Int. J. Mol. Sci. **2014**, *15*, 3495-3506; doi:10.3390/ijms15033495

OPEN ACCESS

International Journal of
Molecular Sciences
ISSN 1422-0067
www.mdpi.com/journal/ijms

Technical Note

PseAAC-General: Fast Building Various Modes of General Form of Chou's Pseudo-Amino Acid Composition for Large-Scale Protein Datasets

Pufeng Du [1,2,3,*], Shuwang Gu [1,2] and Yasen Jiao [1,2]

[1] School of Computer Science and Technology, Tianjin University, Tianjin 300072, China; E-Mails: shuwanggu@gmail.com (S.G.); yasenjiao@gmail.com (Y.J.)

[2] Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin University, Tianjin 300072, China

[3] Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong

* Author to whom correspondence should be addressed; E-Mail: pufengdu@gmail.com; Tel./Fax: +86-22-2368-9450.

Received: 20 January 2014; in revised form: 13 February 2014 / Accepted: 14 February 2014 / Published: 26 February 2014

Abstract: The general form pseudo-amino acid composition (PseAAC) has been widely used to represent protein sequences in predicting protein structural and functional attributes. We developed the program PseAAC-General to generate various different modes of Chou's general PseAAC, such as the gene ontology mode, the functional domain mode, and the sequential evolution mode. This program allows the users to define their own desired modes. In every mode, 544 physicochemical properties of the amino acids are available for choosing. The computing efficiency is at least 100 times that of existing programs, which makes it able to facilitate the extensive studies on proteins and peptides. The PseAAC-General is freely available via SourceForge. It runs on both Linux and Windows.

Keywords: general form; large-scale datasets; pseudo-amino acid composition

1. Introduction

Over the last few years, machine learning has been introduced to predict protein structures and functions. In these studies, one of the keys is to formulate the protein sequences with a mathematical form that can reflect the intrinsic correlation with their structures and functions. To be more specific, this mathematical form should keep representing a protein sequence with a discrete form yet without completely losing its sequence-order information. The pseudo-amino acid compositions (PseAAC), which was originally introduced to predict protein attributes [1], is a typical mathematical form in this regard.

Ever since its first appearance, the PseAAC formulation has been widely applied for studying various problems in protein science, such as predicting eukaryotes and prokaryotes protein subcellular locations [2–11], protein sub-subcellular locations [12–22], membrane protein subcellular locations [23–26], viral protein subcellular locations [27,28], protein structural classes [29–35], secondary structures [36], super-secondary structures [37], quaternary structural attributes [38,39], GPCR classes [40–42], enzyme families [43,44], membrane protein types [45–47], metalloproteinase families [48], risk types of human papillomavirus [49], cell-wall lytic enzymes [50], cyclic proteins [51], allergenic proteins [52], bioluminescent proteins [53], DNA-binding proteins [54], GABA(A) receptor proteins [55], bacterial virulent proteins [56], essential proteins [57], anti-cancer peptides [58], anti-bacterial peptides [59], protein-protein interactions [60], protein solubility [61], drug-target network [62], and many more [63–76]. Recently, it was applied to represent DNA sequences in identifying the recombination spot [77].

Many different types of information, such as gene ontology annotations, functional domain compositions, and sequential evolution information, have been integrated skillfully with the concept of PseAAC to represent protein samples in order to enhance the prediction quality of their attributes. In essence, the protein sample thus formulated were actually various modes of Chou's general form PseAAC, as clearly indicated by Equations 9–14 in a comprehensive review [78]. On the contrary, the Type I PseAAC [1] and Type II PseAAC [79] belong to Chou's special form PseAAC. The modes of Chou's special form PseAAC can be calculated by several programs, such as PseAAC server [80], PseAAC-Builder [81] and the propy package [82].

However, so far no publicly accessible program could calculate Chou's general PseAAC. The current PseAAC-General is a universal software platform for users to generate various modes of general form PseAAC, including several widely used modes, such as the gene ontology mode [3], functional domain mode [83], and sequential evolution mode [18]. It is anticipated that PseAAC-General will become a very useful tool in bioinformatics, computational proteomics, and system biology.

2. Results and Discussion

The current PseAAC-General can generate 13 different modes of general form PseAAC, including conventional amino acid composition, di-peptide composition, tri-peptide composition, Type I PseAAC, Type II PseAAC, the gene ontology mode, the functional domain mode, the sequential evolution mode, the normalized Moreau-Broto autocorrelation coefficients, the Moran autocorrelation coefficients, the Geary autocorrelation coefficients, the composition-transition-distribution (CTD) descriptors and the quasi-sequence order descriptors. In every mode, 544 types of physicochemical properties are available for choosing. Over 20,000 different descriptor values can be calculated.

We list several commonly used modes of general form PseAAC as well as some program features in PseAAC-General program in Table 1. Several modes are uniquely available in PseAAC-General, which include the gene ontology mode, the functional domain mode and the sequential evolution mode. These modes have been mentioned in existing programs [81,82]. However, no program implemented these modes.

Table 1. Comparison of program features.

Program Functions [a]	PseAAC-General	PseAAC-Builder	Propy	PseAAC Server
Physicochemical Properties	544	544	8	6
Output Features				
Type I PseAAC [1]	Y	Y	Y	Y
Type II PseAAC [79]	Y	Y	Y	Y
Amino acid composition	Y	Y	Y	Y
di-Peptide composition	Y	Y	Y	Y
tri-Peptide composition	Y	N	Y	N
Normalized Moreau-Broto autocorrelation [84,85]	Y	N	Y	N
Moran autocorrelation [86]	Y	N	Y	N
Geary autocorrelation [87]	Y	N	Y	N
Composition-Transition-Distribution (CTD) [88]	Y	N	Y	N
Quasi-sequence order [89]	Y	N	Y	N
Gene ontology mode [83]	Y	N	N	N
Functional domain mode [83]	Y	N	N	N
Sequential evolution mode [18]	Y	N	N	N
Other functions				
User defined	Y	N	N	N
Online updates	Y	N	N	N
Graphical User Interface (GUI)	Y	Y	N	Y
Execution efficiency [b]	~17,000 seqs/s	~170 seqs/s	N.A.	~15 seqs/s

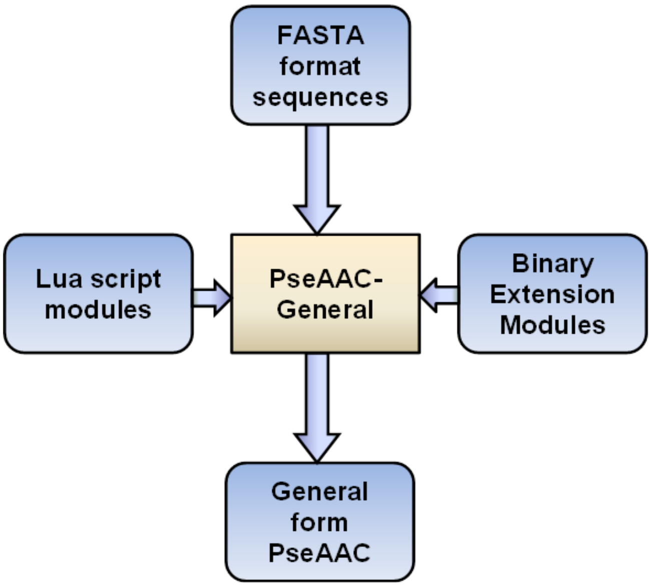
[a] The program functions that were compared. There are three groups of functions, including the physicochemical properties, the sequence features that can be generated and the other function properties of the software. Y = YES; N = NO; [b] the execution time for PseAAC-General and PseAAC-Builder was tested on a dataset containing over 510,000 sequences by the wall-clock time. The execution time for PseAAC-Server was tested on a dataset containing 500 sequences due to the limitation of the service and the internet connection conditions. The execution time for Propy was not tested due the limitation of testing environments. Seqs/s means sequences per second.

PseAAC-General provided two methods for the users to create their own desired modes. The first method is called the Binary Extension Module (BEM). The gene ontology mode and functional domain mode were actually implemented by this method. A set of tools was provided along with the PseAAC-General, so that the users can create their own BEM to represent all kinds of descriptive information, which includes but not limited to the gene ontology annotations and the functional domain compositions.

The other method is the Lua script module. Lua script language is a very simple programming language that has been considered in analyzing sequence annotations [90]. We provided a programming interface that allows the user to use Lua script to access the internal data structures and functions of PseAAC-General. Furthermore, the algorithm modules of PseAAC-General can be replaced by the user-defined Lua script modules. This provides the maximal flexibility for the user-defined mode. Actually, the sequential evolution mode was implemented in this way.

Because of these extension modules, the input to the PseAAC-General is not only the protein sequences. These extension modules should also be loaded if they are needed. We illustrate the data flow of PseAAC-General in Figure 1.

Figure 1. The data flow of pseudo-amino acid composition (PseAAC)-General. The input data is FASTA format sequences. The output data is general form PseAAC. The mode of the general form PseAAC is chosen by the users. For the modes, which are implemented by Binary Extension Modules or Lua script modules, the corresponding modules should be loaded as well.



The usefulness of PseAAC-General is undisputed. In the early days of general form PseAAC, every study had to implement the PseAAC independently. This may bring a number of problems, including but not limited to inconsistent results, different computation efficiency and different basis in comparing predictive performance. PseAAC-General can serve as a standard program that saves time for all these studies. Furthermore, our program eliminates those unforeseen problems that were brought by the different implementations of PseAAC.

PseAAC-General is much faster than existing programs. We tested PseAAC-General by using it to calculate Type I PseAAC with default parameters. On the same machine that we tested PseAAC-Builder [81], it can process about 17,000 sequences per second. This is about 100 times faster than PseAAC-Builder. In other words, PseAAC-General can convert the entire Swiss-Prot database to Type I PseAAC within 30 s, while PseAAC-Builder needs about 40 min.

3. Implementations

PseAAC-General is released under GNU GPL (GNU General Public License). It can be integrated with other programs in the source code level. We have ported PseAAC-General to both Linux and Windows platforms. A GUI (Graphical User Interface) module was provided for both platforms. The users, who do not familiar with the command line, can use PseAAC-General through GUI. However, it should be noted that the most efficient way is the command line, which was designed to follow the GNU command line standard.

PseAAC-General was designed to be a stand-alone program running on the local machine without internet connection requirements. Therefore, we did not include the online sequence retrieving function within the program. On the other hand, the propy package has perfectly implemented the retrieving function. The best choice for the users is to let PseAAC-General work side by side with the propy package. For example, the users can use Propy to retrieve protein sequences and call PseAAC-General to calculate the PseAAC, as python environment has the built-in ability to call external programs, like PseAAC-General. In future versions of PseAAC-General, a similar function will be implemented. PseAAC-General and all its extension modules can be downloaded from its website [91]. To facilitate further studies, all source code of PseAAC-General, including the main program, GUI module and all extension modules, can be freely downloaded from the SourceForge website [92]. We also provided detailed documents within the software package, so that the users can learn not only how to use the existing modes, but also how to create their own modes by building their own extension modules. For the users' convenience to test their own modes, we provided four different testing dataset with different size. These testing datasets can also be downloaded from the website. Along with the testing datasets, we provided simple testing scripts to demonstrate the usage of PseAAC-General in a common case. The users can simply try the testing scripts to learn how to use the program.

Because the gene ontology mode and the functional domain mode should be upgraded along with the Swiss-Prot database, we deployed a cloud-computation based server in Amazon EC2 (Elastic Cloud 2, Amazon.com Inc., Seattle, WA, USA) to automatically upgrade the relevant extension modules on monthly basis.

4. Conclusions

As PseAAC-General is a very powerful and very flexible computation tool, we believe that PseAAC-General will facilitate all studies that apply the general form PseAAC, including those existing modes and those modes in development.

However, as a final reminder, we would like to remind the users to read the manual of PseAAC-General and those literatures describing the algorithm of general form PseAAC carefully before using it. Because of the powerful function and the flexibility of PseAAC-General, using it in your study without knowing the algorithms and technics behind the program and the source code could be very risky.

Acknowledgments

This work was supported by National Science Foundation of China (NSFC 61005041); Specialized Research Fund for the Doctoral Program of Higher Education (SRFDP 20100032120039); Tianjin Natural Science Foundation (No. 12JCQNJC02300); China Post-doctoral Science Foundation (2012T50240 and 2013M530114); HK Scholar Program (XJ2012033) and the Seed Foundation of Tianjin University (No. 60302006 and 60302024).

Author Contributions

P.D. designed the software, partially wrote the code and wrote the manuscript. S.G. and Y.J. partially wrote the code, carried out testing experiments and partially wrote the manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Chou, K.C. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins* **2001**, *43*, 246–255.
2. Lee, K.; Chuang, H.-Y.; Beyer, A.; Sung, M.-K.; Huh, W.-K.; Lee, B.; Ideker, T. Protein networks markedly improve prediction of subcellular localization in multiple eukaryotic species. *Nucleic Acids Res.* **2008**, *36*, e136.
3. Chou, K.-C.; Shen, H.-B. Cell-PLoc: A package of web servers for predicting subcellular localization of proteins in various organisms. *Nat. Protoc.* **2008**, *3*, 153–162.
4. Huang, C.; Yuan, J. Using radial basis function on the general form of Chou's pseudo amino acid composition and PSSM to predict subcellular locations of proteins with both single and multiple sites. *BioSystems* **2013**, *113*, 50–57.
5. Jiang, X.; Wei, R.; Zhang, T.; Gu, Q. Using the concept of Chou's pseudo amino acid composition to predict apoptosis proteins subcellular location: An approach by approximate entropy. *Protein Pept. Lett.* **2008**, *15*, 392–396.
6. Lin, H.; Wang, H.; Ding, H.; Chen, Y.-L.; Li, Q.-Z. Prediction of subcellular localization of apoptosis protein using Chou's pseudo amino acid composition. *Acta Biotheor.* **2009**, *57*, 321–330.
7. Lin, J.; Wang, Y. Using a novel AdaBoost algorithm and Chou's Pseudo amino acid composition for predicting protein subcellular localization. *Protein Pept. Lett.* **2011**, *18*, 1219–1225.
8. Mei, S. Predicting plant protein subcellular multi-localization by Chou's PseAAC formulation based multi-label homolog knowledge transfer learning. *J. Theor. Biol.* **2012**, *310*, 80–87.
9. Pacharawongsakda, E.; Theeramunkong, T. Predict subcellular locations of singleplex and multiplex proteins by semi-supervised learning and dimension-reducing general mode of Chou's PseAAC. *NanoBioscience* **2013**, *12*, 311–320.

10. Wan, S.; Mak, M.-W.; Kung, S.-Y. GOASVM: A subcellular location predictor by incorporating term-frequency gene ontology into the general form of Chou's pseudo-amino acid composition. *J. Theor. Biol.* **2013**, *323*, 40–48.
11. Wang, X.; Li, G.-Z.; Lu, W.-C. Virus-ECC-mPLoc: A multi-label predictor for predicting the subcellular localization of virus proteins with both single and multiple sites based on a general form of Chou's pseudo amino acid composition. *Protein Pept. Lett.* **2013**, *20*, 309–317.
12. Du, P.; Li, Y. Prediction of protein submitochondria locations by hybridizing pseudo-amino acid composition with various physicochemical features of segmented sequence. *BMC Bioinforma.* **2006**, *7*, 518.
13. Du, P.; Yu, Y. SubMito-PSPCP: Predicting protein submitochondrial locations by hybridizing positional specific physicochemical properties with pseudoamino acid compositions. *BioMed Res. Int.* **2013**, *2013*, 263829–263836.
14. Fan, G.-L.; Li, Q.-Z. Predicting protein submitochondria locations by combining different descriptors into the general form of Chou's pseudo amino acid composition. *Amino Acids* **2012**, *43*, 545–555.
15. Mei, S. Multi-kernel transfer learning based on Chou's PseAAC formulation for protein submitochondria localization. *J. Theor. Biol.* **2012**, *293*, 121–130.
16. Huang, C.; Yuan, J.-Q. Predicting protein subchloroplast locations with both single and multiple sites via three different modes of Chou's pseudo amino acid compositions. *J. Theor. Biol.* **2013**, *335*, 205–212.
17. Jiang, X.; Wei, R.; Zhao, Y.; Zhang, T. Using Chou's pseudo amino acid composition based on approximate entropy and an ensemble of AdaBoost classifiers to predict protein subnuclear location. *Amino Acids* **2008**, *34*, 669–675.
18. Shen, H.-B.; Chou, K.-C. Nuc-PLoc: A new web-server for predicting protein subnuclear localization by fusing PseAA composition and PsePSSM. *Protein Eng. Des. Sel.* **2007**, *20*, 561–567.
19. Li, F.-M.; Li, Q.-Z. Predicting protein subcellular location using Chou's pseudo amino acid composition and improved hybrid approach. *Protein Pept. Lett.* **2008**, *15*, 612–616.
20. Li, L.-Q.; Zhang, Y.; Zou, L.-Y.; Zhou, Y.; Zheng, X.-Q. Prediction of protein subcellular multi-localization based on the general form of Chou's pseudo amino acid composition. *Protein Pept. Lett.* **2012**, *19*, 375–387.
21. Nanni, L.; Lumini, A. Genetic programming for creating Chou's pseudo amino acid based features for submitochondria localization. *Amino Acids* **2008**, *34*, 653–660.
22. Zeng, Y.; Guo, Y.; Xiao, R.; Yang, L.; Yu, L.; Li, M. Using the augmented Chou's pseudo amino acid composition for predicting protein submitochondria locations based on auto covariance approach. *J. Theor. Biol.* **2009**, *259*, 366–372.
23. Pierleoni, A.; Martelli, P.L.; Casadio, R. MemLoci: Predicting subcellular localization of membrane proteins in eukaryotes. *Bioinformatics* **2011**, *27*, 1224–1230.
24. Du, P.; Tian, Y.; Yan, Y. Subcellular localization prediction for human internal and organelle membrane proteins with projected gene ontology scores. *J. Theor. Biol.* **2012**, *313*, 61–67.

25. Huang, C.; Yuan, J.-Q. A multilabel model based on Chou's pseudo-amino acid composition for identifying membrane proteins with both single and multiple functional types. *J. Membr. Biol.* **2013**, *246*, 327–334.
26. Zhang, S.-W.; Zhang, Y.-L.; Yang, H.-F.; Zhao, C.-H.; Pan, Q. Using the concept of Chou's pseudo amino acid composition to predict protein subcellular localization: An approach by incorporating evolutionary information and von Neumann entropies. *Amino Acids* **2008**, *34*, 565–572.
27. Cao, J.-Z.; Liu, W.-Q.; Gu, H. Predicting viral protein subcellular localization with Chou's pseudo amino acid composition and imbalance-weighted multi-label K-nearest neighbor algorithm. *Protein Pept. Lett.* **2012**, *19*, 1163–1169.
28. Shen, H.-B.; Chou, K.-C. Virus-mPLoc: A fusion classifier for viral protein subcellular location prediction by incorporating multiple sites. *J. Biomol. Struct. Dyn.* **2010**, *28*, 175–186.
29. Sahu, S.S.; Panda, G. A novel feature representation method based on Chou's pseudo amino acid composition for protein structural class prediction. *Comput. Biol. Chem.* **2010**, *34*, 320–327.
30. Chen, C.; Shen, Z.-B.; Zou, X.-Y. Dual-layer wavelet SVM for predicting protein structural class via the general form of Chou's pseudo amino acid composition. *Protein Pept. Lett.* **2012**, *19*, 422–429.
31. Kong, L.; Zhang, L.; Lv, J. Accurate prediction of protein structural classes by incorporating predicted secondary structure information into the general form of Chou's pseudo amino acid composition. *J. Theor. Biol.* **2013**, *344*, 12–18.
32. Li, Z.-C.; Zhou, X.-B.; Dai, Z.; Zou, X.-Y. Prediction of protein structural classes by Chou's pseudo amino acid composition: Approached using continuous wavelet transform and principal component analysis. *Amino Acids* **2009**, *37*, 415–425.
33. Liao, B.; Xiang, Q.; Li, D. Incorporating secondary features into the general form of Chou's PseAAC for predicting protein structural class. *Protein Pept. Lett.* **2012**, *19*, 1133–1138.
34. Liu, L.; Hu, X.-Z.; Liu, X.-X.; Wang, Y.; Li, S.-B. Predicting protein fold types by the general form of Chou's pseudo amino acid composition: Approached from optimal feature extractions. *Protein Pept. Lett.* **2012**, *19*, 439–449.
35. Qin, Y.-F.; Wang, C.-H.; Yu, X.-Q.; Zhu, J.; Liu, T.-G.; Zheng, X.-Q. Predicting protein structural class by incorporating patterns of over-represented k-mers into the general form of Chou's PseAAC. *Protein Pept. Lett.* **2012**, *19*, 388–397.
36. Chen, C.; Chen, L.; Zou, X.; Cai, P. Prediction of protein secondary structure content by using the concept of Chou's pseudo amino acid composition and support vector machine. *Protein Pept. Lett.* **2009**, *16*, 27–31.
37. Zou, D.; He, Z.; He, J.; Xia, Y. Supersecondary structure prediction using Chou's pseudo amino acid composition. *J. Comput. Chem.* **2011**, *32*, 271–278.
38. Sun, X.-Y.; Shi, S.-P.; Qiu, J.-D.; Suo, S.-B.; Huang, S.-Y.; Liang, R.-P. Identifying protein quaternary structural attributes by incorporating physicochemical properties into the general form of Chou's PseAAC via discrete wavelet transform. *Mol. Biosyst.* **2012**, *8*, 3178–3184.
39. Zhang, S.-W.; Chen, W.; Yang, F.; Pan, Q. Using Chou's pseudo amino acid composition to predict protein quaternary structure: A sequence-segmented PseAAC approach. *Amino Acids* **2008**, *35*, 591–598.

40. Gu, Q.; Ding, Y.S.; Zhang, T.L. Prediction of G-protein-coupled receptor classes in low homology using Chou's pseudo amino acid composition with approximate entropy and hydrophobicity patterns. *Protein Pept. Lett.* **2010**, *17*, 559–567.
41. Qiu, J.-D.; Huang, J.-H.; Liang, R.-P.; Lu, X.-Q. Prediction of G-protein-coupled receptor classes based on the concept of Chou's pseudo amino acid composition: An approach from discrete wavelet transform. *Anal. Biochem.* **2009**, *390*, 68–73.
42. Zia-Ur-Rehman; Khan, A. Identifying GPCRs and their types with Chou's pseudo amino acid composition: An approach from multi-scale energy representation and position specific scoring matrix. *Protein Pept. Lett.* **2012**, *19*, 890–903.
43. Qiu, J.-D.; Huang, J.-H.; Shi, S.-P.; Liang, R.-P. Using the concept of Chou's pseudo amino acid composition to predict enzyme family classes: An approach with support vector machine based on discrete wavelet transform. *Protein Pept. Lett.* **2010**, *17*, 715–722.
44. Zhou, X.-B.; Chen, C.; Li, Z.-C.; Zou, X.-Y. Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. *J. Theor. Biol.* **2007**, *248*, 546–551.
45. Chen, Y.-K.; Li, K.-B. Predicting membrane protein types by incorporating protein topology, domains, signal peptides, and physicochemical properties into the general form of Chou's pseudo amino acid composition. *J. Theor. Biol.* **2013**, *318*, 1–12.
46. Han, G.-S.; Yu, Z.-G.; Anh, V. A two-stage SVM method to predict membrane protein types by incorporating amino acid classifications and physicochemical properties into a general form of Chou's PseAAC. *J. Theor. Biol.* **2013**, *344*, 31–39.
47. Hayat, M.; Khan, A. Discriminating outer membrane proteins with fuzzy K-nearest neighbor algorithms based on the general form of Chou's PseAAC. *Protein Pept. Lett.* **2012**, *19*, 411–421.
48. Mohammad Beigi, M.; Behjati, M.; Mohabatkar, H. Prediction of metalloproteinase family based on the concept of Chou's pseudo amino acid composition using a machine learning approach. *J. Struct. Funct. Genomics* **2011**, *12*, 191–197.
49. Esmaeili, M.; Mohabatkar, H.; Mohsenzadeh, S. Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses. *J. Theor. Biol.* **2010**, *263*, 203–209.
50. Ding, H.; Luo, L.; Lin, H. Prediction of cell wall lytic enzymes using Chou's amphiphilic pseudo amino acid composition. *Protein Pept. Lett.* **2009**, *16*, 351–355.
51. Mohabatkar, H. Prediction of cyclin proteins using Chou's pseudo amino acid composition. *Protein Pept. Lett.* **2010**, *17*, 1207–1214.
52. Mohabatkar, H.; Mohammad Beigi, M.; Abdolahi, K.; Mohsenzadeh, S. Prediction of allergenic proteins by means of the concept of Chou's pseudo amino acid composition and a machine learning approach. *Med. Chem.* **2013**, *9*, 133–137.
53. Fan, G.-L.; Li, Q.-Z. Discriminating bioluminescent proteins by incorporating average chemical shift and evolutionary information into the general form of Chou's pseudo amino acid composition. *J. Theor. Biol.* **2013**, *334*, 45–51.
54. Fang, Y.; Guo, Y.; Feng, Y.; Li, M. Predicting DNA-binding proteins: Approached from Chou's pseudo amino acid composition and other specific sequence features. *Amino Acids* **2008**, *34*, 103–109.

55. Mohabatkar, H.; Mohammad Beigi, M.; Esmaeili, A. Prediction of GABAA receptor proteins using the concept of Chou's pseudo-amino acid composition and support vector machine. *J. Theor. Biol.* **2011**, *281*, 18–23.
56. Nanni, L.; Lumini, A.; Gupta, D.; Garg, A. Identifying bacterial virulent proteins by fusing a set of classifiers based on variants of Chou's pseudo amino acid composition and on evolutionary information. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* **2012**, *9*, 467–475.
57. Sarangi, A.N.; Lohani, M.; Aggarwal, R. Prediction of essential proteins in prokaryotes by incorporating various physico-chemical features into the general form of Chou's pseudo amino acid composition. *Protein Pept. Lett.* **2013**, *20*, 781–795.
58. Hajisharifi, Z.; Piryaiee, M.; Mohammad Beigi, M.; Behbahani, M.; Mohabatkar, H. Predicting anticancer peptides with Chou's pseudo amino acid composition and investigating their mutagenicity via Ames test. *J. Theor. Biol.* **2014**, *341*, 34–40.
59. Khosravian, M.; Faramarzi, F.K.; Beigi, M.M.; Behbahani, M.; Mohabatkar, H. Predicting antibacterial peptides by the concept of Chou's pseudo-amino acid composition and machine learning methods. *Protein Pept. Lett.* **2013**, *20*, 180–186.
60. Zhao, X.-W.; Ma, Z.-Q.; Yin, M.-H. Predicting protein–protein interactions by combing various sequence-derived features into the general form of Chou's pseudo amino acid composition. *Protein Pept. Lett.* **2012**, *19*, 492–500.
61. Niu, X.-H.; Hu, X.-H.; Shi, F.; Xia, J.-B. Predicting protein solubility by the general form of Chou's pseudo amino acid composition: Approached from chaos game representation and fractal dimension. *Protein Pept. Lett.* **2012**, *19*, 940–948.
62. Yu, H.; Chen, J.; Xu, X.; Li, Y.; Zhao, H.; Fang, Y.; Li, X.; Zhou, W.; Wang, W.; Wang, Y.A. Systematic prediction of multiple drug-target interactions from chemical, genomic, and pharmacological data. *PLoS One* **2012**, *7*, e37608.
63. Georgiou, D.N.; Karakasidis, T.E.; Nieto, J.J.; Torres, A. Use of fuzzy clustering technique and matrices to classify amino acids and its impact to Chou's pseudo amino acid composition. *J. Theor. Biol.* **2009**, *257*, 17–26.
64. Gupta, M.K.; Niyogi, R.; Misra, M. An alignment-free method to find similarity among protein sequences via the general form of Chou's pseudo amino acid composition. *SAR QSAR Environ. Res.* **2013**, *24*, 597–609.
65. Lin, H. The modified Mahalanobis discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition. *J. Theor. Biol.* **2008**, *252*, 350–356.
66. Nanni, L.; Brahnam, S.; Lumini, A. Wavelet images and Chou's pseudo amino acid composition for protein classification. *Amino Acids* **2012**, *43*, 657–665.
67. Qiu, J.-D.; Suo, S.-B.; Sun, X.-Y.; Shi, S.-P.; Liang, R.-P. OligoPred: A web-server for predicting homo-oligomeric proteins by incorporating discrete wavelet transform into Chou's pseudo amino acid composition. *J. Mol. Graph. Model.* **2011**, *30*, 129–134.
68. Ren, L.-Y.; Zhang, Y.-S.; Gutman, I. Predicting the classification of transcription factors by incorporating their binding site properties into a novel mode of Chou's pseudo amino acid composition. *Protein Pept. Lett.* **2012**, *19*, 1170–1176.

69. Xiaohui, N.; Nana, L.; Jingbo, X.; Dingyan, C.; Yuehua, P.; Yang, X.; Weiquan, W.; Dongming, W.; Zengzhen, W. Using the concept of Chou's pseudo amino acid composition to predict protein solubility: An approach with entropies in information theory. *J. Theor. Biol.* **2013**, *332*, 211–217.
70. Xie, H.-L.; Fu, L.; Nie, X.-D. Using ensemble SVM to identify human GPCRs *N*-linked glycosylation sites based on the general form of Chou's PseAAC. *Protein Eng. Des. Sel.* **2013**, *26*, 735–742.
71. Yu, L.; Guo, Y.; Li, Y.; Li, G.; Li, M.; Luo, J.; Xiong, W.; Qin, W. SecretP: Identifying bacterial secreted proteins by fusing new features into Chou's pseudo-amino acid composition. *J. Theor. Biol.* **2010**, *267*, 1–6.
72. Zhang, G.-Y.; Fang, B.-S. Predicting the cofactors of oxidoreductases based on amino acid composition distribution and Chou's amphiphilic pseudo-amino acid composition. *J. Theor. Biol.* **2008**, *253*, 310–315.
73. Zhang, G.-Y.; Li, H.-C.; Gao, J.-Q.; Fang, B.-S. Predicting lipase types by improved Chou's pseudo-amino acid composition. *Protein Pept. Lett.* **2008**, *15*, 1132–1137.
74. Liu, B.; Wang, X.; Zou, Q.; Dong, Q.; Chen, Q. Protein remote homology detection by combining Chou's pseudo amino acid composition and profile-based protein representation. *Mol. Inform.* **2013**, *32*, 775–782.
75. Georgiou, D.N.; Karakasidis, T.E.; Nieto, J.J.; Torres, A. A study of entropy/clarity of genetic sequences using metric spaces and fuzzy sets. *J. Theor. Biol.* **2010**, *267*, 95–105.
76. Georgiou, T.N.; Karakasidis, T.E.; Megaritis, A.C. A short survey on genetic sequences, Chou's pseudo amino acid composition and its combination with fuzzy set theory. *Open Bioinforma. J.* **2013**, *7*, 41–48.
77. Chen, W.; Feng, P.-M.; Lin, H.; Chou, K.-C. iRSpot-PseDNC: Identify recombination spots with pseudo dinucleotide composition. *Nucleic Acids Res.* **2013**, *41*, e68.
78. Chou, K.-C. Some remarks on protein attribute prediction and pseudo amino acid composition. *J. Theor. Biol.* **2011**, *273*, 236–247.
79. Chou, K.-C. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics* **2005**, *21*, 10–19.
80. Shen, H.-B.; Chou, K.-C. PseAAC: A flexible web server for generating various kinds of protein pseudo amino acid composition. *Anal. Biochem.* **2008**, *373*, 386–388.
81. Du, P.; Wang, X.; Xu, C.; Gao, Y. PseAAC-Builder: A cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions. *Anal. Biochem.* **2012**, *425*, 117–119.
82. Cao, D.-S.; Xu, Q.-S.; Liang, Y.-Z. Propy: A tool to generate various modes of Chou's PseAAC. *Bioinformatics* **2013**, *29*, 960–962.
83. Chou, K.-C.; Cai, Y.-D. Prediction of protein subcellular locations by GO-FunD-PseAA predictor. *Biochem. Biophys. Res. Commun.* **2004**, *320*, 1236–1239.
84. Feng, Z.P.; Zhang, C.T. Prediction of membrane protein types based on the hydrophobic index of amino acids. *J. Protein Chem.* **2000**, *19*, 269–275.
85. Lin, Z.; Pan, X.M. Accurate prediction of protein secondary structural content. *J. Protein Chem.* **2001**, *20*, 217–220.

86. Horne, D.S. Prediction of protein helix content from an autocorrelation analysis of sequence hydrophobicities. *Biopolymers* **1988**, *27*, 451–477.
87. Sokal, R.R.; Thomson, B.A. Population structure inferred by local spatial autocorrelation: An example from an Amerindian tribal population. *Am. J. Phys. Anthropol.* **2006**, *129*, 121–131.
88. Dubchak, I.; Muchnik, I.; Mayor, C.; Dralyuk, I.; Kim, S.H. Recognition of a protein fold in the context of the Structural Classification of Proteins (SCOP) classification. *Proteins* **1999**, *35*, 401–407.
89. Chou, K.-C. Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. *Biochem. Biophys. Res. Commun.* **2000**, *27*, 477–483.
90. Steinbiss, S.; Gremme, G.; Schärfer, C.; Mader, M.; Kurtz, S. AnnotationSketch: A genome annotation drawing library. *Bioinformatics* **2009**, *25*, 533–534.
91. PseAAC-General. Available online: http://pseb.sf.net (accessed on 19 February 2014).
92. PseAAC-General SourceForge Site. Available online: http://sourceforge.net/projects/pseb/files (accessed on 19 February 2014).