Pse-in-One 2.0: An Improved Package of Web Servers for Generating Various Modes of Pseudo Components of DNA, RNA, and Protein Sequences

Bin Liu1,2,3, Hao Wu1, Kuo-Chen Chou3,4,5

1School of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, China; 2Key Laboratory of Network Oriented Intelligent Computation, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, China; 3Gordon Life Science Institute, Boston, Massachusetts, USA; 4Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu, China; 5Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah, KSA

Correspondence to: Bin Liu, bliu@hit.edu.cn; Kuo-Chen Chou, kcchou@gordonlifescience.org

Keywords: Pseudo Components, DNA Sequences, RNA Sequences, Protein Sequences

Received: April 10, 2017 Accepted: April 25, 2017 Published: April 28, 2017

Copyright © 2017 by authors and Scientific Research Publishing Inc. This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

http://creativecommons.org/licenses/by/4.0/

ABSTRACT

Pse-in-One 2.0 is a package of web-servers evolved from Pse-in-One (Liu, B., Liu, F., Wang, X., Chen, J. Fang, L. & Chou, K.C. Nucleic Acids Research, 2015, 43:W65-W71). In order to make it more flexible and comprehensive as suggested by many users, the updated package has incorporated 23 new pseudo component modes as well as a series of new feature analysis approaches. It is available at http://bioinformatics.hitsz.edu.cn/Pse-in-One2.0/. Moreover, to maximize the convenience of users, provided is also the stand-alone version called “Pse-in-One-Analysis”, by which users can significantly speed up the analysis of massive sequences.

1. INTRODUCTION

With the avalanche of biological sequences generated in the post-genomic age, one of the most challenging problems in computational biology today is how to effectively formulate the sequence of a biological sample (such as DNA, RNA or protein) with a discrete model or a vector that can effectively reflect its sequence pattern information or capture its key features concerned. This is because almost all the existing machine-learning algorithms, such as “Neural Network” or NN algorithm [1-3] “Support Vector Machine” or SVM algorithm [4-12] “Nearest Neighbor” or NN algorithm [13, 14] and “Random Forest” algorithm [15-22] can only handle vectors but not sequence samples as elucidated in a review paper [23]. Unfortunately, if using the sequential model, i.e., the model in which all the samples are represented by their original sequences, it is hardly able to train a machine learning model that can cover all the possible cases concerned, as elaborated in [24].

To avoid completely losing the sequence-order information for proteins, the quasi-sequence-order
approach [25] or PseAAC (pseudo amino acid composition) approach [26-28] or Chou’s PseAAC [29-32] was proposed. Ever since the concept of Chou’s PseAAC [29-31] was proposed, it has rapidly penetrated into many biomedicine and drug development areas [33, 34] and nearly all the areas of computational proteomics [35-212].

Because it has been widely and increasingly used, and also because it would be a trend and future direction to establish user-friendly and publically accessible web-servers for various analysis methods as pointed out in [213], four powerful web-servers were established; they are “PseAAC” [214], “PseAAC-Builder” [29], “propy” [30] and “PseAAC-General” [32]. The former three are for generating various modes of Chou’s special PseAAC; while the 4th one for those of Chou’s general PseAAC [28, 108] including not only all the special modes of feature vectors for proteins but also the higher level feature vectors such as “Functional Domain” mode (see Eqs.9-10 of [108], “Gene Ontology” mode (see Eqs.11-12 of 108), and “Sequential Evolution” or “PSSM” mode (see Eqs.13-14 of [108]).

Encouraged by the successes of using PseAAC to deal with protein/peptide sequences, the concept of PseAAC has been extended to cover DNA/RNA sequences as well via the PseKNC (Pseudo K-tuple Nucleotide Composition) approach [215-223]. Meanwhile, four publically accessible web-servers [215, 217, 224, 225] were developed for generating various pseudo components or feature vectors for DNA/RNA sequences as well.

Particularly, recently a very powerful web-server called Pse-in-One [226] has been established that can be used to generate any desired pseudo components or feature vectors for protein/peptide and DNA/RNA sequences according to the need of users’ studies.

Since then some novel pseudo component modes have been proposed for dealing with various problems in proteomics and genome analysis [7, 10-12, 18-22, 221, 227-263]. In order to incorporate these new and important developments into the Pse-in-One package, an updated version called “Pse-in-One 2.0” has been established.

2. RESULTS AND DISCUSSION

Compared with the original one, the updated version has the following new features and functions.

2.1. Modes of Pseudo Components

Added in are a total of 23 new pseudo component modes, of which 6 for DNA sequences (Table 1), 8 for RNA sequences (Table 2), and 9 for protein sequences (Table 3). These new modes reflect the recent developments of the pseudo components, particularly in extending the coverage scope to those features derived from (1) RNA secondary structures, and (2) the multiple sequence alignments and profiles. As a consequence, Pse-in-One 2.0 covers a total of 51 different features, of which 20 for DNA sequences, 14 for RNA sequences, and 17 for protein sequences. The overall structure can be reflected via the following three sub web-servers.

**PseDAC-General** is for generating the feature vectors of DNA sequences. It contains three categories: nucleotide composition, nucleotide autocorrelation, and pseudo nucleotide composition. Of the 6 new modes, 3 are added into the first category, including IDKmer [224], Mismatch [264], and Subsequence[265]; while the other 3 are added to the second category, including Moran autocorrelation, Geary autocorrelation, and Normalized Moreau-Broto autocorrelation [217].

**PseRAC-General** is aimed to generate the feature vectors for RNA sequences, and it has four categories, of which the “predicted structure composition” is a newly added category for extracting the structure-based features of RNA sequences, in which the following 3 new modes are incorporated: Triplet [266], PseSSC [267] and PseDPC [10]. Triplet is an early approach to use the structure information of RNA sequences and has shown better performance for microRNA identification in comparison with other sequence-based approaches. PseSSC and PseDPC can be used to incorporate the global or long-range structure-order information so as to remarkably improve the prediction quality in identifying the pre-miRNAs. Of the other 5 new modes, 2 are added into the nucleic acid composition category, i.e.,
Table 1. List of the 6 new modes for DNA sequences.

<table>
<thead>
<tr>
<th>Category</th>
<th>Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nucleic acid composition</td>
<td>1) increment of diversity (IDKmer) [226, 270, 291]</td>
</tr>
<tr>
<td></td>
<td>2) The occurrences of kmers, allowing at most m mismatches (Mismatch) [264, 265, 292]</td>
</tr>
<tr>
<td></td>
<td>3) The occurrences of kmers, allowing non-contiguous matches (Subsequence) [265, 292, 293]</td>
</tr>
<tr>
<td></td>
<td>4) Moran autocorrelation (MAC) [268, 294]</td>
</tr>
<tr>
<td>Autocorrelation</td>
<td>5) Geary autocorrelation (GAC) [217, 295]</td>
</tr>
<tr>
<td></td>
<td>6) Normalized Moreau-Broto autocorrelation (NMBAC) [217, 296]</td>
</tr>
</tbody>
</table>

Table 2. List of the 8 new modes for RNA sequences.

<table>
<thead>
<tr>
<th>Category</th>
<th>Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nucleic acid composition</td>
<td>1) The occurrences of kmers, allowing at most m mismatches (Mismatch) [264, 265, 292]</td>
</tr>
<tr>
<td></td>
<td>2) The occurrences of kmers, allowing non-contiguous matches (Subsequence) [265, 292, 293]</td>
</tr>
<tr>
<td></td>
<td>3) Moran autocorrelation (MAC) [217, 294]</td>
</tr>
<tr>
<td>Autocorrelation</td>
<td>4) Geary autocorrelation (GAC) [217, 295]</td>
</tr>
<tr>
<td>Predicted structure composition</td>
<td>5) Normalized Moreau-Broto autocorrelation (NMBAC) [217, 296]</td>
</tr>
<tr>
<td></td>
<td>6) Local structure-sequence triplet element (Triplet) [266]</td>
</tr>
<tr>
<td></td>
<td>7) Pseudo-structure status composition (PseSSC) [226]</td>
</tr>
<tr>
<td></td>
<td>8) Pseudo-distance structure status pair composition (PseDPC) [10]</td>
</tr>
</tbody>
</table>

Table 3. List of the 8 new modes for protein sequences.

<table>
<thead>
<tr>
<th>Category</th>
<th>Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amino acid composition</td>
<td>1) Distance-based Residue (DR) [271]</td>
</tr>
<tr>
<td></td>
<td>2) PseAAC of Distance-Pairs and Reduced Alphabet (Distance Pair) [271]</td>
</tr>
<tr>
<td>Autocorrelation</td>
<td>3) Physicochemical distance transformation (PDT) [270]</td>
</tr>
<tr>
<td></td>
<td>4) Select and combine the n most frequenct amino acids according to their frequencies (Top-n-gram) [269]</td>
</tr>
<tr>
<td>Profile-based features</td>
<td>5) Profile-based Physicochemical distance transformation (PDT-Pofile) [270]</td>
</tr>
<tr>
<td></td>
<td>6) Distance-based Top-n-gram (DT) [271]</td>
</tr>
<tr>
<td></td>
<td>7) Profile-based Auto covariance (AC-PSSM) [272]</td>
</tr>
<tr>
<td></td>
<td>8) Profile-based Cross covariance (CC-PSSM) [272]</td>
</tr>
<tr>
<td></td>
<td>9) Profile-based Auto-cross covariance (ACC-PSSM) [272]</td>
</tr>
</tbody>
</table>
Mismatch [264] and Subsequence [265]; and 3 are added into the autocorrelation category, i.e., Moran autocorrelation, Geary autocorrelation, and Normalized Moreau-Broto autocorrelation [268].

**PseAAC-General** is designed to generate the feature vectors for protein sequences. For this sub-web-server, we have created a special category called “profile-based” category, into which 6 new modes are added; they are “top-n-gram” [269], "PDT-Profile" [270], “DT” [271], “AC-PSSM”, “CC-PSSM” and “ACC-PSSM” [272]. The top-n-gram combines the n most frequent amino acids in each amino acid frequency profile; PDT-Profile is the abbreviation for “Profile-based physicochemical distance transformation” and it is similar to PDT except that PDT-Profile extracts the evolutionary information from the frequency profile; DT is the abbreviation for “distance-based Top-n-gram” and this method extends Top-n-gram by considering the distances between Top-n-gram pairs; AC-PSSM, CC-PSSM and ACC-PSSM incorporate the position-specific score matrix (PSSM) into the methods of AC, CC and ACC [272, 273]. These profile-based methods can significantly improve the protein remote homology detection [7, 8], protein fold recognition and so forth. Moreover, added into the amino acid composition category are 3 new modes: they are “DR” [274], “Distance Pair” [271], and “PDT” [270]. DR is the abbreviation for “Distance-based Residue”. It is sequence-based method, in which the generated feature vector for protein sequence is based on the distance between residue pairs and has shown better performance for protein remote homology detection. “Distance Pair” method incorporates the amino acid distance pair coupling information and the amino acid reduced alphabet profile into the general pseudo amino acid composition (PseAAC) [108] vector, which is very useful for analysing DNA-binding proteins [15, 170, 189, 275]. PDT is the abbreviation for “physicochemical distance transformation”, which can incorporate considerable sequence-order information or important patterns of protein/peptide sequences into Pseudo components [28], which is very useful for conducting various proteome analyses [17, 23, 215-217, 224, 225, 231, 235, 276-289] and genome analysis as well [216, 218, 220, 223, 229, 255, 277, 290].

For more information about the three sub-webservers, see Supporting information S1.

### 2.2. New Facility

Added into the updated version is also a new facility called “Pse-in-One-Analysis”, by which the feature vectors for the input DNA, RNA, or protein sequences can be automatically generated according to the selected modes and parameters. And the results will be sent to the users via their e-mail addresses. The users can also see the result by revisiting the link concerned. Moreover, provided are also the feature vector visualization and the predicted RNA secondary structure visualization functions, which are very useful for the feature analysis and interpretation. See Supporting Information S2 for detailed information in this regard.

The stand-alone version of **Pse-in-One 2.0** is available. Users can easily download it into their own computer for conducting high throughput analysis of massive biological sequences.

By means of the new facility **Pse-in-One-Analysis** or **Pse-Analysis** [254], all the tedious jobs in developing a predictor, such as selecting optimal features and parameters as well as evaluating anticipated prediction quality, can be automatically fulfilled by the computer as elaborated in [254]. It will save scientists a lot of time, one big step forward to realize the dream of using robots or computers to conduct genome/proteome analyses.

### 2.3. New Kits

Newly provided in **Pse-in-One 2.0** are also some useful kits, including automatic notification of results by e-mail, RNA secondary structure visualization, etc. Meanwhile, some bugs have been fixed to make the web-server work more smoothly and fully consistent.

A flowchart of Pse-in-One 2.0 is given in Figure 1.

### 3. CONCLUSION

Evolved from the original **Pse-in-One** package [226], **Pse-in-One 2.0** is much more flexible and
Figure 1. The flowchart of Pse-in-One 2.0. The first two steps are implemented in Pse-in-One 2.0 webserver. The last two steps are implemented in Pse-in-One-Analysis. The output of the webserver can be directly used as the input of Pse-in-One-Analysis package.

powerful than the former. In comparison with the 2015 version that has been widely used in bioinformatics and computational biology and biomedicine within a very short period of time, the new version is even more powerful for conducting various genome analyses and proteome analyses. Science is rapidly developing, particularly in life science. Once having new and important developments, the future version for the Pse-in-One series will be announced via a publication or web-page.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (No. 61672184), the Natural Science Foundation of Guangdong Province (2014A030313695), Guangdong Natural Science Funds for Distinguished Young Scholars (2016A030306008), and Scientific Research Foundation in Shenzhen (Grant No. JCYJ 2015062611042522).

SUPPORTING INFORMATION

Supporting Information S1. Pse-in-One 2.0 Description. This document describes totally 51 different modes, including 20 modes for DNA sequences, 14 modes for RNA sequences, and 17 modes for protein sequences. The document is available at http://bioinformatics.hitsz.edu.cn/Pse-in-One2.0/static/download/Pse-in-One%202.0_description.pdf.

REFERENCES


Composition and PSSM to Predict Subcellular Locations of Proteins with Both Single and Multiple Sites. *Biosystems, 113*, 50-57. https://doi.org/10.1016/j.biosystems.2013.04.005


https://doi.org/10.4236/ns.2017.94007


260. Meher, P.K., Sahu, T.K., Saini, V. and Rao, A.R. (2017) Predicting Antimicrobial Peptides with Improved Accuracy by Incorporating the Compositional, Physico-Chemical and Structural Features into Chou’s General PseAAC. *Scientific Reports*, **7**, 42362. [https://doi.org/10.1038/srep42362](https://doi.org/10.1038/srep42362)


Submit or recommend next manuscript to SCIRP and we will provide best service for you:
Accepting pre-submission inquiries through Email, Facebook, LinkedIn, Twitter, etc.
A wide selection of journals (inclusive of 9 subjects, more than 200 journals)
Providing 24-hour high-quality service
User-friendly online submission system
Fair and swift peer-review system
Efficient typesetting and proofreading procedure
Display of the result of downloads and visits, as well as the number of cited articles
Maximum dissemination of your research work
Submit your manuscript at: http://papersubmission.scirp.org/
Or contact ns@scirp.org