K$
3B	WAG matrix	2-19	$RAAC^K$
4	Protein Blocks	5,8,9,11,13	$RAAC^K$
5	BLOSUM50 matrix	3,4,8,10,15	$RAAC^K$
6	Multiple cluster	4,5a,5b,5c	$RAAC^K$
7	Metric multi-dimensional scaling (MMDS)	2-19	$RAAC^K$
8	Grantham Distance Matrix(Saturation)	2-19	$RAAC^K$
9	Grantham Distance Matrix (Grantham)	2-19	$RAAC^K$
10	BLOSUM matrix for SWISS-PROT	2-19	$RAAC^K$
11	BLOSUM matrix for SWISS-PROT	2-19	$RAAC^K$
12	BLOSUM matrix for DAPS	2-18	$RAAC^K$
13	Coarse-graining substitution matrices	4,12,17	$RAAC^K$
14	Alphabet Simplifer	2-19	$RAAC^K$
15	MJ matrix	2-16	$RAAC^K$
16	BLOSUM50 matrix	2-16	$RAAC^K$

$RAAC^K$: K-tuple of reduced amino acid cluster (RAAC), For example, Type 1, Cluster =10(RAAC) and K-tuple=2 (K=2), Dimension= $RAAC^K=10^2=100$.

3 Reduced amino acid composition

The proposed server can generate two different types of PseKRAAC for protein sequences analysis: I. g-gap and II. λ−correlation PseKRAAC(Chou, 2001). Suppose a protein sequence P with L amino acid residues as follows,

$$P = R_1R_2R_3R_4R_5 \ldots R_{L-3}R_{L-2}R_{L-1}R_L$$

where R_1 represents the amino acid residue at the sequence position 1, R_2 represents the amino acid residue at position 2 and so on. For each K-tuple of reduced amino acid cluster (RAAC), the feature vector of the protein sequence contains $RAAC^K$ dimensions.

I. g-gap PseKRAAC

The g-gap PseKRAAC is used to represent a protein sequence with a vector containing $RAAC^K$ components, where g represents the gap between each K-tuple peptides(Liu, et al., 2015; Wang, et al., 2016). A g-gap of *n* reflects the sequence-order information for all dipeptides with the starting residues separated by *n* residues. Supplementary Figure 1A shows the schematic drawing of g-gap definition of dipeptide (K=2).

II. λ-correlation PseKRAAC

The λ-correlation PseKRAAC, also called parallel correlation PseKRAAC, is used to represent a protein sequence with a vector containing $RAAC^K$ components, where λ is an integer that represents the correlation tier, and is less than L – K. The n^{th}-tier correlation factor (λ=n) reflects the sequence-order correlation between n^{th} most nearest residue. Supplementary Figure 1B shows the schematic drawing of λ-correlation definition of dipeptide (K=2).

4 Server description

A step-by-step server guide on how to use PseKRAAC can refer to the Supplementary Data. Compared to the original Chou's PseAAC server, PseKRAAC server offers following important improvements and advantages: First, by implementing the concept of reduced amino acid alphabet for amino acid clustering, the complexity of protein composition is significantly simplified, which leads to decrease chance of overfitting, lower computational handicap, and reduce information redundancy.

Also, PseKRAAC delivers more capability for protein research by incorporating three crucial parameters that describes protein composition: K-tuple peptide functionality for K up to 3, λ-correlation PseKRAAC, and g-gap PseKRAAC for protein characterization. Users can increase the λ value in PseAAC webserver to cover more global sequence-pattern effects or increase the K value to count more local sequence-pattern effects. Finally, PseKRAAC provides easier application for inputting sequences by accepting protein sequences in FASTA format via directly enter into the input text box or upload it as a FASTA file. The server is also capable of outputting result files in LIBSVM, CSV and FASTA format for further analysis. When uploading FASTA outputting files to other PseAAC webservers, the user can easily generating more various modes of PseAAC(Chou, 2005).

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 Scientific Foundation of China (No: 61561036, 31501078) and the Specialized Research Fund for the Doctoral Program of Higher Education (20131501120009).

Conflict of Interest: none declared.

References

Cao, D.S., Xu, Q.S. and Liang, Y.Z. propy: a tool to generate various modes of Chou's PseAAC. *Bioinformatics (Oxford, England)* 2013;29(7):960-962.

Chen, W., Lin, H. and Chou, K.C. Pseudo nucleotide composition or PseKNC: an effective formulation for analyzing genomic sequences. *Molecular bioSystems* 2015;11(10):2620-2634.

Chou, K.C. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins* 2001;43(3):246-255.

Chou, K.C. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics (Oxford, England)* 2005;21(1):10-19.

Du, P., Gu, S. and Jiao, Y. PseAAC-General: fast building various modes of general form of Chou's pseudo-amino acid composition for large-scale protein datasets. *International journal of molecular sciences* 2014;15(3):3495-3506.

Du, P., *et al.* PseAAC-Builder: a cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions. *Analytical biochemistry* 2012;425(2):117-119.

Feng, P.M., *et al.* iHSP-PseRAAAC: Identifying the heat shock protein families using pseudo reduced amino acid alphabet composition. *Analytical biochemistry* 2013;442(1):118-125.

Liu, B., *et al.* Identification of microRNA precursor with the degenerate K-tuple or Kmer strategy. *Journal of theoretical biology* 2015;385:153-159.

Liu, B., *et al.* repDNA: a Python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effects. *Bioinformatics (Oxford, England)* 2015;31(8):1307-1309.

Liu, B., *et al.* Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic acids research* 2015;43(W1):W65-71.

Liu, B., *et al.* iDNA-Prot|dis: identifying DNA-binding proteins by incorporating amino acid distance-pairs and reduced alphabet profile into the general pseudo amino acid composition. *PloS one* 2014;9(9):e106691.

Peterson, E.L., *et al.* Reduced amino acid alphabets exhibit an improved sensitivity and selectivity in fold assignment. *Bioinformatics (Oxford, England)* 2009;25(11):1356-1362.

Wang, J. and Wang, W. A computational approach to simplifying the protein folding alphabet. *Nature structural biology* 1999;6(11):1033-1038.

Wang, R., Xu, Y. and Liu, B. Recombination spot identification Based on gapped k-mers. *Scientific reports* 2016;6:23934.

Zhong, W.Z. and Zhou, S.F. Molecular science for drug development and biomedicine. *International journal of molecular sciences* 2014;15(11):20072-20078.