Advancement in Scientific and Engineering Research
Vol. 4(2), pp. 31-36, October 2019
doi: 10.33495/aser_v4i2.19.105
ISSN: 2384-7336
Review

An insightful recollection since the birth of Gordon Life Science Institute about 17 years ago

Kuo-Chen Chou[1,2]

[1]Gordon Life Science Institute, Boston, Massachusetts 02478, USA
[2]Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu, 610054, China.

*Corresponding E-mail: kcchou38@gmail.com

Accepted 15 October, 2019.

Abstract. Gordon Life Science Institute is the first Internet Research Institute ever established in the world. Recollected in this minireview is its establishing and developing processes, as well as its philosophy and accomplishments.

Keywords: Reform and opening, free communication; Sweden, cradle; San Diego, Boston, door-opening.

Running Title: Gordon Life Science Institute

INTRODUCTION

The Gordon Life Science Institute was established in 2003 at San Diego of California, USA. Its founder is Professor Dr. Kuo-Chen Chou. Its mission is to develop and apply new mathematical tools and physical concepts for understanding biological phenomena. For more detailed about the growth process of Gordon Life Science Institute as well as its novel philosophy, see an article at https://gordonlifescience.org/GordonLifeScience.html.

The Gordon Life Science Institute is a newly emerging academic organization in the Age of Information and Internet. Founded by Professor Dr. Kuo-Chen Chou at San Diego of California, right after he was retired from Pfizer Global Research and Development in 2003. Its mission is to develop and apply new mathematical tools and physical concepts for understanding biological phenomena.

The Institute's name reflects an interesting historical story. After the Cultural Revolution, China started to open its door, the founder was invited by Professor Sture Forsén, the Chairman of Nobel Prize Committee, to work in Chemical Center of Lund University as a Visiting Professor. To make Swedish people easier to pronounce his name, Professor Chou used "Gordon" as his name in Sweden. About a quarter of century later, the same name was used for the Institute, meaning that "Reform and Opening" (改革开放) and "Free Communication" (自由交换信息) can stimulate a lot of great creativities. The current liaison site of Gordon Life Science Institute is in Boston of Massachusetts, USA; gls@gordonlifescience.org.

MISSION AND ORGANIZATION

The Institute has no physical boundaries. Its members do not have to work in a same building or campus. Distributed over different countries of the world, they shall freely collaborate, exchange ideas, and share information and findings via a variety of modern communication methods. This versatile system allows the members to focus completely on science without having to cope with troubles in obtaining visas and in paying for relocation

expenses, among many others.

The Gordon Life Science Institute is a non-profit organization. It is a gift to science and human beings. Its founding principle is to pursue the excellence in science: anyone who has proved his/her creativity in science can become a member regardless of his/her age, occupation, and nationality. Accordingly, the Institute has provided an ideal society or organization for those scientists who are really dedicated themselves to science and loving science more than anything else. In the friendly door-opened Institute, these scientists can maximize their time and energy to engage in their scientific creativity. Members of the Institute believe science would be more truthful and wonderful if scientists do not have to spend a lot of time on funding application. We also note that great scientific findings and creations in history were often made by those who were least supported or funded but driven by interesting imagination and curiosity.

Accomplishments

Up to March 2019, the Institute has 26 members. Among them 5 have been selected by Thompson Reuter and Clarivate Analytics as the "Highly Cited Researcher": (**1**) Kuo-Chen Chou for continuously 5 years (2014, 2015, 2016, 2017, and 2018), (**2**) Hong-Bin Shen (2014 and 2015), (**3**) Wei Chen (2018), (**4**) Hao Lin (2018), and (**5**) Xoan Xiao (2018). Listed below are just some represented works produced by the Gordon Life Science Institute.

Extension of special PseAAC to the general one

With the explosive growth of biological sequences in the post-genomic era, one of the most challenging problems in computational biology is how to express a biological sequence with a discrete model or a vector, yet still keep considerable sequence-order information or key pattern characteristic. This is because all the existing machine-learning algorithms (such as "Optimization" algorithm (Zhang, 1992), "Covariance Discriminant" or "CD" algorithm (Chou and Elrod, 2002), (Chou and Cai, 2003), "Nearest Neighbor" or "NN" algorithm (Hu, et al., 2011), and "Support Vector Machine" or "SVM" algorithm (Hu, et al., 2011, Cai, et al., 2006) can only handle vectors as elaborated in a comprehensive review (Chou, 2015). However, a vector defined in a discrete model may completely lose all the sequence-pattern information. To avoid completely losing the sequence-pattern information for proteins, the pseudo amino acid composition (Chou, 2001) or PseAAC (Chou, 2005) was proposed. Ever since then, it has been widely used in nearly all the areas of computational proteomics (Chou and Cai, 2003, Guo 2002, Georgiou et al., 2009) (Ding et al., 2009, Gao et al, 2009, Li et al., 2009, Xiao et al ., 2018). Because it has been widely and increasingly used, four powerful open access soft-wares, called 'PseAAC' (Shen, 2008), 'PseAAC-Builder' (Du et al., 2012), 'propy' (Cao et al., 2013) and 'PseAAC-General' (Du et al., 2014), were established: the former three are for generating various modes of Chou's special PseAAC (Chou, 2009) while the 4th one for those of Chou's general PseAAC (Chou, 2011), 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 (Chou, 2011), "Gene Ontology" mode (see Eqs.11-12 of (Chou, 2011), and "Sequential Evolution" or "PSSM" mode (see Eqs.13-14 of (Chou, 2011). For more information about the PseAAC, please visit an insightful Wikipedia article at https://en.wikipedia.org/wiki/Pseudo_amino_acid_composition.

Extension of PseAAC to PseKNC

Encouraged by the successes of using PseAAC to deal with protein/peptide sequences, the concept of PseKNC (Pseudo K-tuple Nucleotide Composition) (Chen et al., 2014) was developed for generating various feature vectors for DNA/RNA sequences that have proved very useful as well (Chen et al., 2014). Particularly, in 2015 a very powerful web-server called 'Pse-in-One' and its updated version 'Pse-in-One2.0' have been established that can be used to generate any desired feature vectors for protein/peptide and DNA/RNA sequences according to the need of users' studies". For more information about the PseKNC, please visit an insightful Wikipedia article at https://en.wikipedia.org/wiki/Pseudo_K-tuple_nucleotide_composition.

Distorted key theory for peptide drugs

According to Fisher's "lock and key" model, Koshland's "induced fit" theory, and the "rack mechanism", the prerequisite condition for a peptide to be cleaved by the disease-causing enzyme is a good fit and tightly binding with the enzyme's active site (Figure 1). However, such a peptide, after a modification on its scissile bond with some simple chemical procedure, will no longer be cleavable by the enzyme but it can still tightly bind to its active site. An illustration about the distorted key theory is given in Figure 2, where panel (**a**) shows an effective binding of a cleavable peptide to the active site of HIV protease, while panel (**b**) the peptide has become a non-cleavable one after its scissile bond is modified although it can still bind to the active site. Such a modified peptide, or ''distorted key", will automatically become an inhibitor candidate against HIV protease. Even for non-peptide inhibitors, the information derived from the cleavable peptides can also provide useful insights about the key binding groups and fitting conformation in the sense of

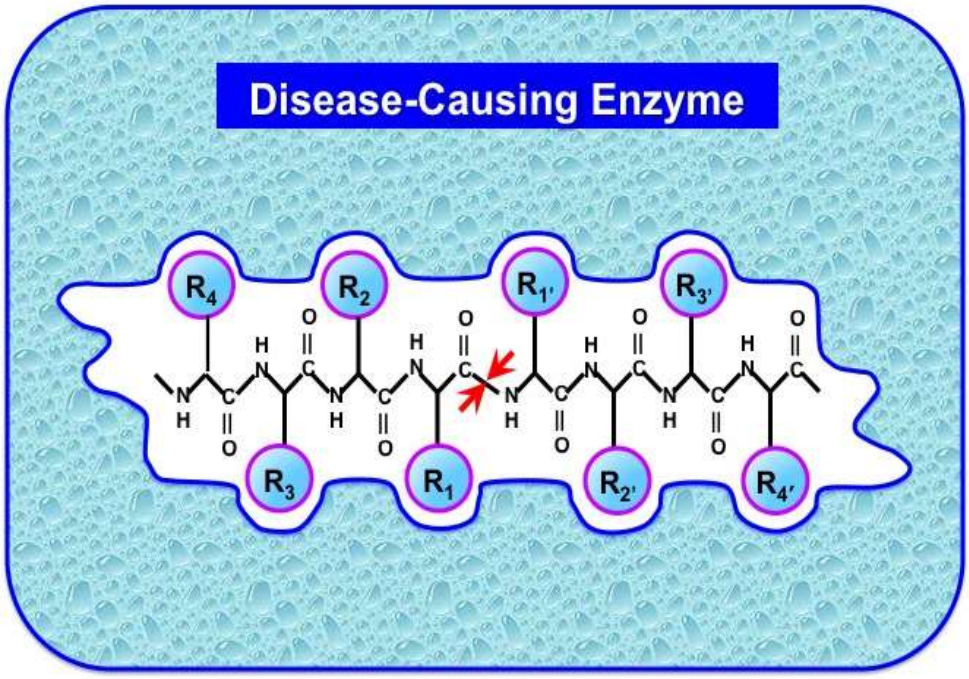

Figure 1. A schematic illustration to show a peptide in good fitting and tightly binding with the enzyme's active site before it is cleaved by the latter. Adapted from (Chou, 1996) with permission.

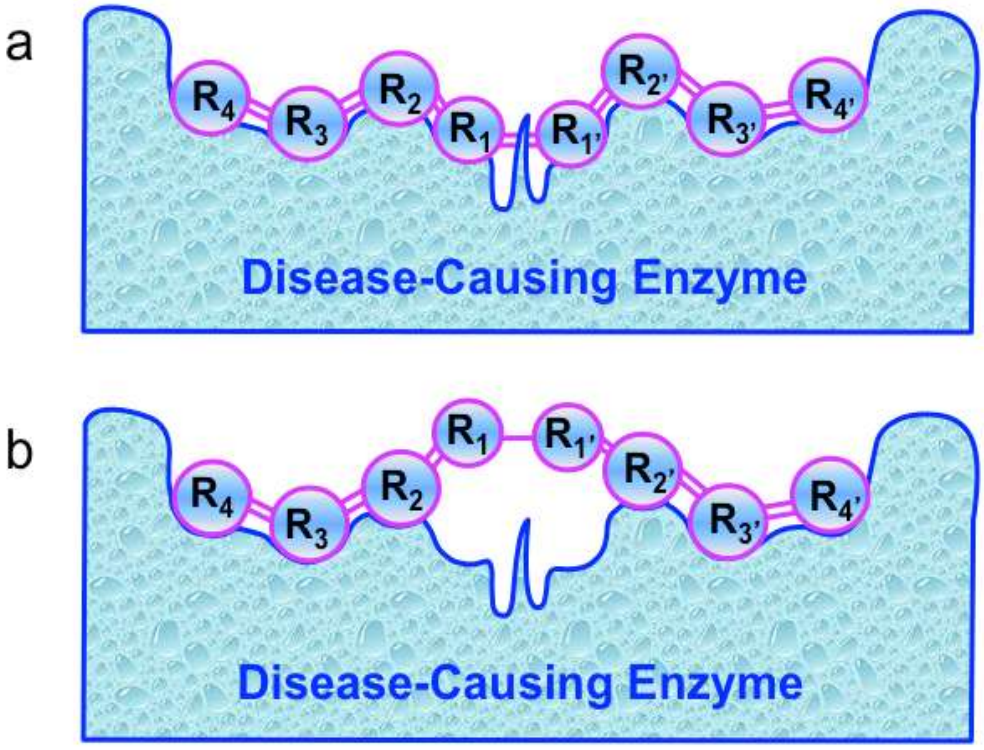

Figure 2. Schematic drawing to illustrate the "Distorted Key" theory, where panel (**a**) shows an effective binding of a cleavable peptide to the active site of a disease-causing enzyme, while panel (**b**) the same peptide has become a non-cleavable one after its scissile bond is modified although it can still bind to the active site. Such a modified peptide, or ''distorted key", will automatically become an inhibitor candidate against the disease-causing enzyme. Adapted from with permission.

microenvironment. Besides, peptide drugs usually have no toxicity in vivo under the physiological concentration. For more discussion about the distorted key theory, see a comprehensive review paper. It was based on such a distorted key theory that many investigators were enthusiastic to develop various methods for predicting the protein cleavage sites by disease-causing. Furthermore, a web-server called "HIVcleave" (Shen, 2008) has been established for predicting HIV protease cleavage sites in proteins. Its website address is at http://chou.med.harvard.edu/bioinf/HIV/. For more discussions about the "distorted key theory", see an insightful Wikipedia article at https://en.wikipedia.org/wiki/Chou%27s_distorted_key_theory_for_peptide_drugs.

Introduction of wenxiang diagram

Using graphic approaches to study biological and medical systems can provide an intuitive vision and useful insights for helping analyze complicated relations therein, as indicated by many previous studies on a series of important biological topics**.** Its importance can also be seen in an insightful Wikipedia Article at https://en.wikipedia.org/wiki/Graph_theory_in_enzymatic_kinetics. Wenxiang diagram] is a special kind of graphical approach, which is very useful for in-depth studying protein-protein interaction mechanism. For more about the wenxiang diagram, see an insightful Wikipedia article at https://en.wikipedia.org/wiki/Wenxiang_diagram.

Predictors for multi-label systems

Information of subcellular localization for a protein is indispensable for revealing its biological function. Therefore, one of the fundamental goals in molecular cell biology and proteomics is to determine the subcellular locations of proteins in an entire cell. Before 2007, most efforts in this regard were focused on the single-label system by assuming that each of the constitute proteins in a cell had one, and only one, subcellular.

However, with more experimental data uncovered, it has been found that many proteins may simultaneously occur or move between two or more location sites in a cell and hence need multiple labels to mark them. Proteins with multiple locations are also called multiplex proteins, which are often the special targets for drug development. Therefore, how to deal with this kind of multi-label systems is a critical challenge. To take the challenge, the Institute has developed the following four series of predictors. All these predictors have yielded very high success rates, both globally and locally, as summarized in a comprehensive review paper. For more about protein subcellular localization prediction, see an insightful Wikipedia article at https://en.wikipedia.org/wiki/Protein_subcellular_localization_prediction.

Five-step rule

The Institute was the birth place of the famous 5-steps rule (Chou 2011), which has been used in nearly all the areas of computational biology (Cheng and Xiao, 2017, Cheng and Xiao, 2017, Xiao et al., 2017, Cheng et al., 2018, Chou et al., 2019, Ghauri et al., 2018, Awais et al., 2019, Cheng et al., 2019, Hussain et al., 2019, Kabir et al., 2019, Ning et al., 2019, Wang et al., 2019, Xiao et al., 2019, Lin et al., 2014, Liu et al., 2016, Feng et al., 2019, Cheng et al., 2017, Zhang et al. , 2018), material science (Zhai et al.,2018), and even the commercial science (e.g., the bank account systems). The only difference between them is how to formulate the statistical samples or events with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted. It just like the case of many machine-learning algorithms. They can be widely used in nearly all the areas of statistical analysis. For more about protein subcellular localization prediction, see an insightful Wikipedia article at https://en.wikipedia.org/wiki/5-step_rules

Working in that Institute filled with this kind of philosophy and atmosphere, the scientists would be more prone to be stimulated by the eight pioneering papers from the then Chairman of Nobel Prize Committee Sture Forsen (Chou, S. Forsen, 1980, Chou, S. Forsen, 1981) and many of their follow-up papers (Jia et al., 2015, Jia et al., 2016, Zhou, 2011, Chou, 2019, Chou, 1983, Chou, 2019), so as to render them substantially more creative and productive.

CONCLUSION AND PERSPECTIVE

In comparison with the conventional institutes, Gordon Life Science Institute has the following unique advantages: it can (1) attract those scientists who are really loving science more than anything else; (2) maximize their creativity in science and minimize the distraction or disturbance caused by the relocation and various followed-up tedious things; (3) provide them with an ideal environment to completely focus on doing science; (4) drive their motivation by interesting imagination and curiosity; and (5) guide their scientific results more truthful and wonderful.

Accordingly, it would not be surprised to see that a total of five members of Gordon Life Scientist have been selected by Clarivate Analytics as Highly Cited Researcher or HCR (https://hcr.clarivate.com/resources/archived-lists/), indicating that for the ratio of HCR per member, the "Gordon Life Science Institute" has already become the top one in the world. It is expected that more significant accomplishments will be achieved by the Gordon Life Science Institute for many years to come.

REFERENCES

Awais M, Hussain W, Khan YD, Rasool N, Khan SA (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 Trans. Comput. Biol. Bioinform doi:10.1109/TCBB.2019.2919025 (2019).

Cai YD, Feng KY, Lu WC (2006). Using LogitBoost classifier to predict protein structural classes. J. Theo. Biol. 238: 172-176.

Cao DS, Xu QS, Liang YZ (2013). propy: a tool to generate various modes of Chou's PseAAC. Bioinformatics 29: 960-962.

Cheng X, Lin WZ, Xiao X (2019). pLoc_bal-mAnimal: predict subcellular localization of animal proteins by balancing training dataset and PseAAC. Bioinformatics 35: 398-406.

Cheng X, Xiao X (2017). pLoc-mPlant: predict subcellular localization of multi-location plant proteins via incorporating the optimal GO information into general PseAAC. Molecular BioSystems 13:1722-1727.

Cheng X, Xiao X (2017). pLoc-mVirus: predict subcellular localization of multi-location virus proteins via incorporating the optimal GO information into general PseAAC. Gene (Erratum: ibid., 2018, Vol. 644, 156-156) 628 (2017): 315-321.

Cheng X, Xiao, X (2018). pLoc-mEuk: Predict subcellular localization of multi-label eukaryotic proteins by extracting the key GO information into general PseAAC. Genomics 110 (2018): 50-58.

Cheng X, Zhao SG, Lin WZ, Xiao X (2017). pLoc-mAnimal: predict subcellular localization of animal proteins with both single and multiple sites. Bioinformatics 33:3524-3531.

Chou KC (1983). Low-frequency vibrations of helical structures in protein molecules. Biochem. J. 209: 573-580.

Chou KC (1996). Review: Prediction of human immunodeficiency virus protease cleavage sites in proteins. Analytical Biochem. 233:1-14.

Chou KC (2001). Prediction of protein cellular attributes using pseudo amino acid composition. PROTEINS: Structure, Function, and Genetics (Erratum: ibid., 2001, Vol. 44, 60) 43 (2001) 246-255.

Chou KC (2005). Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21: 10-19.

Chou KC (2009). Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. Current Proteomics 6:262-274.

Chou KC (2011). Some remarks on protein attribute prediction and pseudo amino acid composition (50th Anniversary Year Review, 5-steps rule). J. Theo. Biol. 273: 236-247.

Chou KC (2015). Impacts of bioinformatics to medicinal chemistry. Medicinal Chem. 11: 218-234.

Chou KC (2019). Progresses in predicting post-translational modification. Int. J. Peptide Res. Therapeut. DOI: 10.1007/s10989-019-09893-5.

Chou KC (2019). Recent progresses in predicting protein subcellular localization with artificial intelligence tools developed via the 5-steps rule. Medicinal Chemistry Submitted.

Chou KC, Cai YD (2003). Prediction and classification of protein subcellular location: sequence-order effect and pseudo amino acid composition. J. Cell. Biochem. (Addendum, ibid. 2004, 91, 1085) 90 (2003) 1250-1260.

Chou KC, Cheng X, Xiao X (2019). pLoc_bal-mEuk: predict subcellular localization of eukaryotic proteins by general PseAAC and quasi-balancing training dataset. Med. Chem. 15 (2019) 472-485.

Chou KC, Elrod DW (2002). Bioinformatical analysis of G-protein-coupled receptors. J. Proteome Res. 1:429-433.

Chou KC, Forsen S (1980). Graphical rules for enzyme-catalyzed rate laws. Biochem. J. 187: 829-835.

Chou KC, Forsen SU(1981). Graphical rules of steady-state reaction systems. Can. J. Chem. 59: 737-755.

Ding H, Luo L, Lin H (2009). Prediction of cell wall lytic enzymes using Chou's amphiphilic pseudo amino acid composition. Protein & Peptide Letters 16: 351-355.

Du P, Gu S, Jiao Y (2014). PseAAC-General: Fast building various modes of general form of Chou's pseudo amino acid composition for large-scale protein datasets. Int. J. Molecular Sci. 15: 3495-3506.

Du P, Wang X, Xu C, Gao Y (2012). PseAAC-Builder: A cross-platform stand-alone program for generating various special Chou's pseudo amino acid compositions. Analytical Biochem. 425: 117-119.

Feng P, Yang H, Ding H, Lin H, Chen W (2019). iDNA6mA- PseKNC: Identifying DNA N(6)-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC. Genomics 111: 96-102.

Gao QB, Jin ZC, Ye ZF, Wu C, He J (2009). Prediction of nuclear receptors with optimal pseudo amino acid composition. Anal. Biochem. 387: 54-59.

Georgiou DN, Karakasidis TE, Nieto JJ, Torres A (2009). Use of fuzzy clustering technique and matrices to classify amino acids and its impact to Chou's pseudo amino acid composition. J. Theo. Biol. 257: 17-26.

Ghauri AW, Khan YD, Rasool N, Khan SA (2018). pNitro-Tyr-PseAAC: Predict nitrotyrosine sites in proteins by incorporating five features into Chou's general PseAAC. Curr Pharm Des 24: 4034-4043.

Guo ZM (2002). Prediction of membrane protein types by using pattern recognition method based on pseudo amino acid composition. Master Thesis, Bio-X Life Science Research Center, Shanghai Jiaotong University.

Hu L, Huang T, Shi X, Lu WC, Cai, YD (2011). Predicting functions of proteins in mouse based on weighted protein-protein interaction network and protein hybrid properties PLoS ONE 6: e14556.

Hussain W, Khan SD, Rasool N, Khan SA (2019). SPalmitoylC-PseAAC: A sequence-based model developed via Chou's 5-steps rule and general PseAAC for identifying S-palmitoylation sites in proteins. Anal. Biochem. 568:14-23.

Jia J, Liu Z, Xiao X (2015). iPPI-Esml: an ensemble classifier for identifying the interactions of proteins by incorporating their physicochemical properties and wavelet transforms into PseAAC. J. Theo. Biol. 377: 47-56.

Jia J, Liu Z, Xiao X (2016). Identification of protein-protein binding sites by incorporating the physicochemical properties and stationary wavelet transforms into pseudo amino acid composition (iPPBS-PseAAC). J. Biomol. Struct. Dyn. (JBSD) 34: 1946-1961.

Kabir M, Ahmad SA, Iqbal M, Hayat M (2019). 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 doi:10.1016/j.ygeno.2019.02.006.

Li ZC, Zhou XB, Dai Z, Zou XY (2009). Prediction of protein structural classes by Chou's pseudo amino acid composition: approached using continuous wavelet transform and principal component analysis. Amino Acids 37: 415-425.

Lin H, Deng EZ, Ding H, Chen W (2014). iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. Nucleic Acids Res. 42:12961-12972.

Liu B, Fang L, Long R, Lan X (2016). iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition. Bioinformatics 32: 362-369.

Ning Q, Ma Z, 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. J. Theor. Biol. 470: 43-49.

Shen HB (2008). HIVcleave: a web-server for predicting HIV protease cleavage sites in proteins. Anal. Biochem. 375: 388-390.

Shen HB (2008). PseAAC: a flexible web-server for generating various kinds of protein pseudo amino acid composition. Anal. Biochem. 373:386-388.

Wang L, Zhang R, Mu Y (2019). Fu-SulfPred: Identification of Protein S-sulfenylation Sites by Fusing Forests via Chou's General PseAAC. J Theor. Biol. 461: 51-58.

Xiao X, Cheng X, Chen G, Mao Q (2018). pLoc_bal-mVirus: Predict Subcellular Localization of Multi-Label Virus Proteins by Chou's General PseAAC and IHTS Treatment to Balance Training Dataset. Med Chem 15: 496-509.

Xiao X, Cheng X, Chen G, Mao Q (2019). pLoc_bal-mGpos: predict subcellular localization of Gram-positive bacterial proteins by quasi-balancing training dataset and PseAAC. Genomics 111: 886-892.

Xiao X, Cheng X, Su S, Nao Q (2017). pLoc-mGpos: Incorporate key gene ontology information into general PseAAC for predicting

subcellular localization of Gram-positive bacterial proteins. Natural Sci. 9: 331-349.

Zhai X, Chen M, Lu W (2018). Accelerated search for perovskite materials with higher Curie temperature based on the machine learning methods. Computational Materials Sci. 151: 41-48.

Zhang CT (1992). An optimization approach to predicting protein structural class from amino acid composition. Protein Science 1:401-408.

Zhang Y, Xie R, Wang J, Leier A, Marquez-Lago TT, Akutsu T, Webb GI, Song J (2018). Computational analysis and prediction of lysine malonylation sites by exploiting informative features in an integrative machine-learning framework. Brief in Bioinform doi: 10.1093/bib/bby079.