https://www.scirp.org/journal/ns Natural Science, 2020, Vol. 12, (No. 3), pp: 74-79

Proposing 5-Steps Rule Is a Notable Milestone for Studying Molecular Biology

Kuo-Chen Chou

Gordon Life Science Institute, Boston, Massachusetts 02478, United States of America

Correspondence to: Kuo-Chen Chou, kcchou@gordonlifescience.org, kcchou38@gmail.com

Keywords: 5-Steps Rule, Cradle, Global and Local Metrics, Multi-Label System, Web-Server

Received: February 21, 2020 **Accepted:** March 7, 2020 **Published:** March 10, 2020

Copyright © 2020 by author(s) 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/

Open Access

ABSTRACT

In this current minireview, the cradle of the "5-steps rule" or "5-step rules", along with its essence and advances, has been recalled. Born in 2011, its impacts on molecular biology are both substantial and rapid, fully indicating the "5-steps rule" is no double a remarkable and profound milestone in molecular biology.

1. INTRODUCTION

Since it was proposed in 2011, the "5-steps rule" or "5-step rules" has been widely used in molecular biology, both theoretical and experimental. Its original source was usually referred by citing a review paper for celebrating the 50th anniversary year of Journal of Theoretical Biology [1].

Interestingly, no such a clear-cut term as "5-step" can be found in the entire aforementioned paper. Why? This is because: it is the idea of the "5-steps rule" that would become crystal clear after carefully reading through the whole paper. Accordingly, the paper [1] is actually the cradle of the "5-steps rule".

2. THE ESSENCE OF 5-STEPS RULE

In order to quantitatively predict, or develop a useful predictor for, a molecular biology system, the following five guidelines should be observed: 1) select or construct a valid benchmark dataset to train and test the predictor; 2) represent the samples with an effective formulation that can truly reflect their intrinsic correlation with the target to be predicted; 3) introduce or develop a powerful algorithm to conduct the prediction; 4) properly perform cross-validation tests to objectively evaluate the anticipated prediction accuracy; 5) establish a user-friendly web-server for the predictor that is accessible to the public. The predictors established in compliance with these steps have the following notable merits: a) crystal clear in logic development; b) completely transparent in operation; c) easily to repeat the reported results by other investigators; d) with high potential in stimulating other predictors; e) very convenient to be used by the majority of experimental scientists.

3. RESULT AND DISCUSSION

It is without exaggeration to say that the "5-steps rule" has been used at a very deeper levels of many molecular biology systems, as clearly and remarkably indicated by a series of the following reports: 1) "prediction of S-sulfenylation sites [2], 2) "identify phosphohistidine sites in proteins by blending statistical moments and position relative features [3], 3) "identify tyrosine sulfation sites by incorporating statistical moments" [4], 4) "prediction of S-sulfenylation sites using statistical moments [5], 5) "reveal active compound and mechanism of shuangsheng pingfei san on idiopathic pulmonary fibrosis [6], 6) "exploring DNA-binding proteins by integrating multi-scale sequence information" [7], 7) "predict splice junctions with interpretable bidirectional long short-term memory networks" [8], 8) "identify hydroxylation sites in proteins by extracting enhanced position and sequence variant feature" [9], 9) "a sequence model for identifying S-palmitoylation sites in proteins" [10], 10) "a sequence-based model for identifying S-prenylation sites in proteins" [11], 11) "a two-level computation model based on deep learning algorithm for identification of piRNA and their functions [12], 12) "deep learning-based recombination spots prediction by incorporating secondary sequence information coupled with physio-chemical properties" [13], 13) "a study for therapeutic treatment against Parkinson's disease" [14], 14) "identifying DNA N(6)-methyladenine sites in rice genome using continuous bag of nucleobases" [15], 15) "identifying enhancers using hidden information of DNA sequences" [16], 16) "identifying molecular functions of cytoskeleton motor proteins using 2D convolutional neural network" [17], 17) "identifying cancer targets based on machine learning methods" [18], 18) "identifying DNase I hypersensitive sites using multi-features fusion and F-score features selection" [19], 19) "an improved bioinformatics tool for identifying DNA 6 mA modifications" [20], 20) "identify lysine crotonylation sites by blending position relative statistical features" [21], 21) "identifying RNA N6-methyladenosine sites using deep learning mode" [22], 22) "detecting formylation sites from protein sequences using K-nearest neighbor algorithm" [23], 23) "identification of DNA N6-methyladenine sites in the rice genome by intelligent computational model" [24], 24) "calcium pattern assessment in patients with severe aortic stenosis" [25], 25) "identifying FL11 subtype by characterizing tumor immune microenvironment in prostate adenocarcinoma" [26], 26) "a sequence-based tool for the prediction and analysis of quorum sensing peptides" [27], 27) "evaluate the stability of tautomers: susceptibility of 2-[(Phenylimino)-methyl]-cyclohexane-1,3-diones to tautomerization based on the calculated Gibbs free energies" [28], 28) "prediction of lysine formylation sites using the composition of k-spaced amino acid pairs" [29], 29) "a two-level sequence-based predictor for identifying nuclear receptors and their families" [30], 30) "a two-layer predictor for identifying proteases and their types" [31], 31) "classifying anticancer peptides using discriminative intelligent model" [32], 32) "a tool for protein physicochemical descriptor generation" [33], 33) "model feedback in lung cancer" [34].

It is instructive to point out that in the systems of molecular biology there exist many multi-label ones where each of the individual constituents or samples considered may need two or more labels for distinction. For this kind of multi-label systems, two kinds of metrics are needed: one is the global set of metrics to indicate the global accuracy of the prediction method or predictor developed, while the other is the local metrics to indicate its local accuracy [35]. For the concrete mathematical formulations of the two sets of metrics, as well as their biological implications, refer to a recent paper [36].

4. CONCLUSION AND PERSPECTIVE

The "5-steps rule" has played substantial roles in stimulating in-depth studies of molecular biology, both theoretical and experimental. It is indeed a remarkable and profound milestone for molecular biology.

Although at the present the reports in this regard from theoretical scientists are more than those from experimental scientists, it is anticipated that, with more experimental data available in future, this kind of reports from experimental scientists will be increasing as well. Particularly, the combined reports between experimental and theoretical approaches, or their compliments to each other, will increasingly appear.

It is anticipated that more impacts will be realized by the "5-steps rule", as indicated by some very

impressive papers [35-41] and a series of very recent papers (see, e.g., [42-59]).

CONFLICTS OF INTEREST

The author declares no conflicts of interest regarding the publication of this paper.

REFERENCES

1. Chou, K.C. (2011) Some Remarks on Protein Attribute Prediction and Pseudo Amino Acid Composition (50th Anniversary Year Review, 5-Steps Rule). *Journal of Theoretical Biology*, **273**, 236-247. https://doi.org/10.1016/j.jtbi.2010.12.024
2. Butt, A.H. and Khan, Y.D. (2018) Prediction of S-Sulfenylation Sites Using Statistical Moments Based Features via Chou's 5-Step Rule. *International Journal of Peptide Research and Therapeutics*.
3. Awais, M., Hussain, W., Khan, Y.D., Rasool, N., Khan, S.A. and Chou, K.C. (2019) iPhosH-PseAAC: Identify Phosphohistidine Sites in Proteins by Blending Statistical Moments and Position Relative Features According to the Chou's 5-Step Rule and General Pseudo Amino Acid Composition. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. https://www.ncbi.nlm.nih.gov/pubmed/31144645 https://doi.org/10.1109/TCBB.2019.2919025
4. Barukab, O., Khan, Y.D., Khan, S.A. and Chou, K.C. (2019) iSulfoTyr-PseAAC: Identify Tyrosine Sulfation Sites by incorporating Statistical Moments via Chou's 5-Steps Rule and Pseudo Components. *Current Genomics*, **20**, 306-320. http://www.eurekaselect.com/174277/article
5. Butt, A.H. and Khan, Y.D. (2019) Prediction of S-Sulfenylation Sites Using Statistical Moments Based Features via Chou's 5-Step Rule. *International Journal of Peptide Research and Therapeutics*.
6. Chen, Y. and Fan, X. (2019) Use Chou's 5-Steps Rule to Reveal Active Compound and Mechanism of Shuangsheng Pingfei San on Idiopathic Pulmonary Fibrosis. *Current Molecular Medicine*, **20**, 220-230.
7. Du, X., Diao, Y., Liu, H. and Li, S. (2019) MsDBP: Exploring DNA-Binding Proteins by Integrating Multi-Scale Sequence Information via Chou's 5-Steps Rule. *Journal of Proteome Research*, **18**, 3119-3132. https://doi.org/10.1021/acs.jproteome.9b00226
8. Dutta, A., Dalmia, A., Singh, K.K. and Anand, A. (2019) Using the Chou's 5-Steps Rule to Predict Splice Junctions with Interpretable Bidirectional Long Short-Term Memory Networks. *Computers in Biology and Medicine*, **116**, Article ID: 103558. https://doi.org/10.1016/j.compbiomed.2019.103558
9. Ehsan, A., Mahmood, M.K., Khan, Y.D., Barukab, O.M., Khan, S.A. and Chou, K.C. (2019) iHyd-PseAAC (EPSV): Identify Hydroxylation Sites in Proteins by Extracting Enhanced Position and Sequence Variant Feature via Chou's 5-Step Rule and General Pseudo Amino Acid Composition. *Current Genomics*, **20**, 124-133. https://doi.org/10.2174/1389202920666190325162307
10. Hussain, W., Khan, S.D., Rasool, N., Khan, S.A. and Chou, K.C. (2019) SPalmitoylC-PseAAC: A Sequence-Based Model Developed via Chou's 5-Steps Rule and General PseAAC for Identifying S-palmitoylation Sites in Proteins. *Analytical Biochemistry*, **568**, 14-23. https://doi.org/10.1016/j.ab.2018.12.019
11. Hussain, W., Khan, Y.D., Rasool, N., Khan, S.A. and Chou, K.C. (2019) SPrenylC-PseAAC: A Sequence-Based Model Developed via Chou's 5-Steps Rule and General PseAAC for Identifying S-prenylation Sites in Proteins. *Journal of Theoretical Biology*, **468**, 1-11. https://doi.org/10.1016/j.jtbi.2019.02.007
12. Khan, S., Khan, M., Iqbal, N., Hussain, T., Khan, S.A. and Chou, K.C. (2019) A Two-Level Computation Model Based on Deep Learning Algorithm for Identification of piRNA and Their Functions via Chou's 5-Steps Rule. International Journal of Peptide Research and Therapeutics. https://link.springer.com/article/10.1007%2Fs10989-019-09887-3

13. Khan, Z.U., Ali, F., Khan, I.A., Hussain, Y. and Pi, D. (2019) iRSpot-SPI: Deep Learning-Based Recombination Spots Prediction by Incorporating Secondary Sequence Information Coupled with Physio-Chemical Properties via Chou's 5-Step Rule and Pseudo Components. *Chemometrics and Intelligent Laboratory Systems*, **189**, 169-180. https://doi.org/10.1016/j.chemolab.2019.05.003

14. Lan, J., Liu, J., Liao, C., Merkler, D.J., Han, Q. and Li, J. (2019) A Study for Therapeutic Treatment against Parkinson's Disease via Chou's 5-Steps Rule. *Current Topics in Medicinal Chemistry*, **19**, 2318-2333. http://www.eurekaselect.com/175887/article

15. Le, N.Q.K. (2019) iN6-methylat (5-Step): Identifying DNA N(6)-methyladenine Sites in Rice Genome Using Continuous Bag of Nucleobases via Chou's 5-Step Rule. *Molecular Genetics and Genomics*: *MGG*, **294**, 1173-1182. https://doi.org/10.1007/s00438-019-01570-y

16. Le, N.Q.K., Yapp, E.K.Y., Ho, Q.T., Nagasundaram, N., Ou, Y.Y. and Yeh, H.Y. (2019) iEnhancer-5Step: Identifying Enhancers Using Hidden Information of DNA Sequences via Chou's 5-Step Rule and Word Embedding. *Analytical Biochemistry*, **571**, 53-61. https://doi.org/10.1016/j.ab.2019.02.017

17. Le, N.Q.K., Yapp, E.K.Y., Ou, Y.Y. and Yeh, H.Y. (2019) iMotor-CNN: Identifying Molecular Functions of Cytoskeleton Motor Proteins Using 2D Convolutional Neural Network via Chou's 5-Step Rule. *Analytical Biochemistry*, **575**, 17-26. https://doi.org/10.1016/j.ab.2019.03.017

18. Liang, R., Xie, J., Zhang, C., Zhang, M., Huang, H., Huo, H., Cao, X. and Niu, B. (2019) Identifying Cancer Targets Based on Machine Learning Methods via Chou's 5-Steps Rule and General Pseudo Components. *Current Topics in Medicinal Chemistry*, **19**, 2301-2317. https://doi.org/10.2174/1568026619666191016155543

19. Liang, Y. and Zhang, S. (2019) Identifying DNase I Hypersensitive Sites Using Multi-Features Fusion and F-Score Features Selection via Chou's 5-Steps Rule. *Biophysical Chemistry*, **253**, Article ID: 106227. https://doi.org/10.1016/j.bpc.2019.106227

20. Liu, Z., Dong, W., Jiang, W. and He, Z. (2019) csDMA: An Improved Bioinformatics Tool for Identifying DNA 6 mA Modifications via Chou's 5-Step Rule. *Scientific Reports*, **9**, Article No. 13109. https://doi.org/10.1038/s41598-019-49430-4

21. Malebary, S.J., Rehman, M.S.U. and Khan, Y.D. (2019) iCrotoK-PseAAC: Identify Lysine Crotonylation Sites by Blending Position Relative Statistical Features According to the Chou's 5-Step Rule. *PLoS ONE*, **14**, e0223993. https://doi.org/10.1371/journal.pone.0223993

22. Nazari, I., Tahir, M., Tayari, H. and Chong, K.T. (2019) iN6-Methyl (5-Step): Identifying RNA N6-Methyladenosine Sites Using Deep Learning Mode via Chou's 5-Step Rules and Chou's General PseKNC. *Chemometrics and Intelligent Laboratory Systems*, **193**, Article ID: 103811. https://doi.org/10.1016/j.chemolab.2019.103811

23. Ning, Q., Ma, Z. and Zhao, X. (2019) dForml(KNN)-PseAAC: Detecting Formylation Sites from Protein Sequences Using K-Nearest Neighbor Algorithm via Chou's 5-Step Rule and Pseudo Components. *Journal of Theoretical Biology*, **470**, 43-49. https://doi.org/10.1016/j.jtbi.2019.03.011

24. Tahir, M., Tayara, H. and Chong, K.T. (2019) iDNA6mA (5-Step Rule): Identification of DNA N6-Methyladenine Sites in the Rice Genome by Intelligent Computational Model via Chou's 5-Step Rule. *CHEMOLAB*, **189**, 96-101. https://doi.org/10.1016/j.chemolab.2019.04.007

25. Wiktorowicz, A., Wit, A., Dziewierz, A., Rzeszutko, L., Dudek, D. and Kleczynski, P. (2019) Calcium Pattern Assessment in Patients with Severe Aortic Stenosis via the Chou's 5-Steps Rule. *Current Pharmaceutical Design*, **25**, 3769-3775.

26. Yang, L., Lv, Y., Wang, S., Zhang, Q., Pan, Y., Su, D., Lu, Q. and Zuo, Y. (2019) Identifying FL11 Subtype by Characterizing Tumor Immune Microenvironment in Prostate Adenocarcinoma via Chou's 5-Steps Rule. *Genomics*, **112**, 1500-1515.

27. Charoenkwan, P., Schaduangrat, N., Nantasenamat, C., Piacham, T. and Shoombuatong, W. (2020) iQSP: A Sequence-Based Tool for the Prediction and Analysis of Quorum Sensing Peptides via Chou's 5-Steps Rule and Informative Physicochemical Properties. *International Journal of Molecular Sciences*, **21**, 75. https://doi.org/10.3390/ijms21010075

28. Dobosz, R., Mucko, J. and Gawinecki, R. (2020) Using Chou's 5-Step Rule to Evaluate the Stability of Tautomers: Susceptibility of 2-[(Phenylimino)-methyl]-cyclohexane-1,3-diones to Tautomerization Based on the Calculated Gibbs Free Energies. *Energies*, **13**, 183. https://doi.org/10.3390/en13010183

29. Ju, Z. and Wang, S.Y. (2020) Prediction of Lysine Formylation Sites Using the Composition of k-Spaced Amino Acid Pairs via Chou's 5-Steps Rule and General Pseudo Components. *Genomics*, **112**, 859-866. https://doi.org/10.1016/j.ygeno.2019.05.027

30. Kabir, M., Ahmad, S., Iqbal, M. and Hayat, M. (2020) iNR-2L: A Two-Level Sequence-Based Predictor Developed via Chou's 5-Steps Rule and General PseAAC for Identifying Nuclear Receptors and Their Families. *Genomics*, **112**, 276-285. https://doi.org/10.1016/j.ygeno.2019.02.006

31. Khan, Y.D., Amin, N., Hussain, W., Rasool, N., Khan, S.A. and Chou, K.C. (2020) iProtease-PseAAC(2L): A Two-Layer Predictor for Identifying Proteases and Their Types Using Chou's 5-Step-Rule and General PseAAC. *Analytical Biochemistry*, **588**, Article ID: 113477. https://doi.org/10.1016/j.ab.2019.113477

32. Akbar, S., Rahman, A.U. and Hayat, M. (2020) cACP: Classifying Anticancer Peptides Using Discriminative Intelligent Model via Chou's 5-Step Rules and General Pseudo Components. *Chemometrics and Intelligent Laboratory* (*CHEMOLAB*), **196**, Article ID: 103912. https://doi.org/10.1016/j.chemolab.2019.103912

33. Vishnoi, S., Garg, P. and Arora, P. (2020) Physicochemical n-Grams Tool: A Tool for Protein Physicochemical Descriptor Generation via Chou's 5-Step Rule. *Chemical Biology & Drug Design*, **95**, 79-86. https://doi.org/10.1111/cbdd.13617

34. Vundavilli, H., Datta, A., Sima, C., Hua, J., Lopes, R. and Bittner, M. (2020) Using Chou's 5-Steps Rule to Model Feedback in Lung Cancer. *IEEE Journal of Biomedical and Health Informatics*. https://doi.org/10.1109/JBHI.2019.2958042

35. Chou, K.C. (2019) Advance in Predicting Subcellular Localization of Multi-Label Proteins and Its Implication for Developing Multi-Target Drugs. *Current Medicinal Chemistry*, **26**, 4918-4943. http://www.eurekaselect.com/172010/article https://doi.org/10.2174/0929867326666190507082559

36. Chou, K.C. (2019) Two Kinds of Metrics for Computational Biology. *Genomics*. https://www.sciencedirect.com/science/article/pii/S0888754319304604?via%3Dihub

37. Chou, K.C. (2019) Proposing Pseudo Amino Acid Components Is an Important Milestone for Proteome and Genome Analyses. *International Journal for Peptide Research and Therapeutics*. https://link.springer.com/article/10.1007%2Fs10989-019-09910-7

38. Chou, K.C. (2019) An Insightful Recollection for Predicting Protein Subcellular Locations in Multi-Label Systems. *Genomics*. https://www.sciencedirect.com/science/article/pii/S0888754319304604?via%3Dihub

39. Chou, K.C. (2019) Progresses in Predicting Post-Translational Modification. *International Journal of Peptide Research and Therapeutics*. https://link.springer.com/article/10.1007%2Fs10989-019-09893-5

40. Chou, K.C. (2019) An Insightful Recollection since the Distorted Key Theory Was Born about 23 Years Ago. *Genomics*. https://www.sciencedirect.com/science/article/pii/S0888754319305543?via%3Dihub

41. Chou, K.C. (2019) Artificial Intelligence (AI) Tools Constructed via the 5-Steps Rule for Predicting Post-Translational Modifications. *Trends in Artificial Intelligence* (*TIA*), **3**, 60-74. https://doi.org/10.36959/643/304

42. Liu, B. (2018) BioSeq-Analysis: A Platform for DNA, RNA, and Protein Sequence Analysis Based on Machine Learning Approaches. *Briefings in Bioinformatics*, **20**, 1280-1294.

43. Liu, B., Gao, X. and Zhang, H. (2019) BioSeq-Analysis2.0: An Updated Platform for Analyzing DNA, RNA and Protein Sequences at Sequence Level and Residue Level Based on Machine Learning Approaches. *Nucleic Acids Research*, **47**, e127. https://doi.org/10.1093/nar/gkz740

44. Chou, K.C. (2019) The Cradle of Gordon Life Science Institute and Its Development and Driving Force. *Int J Biol Genetics*, **1**, 1-28.

45. Chou, K.C. (2019) Showcase to Illustrate How the Web-Server iDNA6mA-PseKNC Is Working. *Journal of Pathology Research Reviews & Reports*, **1**, 1-15.

46. Chou, K.C. (2019) The pLoc_bal-mPlant Is a Powerful Artificial Intelligence Tool for Predicting the Subcellular Localization of Plant Proteins Purely Based on Their Sequence Information. *International Journal of Nutrition Sciences*, **4**, 1037.

47. Chou, K.C., Cheng, X. and Xiao, X. (2019) pLoc_bal-mEuk: Predict Subcellular Localization of Eukaryotic Proteins by General PseAAC and Quasi-Balancing Training Dataset. *Medicinal Chemistry*, **15**, 472-485. https://doi.org/10.2174/1573406415666181218102517

48. Chou, K.C. (2019) Showcase to Illustrate How the Web-Server iNitro-Tyr Is Working. *Glo J of Com Sci and Infor Tec.*, **2**, 1-16.

49. Chou, K.C. (2019) The pLoc_bal-mAnimal Is a Powerful Artificial Intelligence Tool for Predicting the Subcellular Localization of Animal Proteins Based on Their Sequence Information Alone. *Scientific Journal of Biometrics & Biostatistics*, **2**, 1-13.

50. Chou, K.C. (2020) Showcase to Illustrate How the webserver pLoc_bal-mEuk Is Working. *Biomedical Journal of Scientific & Technical Research*.

51. Chou, K.C. (2020) The pLoc_bal-mGneg Predictor Is a Powerful Web-Server for Identifying the Subcellular Localization of Gram-Negative Bacterial Proteins Based on Their Sequences Information Alone. *ijSci*, **9**, 27-34. https://doi.org/10.18483/ijSci.2248

52. Chou, K.C. (2020) How the Artificial Intelligence Tool iRNA-2methyl Is Working for RNA 2'-Omethylation Sites. *Journal of Medical Care Research and Review*, **3**, 348-366.

53. Chou, K.-C. (2020) Showcase to Illustrate How the Web-Server iKcr-PseEns Is Working. *Journal of Medical Care Research and Review*, **3**, 331-347.

54. Chou, K.C. (2020) The pLoc_bal-mVirus Is a Powerful Artificial Intelligence Tool for Predicting the Subcellular Localization of Virus Proteins According to Their Sequence Information Alone. *J Gent & Genome*, **4**.

55. Chou, K.C. (2019) How the Artificial Intelligence Tool iSNO-PseAAC Is Working in Predicting the Cysteine S-nitrosylation Sites in Proteins. *Journal of Stem Cell Research and Medicine*, **4**, 1-9.

56. Chou, K.C. (2020) Showcase to Illustrate How the Web-Server iRNA-Methyl Is Working. *J Mol Genet*, **3**, 1-7.

57. Chou, K.C. (2020) How the Artificial Intelligence Tool iRNA-PseU Is Working in Predicting the RNA Pseudouridine Sites. *Biomedical Journal of Scientific & Technical Research*.

58. Chou, K.C. (2020) Showcase to Illustrate How the Web-Server iSNO-AAPair Is Working. *J Gent & Genome*, **4**.

59. Chou, K.C. (2020) The pLoc_bal-mHum Is a Powerful Web-Serve for Predicting the Subcellular Localization of Human Proteins Purely Based on Their Sequence Information. *Advances in Bioengineering and Biomedical Science Research*, **3**, 1-5.