中国在翻译后修饰的生物信息学研究领域的进展与前瞻

刘泽先[1]，蔡煜东[2]，郭雪江[3]，李骜[4]，李婷婷[5]，邱建丁[6]，任间[7]，施绍萍[8]，宋江宁[9]，王明会[4]，谢鹭[10]，薛宇[1]，张子丁[11]，赵兴明[12]

1. 华中科技大学生命科学与技术学院生物医学工程系，武汉 430074；

2. 上海大学系统生物技术研究所，上海 200444；

3. 南京医科大学组织胚胎学系，生殖医学国家重点实验室，南京 210029；

4. 中国科学技术大学信息科学与技术学院，合肥 230027；

5. 北京大学医学部基础医学院医学信息学系，北京 100191；

6. 南昌大学化学学院化学系，南昌 330031；

7. 中山大学生命科学学院，生物防治国家重点实验室，广州 510275；

8. 南昌大学理学院数学系，南昌 330031；

9. 中国科学院天津工业生物技术研究所，工业酶国家工程实验室与系统微生物工程重点实验室，天津 300308；

10. 上海科学院，上海生物信息技术研究中心，上海 2012013；

11. 中国农业大学生物学院，农业生物技术国家重点实验室，北京 100193；

12. 同济大学电子与信息工程学院计算机系，上海 201804

摘要：翻译后修饰在调控蛋白质构象变化、活性以及功能方面具有重要作用，并参与了几乎所有细胞通路和过程。蛋白质翻译后修饰的鉴定是阐明细胞内分子机理的基础。相对于劳动密集的、耗费时间的实验工作，利用各种生物信息学方法开展翻译后修饰预测，能够提供准确、简便和快速的研究方案，并产生有价值的信息为进一步实验研究提供参考。文章主要综述了中国生物信息学者在翻译后修饰生物信息学领域所取得的研究进展，包括修饰底物与位点预测的计算方法学设计与完善、在线或本地化工具的设计与维护、修饰相关数据库及数据资源的构建及基于修饰蛋白质组学数据的生物信息学分析。通过比较国内外的同类研究，发现优势和不足，并对未来的研究作出前瞻。

收稿日期: 2014-12-31；**修回日期:** 2015-02-05

基金项目: 国家自然科学基金项目（编号：31471252，31371337，81222006，61471331），国家重点基础研究发展规划(973 计划)项目 （编号：2011CB910600）和国家高技术研究发展计划（863 计划）项目（编号：2012AA020201）资助

作者简介: 刘泽先，博士，研究方向：生物信息学与计算蛋白质组学。E-mail: lzx@hust.edu.cn

通讯作者: 薛宇，博士，教授，研究方向：生物信息学与计算蛋白质组学。E-mail: xueyu@hust.edu.cn

关键词：翻译后修饰；共价修饰；修饰组学；磷酸化

Post-translational modification (PTM) bioinformatics in China: progresses and perspectives

Zexian Liu[1], Yudong Cai[2], Xuejiang Guo[3], Ao Li[4], Tingting Li[5], Jianding Qiu[6], Jian Ren[7], Shaoping Shi[8], Jiangning Song[9], Minghui Wang[4], Lu Xie[10], Yu Xue[1], Ziding Zhang[11], Xing-Ming Zhao[12]

1. *Department of Biomedical Engineering, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China;*

2. *Institute of Systems Biology, Shanghai University, Shanghai 200444, China;*

3. *State Key Laboratory of Reproductive Medicine, Department of Histology and Embryology, Nanjing Medical University, Nanjing 210029, China;*

4. *School of Information Science and Technology, University of Science and Technology of China, Hefei 232207, China;*

5. *Department of Biomedical Informatics, School of Basic Medical Sciences, Peking University Health Science Center, Beijing 100191, China;*

6. *Department of Chemistry, School of chemistry, Nanchang University, Nanchang 330031, China；*

7. *State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen University, Guangzhou 510275, China;*

8. *Department of Mathematics, college of science, Nanchang University, Nanchang 330031, China;*

9. *National Engineering Laboratory for Industrial Enzymes and Key Laboratory of Systems Microbial Biotechnology, Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, Tianjin 300308, China;*

10. *Shanghai Center for Bioinformation Technology, Shanghai Academy of Science and Technology, Shanghai 201203, China;*

11. *State Key Laboratory of Agrobiotechnology, College of Biological Sciences, Beijing 100193, China;*

12. *Department of Computer Science, School of Electronics and Information Engineering, Tongji University, Shanghai 201804, China*

Abstract: Post-translational modifications (PTMs) are essential for regulating conformational changes, activities and functions of proteins, and are involved in almost all cellular pathways and processes. Identification of protein PTMs is the basis for understanding cellular and molecular mechanisms. In contrast with labor-intensive and time-consuming experiments, the PTM prediction using various bioinformatics approaches can provide accurate, convenient, and efficient strategies and generate valuable information for further experimental consideration. In this review, we summarize the current progresses made by Chineses bioinformaticians in the field of PTM Bioinformatics, including the design and improvement of computational algorithms for predicting PTM substrates and sites, design and maintenance of online and offline tools, establishment of PTM-related databases and resources, and bioinformatics analysis of PTM proteomics data. Through comparing similar studies in China and other countries, we demonstrate both advantages and limitations of current PTM bioinformatics as well as perspectives for future studies in China.

Keywords: post-translational modification; covalent modification; PTM proteomics; phosphorylation

细胞是生命活动的最小单元，蛋白质是构成细胞的基本要素。蛋白质翻译后修饰（Post-translational modification, PTM）是调控蛋白质功能的重要机制，在生物学过程和信号通路中发挥着不可替代的作用，并可逆地决定了细胞的动力学和可塑性[1]。翻译后修饰通常发生在编码基因 DNA 序列转录为 mRNA，并翻译成蛋白质序列之后。由于翻译后修饰过程的生化机理为共价键的形成或断裂，因此翻译后修饰也称为共价修饰（Covalent modification）[1]。翻译后修饰可发生在特定氨基酸的侧链上，如常见的磷酸化（Phosphorylation）[2]、泛素化（Ubiquitination）[3]和乙酰化（Acetylation）[4]等，也可发生在蛋白质序列的主链上，如蛋白酶切（Proteolytic cleavage）和内含肽剪切（Intein splicing）等[5,6]。翻译后修饰过程通常受到特定酶（Enzyme）的催化，如蛋白激酶（Protein kinase）催化磷酸化反应[7]，E1 泛素活化酶（Ubiquitin-activating enzyme）、E2 泛素结合酶（Ubiquitin-conjugating enzyme）和 E3 泛素连接酶（Ubiquitin ligase）协同催化泛素化过程[8]，而组蛋白乙酰转移酶（Histone acetyltransferase, HAT）则催化乙酰化修饰[9]。修饰底物也能在特定酶的作用下发生去修饰反应，如蛋白磷酸酶（Protein phosphatase）催化去磷酸化[7]，去泛素酶（Deubiquitinating enzyme, DUB）执行去泛素化反应[8]，而组蛋白去乙酰化酶（Histone deacetylase, HDAC）则催化去乙酰化过程[9]。系统分析发现，人类基因组中存在 516 个编码激酶的基因和 160 个编码磷酸酶的基因[7]，以及 720 个 E1、E2、E3 泛素酶和 126 个去泛素酶基因[8]。研究表明，不同的酶通常仅调控有限的、特定的底物，并且具有特定的底物识别特异性[2]。由于去修饰酶的数量较少，其识别底物的特异性较低，因此目前研究者的兴趣主要集中于修饰酶催化的修饰过程，对于去修饰过程的关注则相对较少。

在构成蛋白质序列的 20 种常见氨基酸中，有 15 种氨基酸上可发生修饰，而 5 种疏水性氨基酸——亮氨酸（Leucine, L）、异亮氨酸（Isoleucine, I）、缬氨酸（Valine, V）、丙氨酸(Alanine, A)和苯丙氨酸(Phenylalanine, F)尚未发现修饰的存在[10]。蛋白质翻译后修饰的种类很多，已鉴定并被认可的修饰约有 470 种（http://www.uniprot.org/docs/ptmlist）。虽然每一种修饰都可能具有潜在的调控功能，但不同的修饰往往倾向参与特定的生物学过程。例如，磷酸化更倾向于参与细胞信号转导通路与细胞周期，泛素化则主要通过促进底物的降解来调控蛋白的生命周期，而近期的研究表明乙酰化除了参与转录调控，也在细胞代谢过程中发挥重要的作用（表 1）。常见的几种翻译后修饰及其主要调控功能可参见表 1。

传统的实验学方法，例如通过定点突变潜在的修饰位点，利用体外或体内的实验方法鉴定修饰底物和位点，需要耗费大量的时间、精力和资金。近年来，基于高通量质谱等技术的修饰组学研究发展迅速，但由于实验步骤繁琐、生物样本不稳定等因素，修饰组学鉴定的可重复性仍有待提高。因此，

表 1 12 种常见的蛋白质共价修饰的相关信息以及主要的调控功能

翻译后修饰	中文名称	主要识别位点/模体	功能
Phosphorylation[2]	磷酸化	S/T, Y	普遍
Uiquitination[3]	泛素化	K	蛋白质降解、信号转导
Glycosylation[11]	糖基化	N, W	蛋白质折叠、细胞黏附
Acetylation[4]	乙酰化	K	转录调控、DNA 损伤修复
Nitration[12]	硝基化	Y	蛋白质功能损伤
Sumoylation[13]	SUMO 化	K	转录调控、信号转导
Palmitoylation[14]	棕榈酰化	C	信号转导、细胞凋亡
Methylation[15]	甲基化	R, K	转录调控、信号转导
Sulfation[16]	硫化	Y	类固醇合成、异物代谢
S-nitrosylation[17]	亚硝基化	C	信号转导、钙循环
Caspase cleavage[5]	Caspase 酶切	DXXD↓	细胞凋亡
Calpain cleavage[6]	Calpain 酶切	L/VX↓	细胞周期、胃酸分泌

发展生物信息学方法，能够准确、快速、有效地预测潜在的修饰底物和位点，为进一步的实验研究提供有价值的参考信息[18~20]。在最近几年的工作中，本文作者已详细评述了国际翻译后修饰生物信息学领域的研究进展[18~20]：介绍并总结了 22 个磷酸化相关的数据资源与数据库，30 个磷酸化底物、位点、磷酸化结合模体（Phospho-binding motif）的预测工具[19]；评述了基于磷酸化蛋白质组学数据的计算分析[18]，包括磷酸化模体的发现、磷酸化网络的系统模拟、遗传变异影响磷酸化的识别，以及磷酸化的分子进化等；重点总结了基于磷酸化组学的网络医学（Phosphoproteomics-based network medicine）的研究进展[20]，包括从尚未注释的磷酸化组学数据中重建磷酸化介导的信号网络、网络磷酸化分子标记（Network phospho-signature）的发现，以及利用这些分子标记来分类癌症和预测药物反应（Drug response）等。近期，我们收集了国际和国内翻译后修饰相关的 230 多个数据库和计算工具，并予以简介和推荐（http://www.biocuckoo.org/link.php）。

本文主要介绍了中国在翻译后修饰的生物信息学领域的研究进展，包括翻译后修饰位点预测的计算方法学研究，翻译后修饰底物及位点预测的在线工具、软件和计算平台的设计，翻译后修饰相关数据资源与数据库的构建，以及基于磷酸化蛋白质组数据的生物信息学分析等，并与国外同类工作进行比较，发现优势和不足，总结研究的经验和教训，同时对未来本领域的研究方向和内容进行展望，提出新的科学与技术问题，以推动我国在翻译后修饰生物信息学领域的快速发展。其中根据质谱鉴定获得的质荷比信息谱分析获得修饰肽段的质谱数据处理研究不在本文讨论的范围之内。

1 翻译后修饰位点预测的计算方法学研究

蛋白质序列上的翻译后修饰残基位点能否预测，是本领域早期研究中面临的最大挑战。若修饰底物和位点不能准确的预测，则后续的计算分析也毫无意义。因此修饰位点预测是本领域最基础的核心问题。在早期研究中，对于该问题的真伪充满争议。例如，Blom 等[21~23]的系列工作表明，蛋白质磷酸化修饰在一级序列上存在特异性，不同激酶可能通过识别特定的序列模体（Sequence motif）来修饰底物，因此磷酸化位点可以预测。此外，实验证据表明大多数 SUMO 化修饰位点符合 ψ-K-X-E 模体（ψ为疏水性氨基酸，X 为任意氨基酸），因此 SUMO 化修饰具有高度的特异性，SUMO 化位点也可以准确的预测[24]。但对于棕榈酰化（Palmitoylation）来说，由于没有发现任何有明显特征的序列模体，主流学者一致认为棕榈酰化位点的预测将非常困难。Zhou 等[25]于 2006 年假设棕榈酰化修饰可能识别多个不同的序列模体，因此先将已知的棕榈酰化按照序列相似性分成 3 个组，再利用聚类和打分策略（Clustering and Scoring Strategy）设计了第一个棕榈酰化位点预测工具 CSS-Palm，具有较高的准确性，从而证明棕榈酰化位点具有可预测性。近年来，国际与国内学者不断地设计有效的计算方法，针对大多数常见修饰已构建了相应的计算工具。因此，翻译后修饰位点能否预测这个问题，在当前已基本不再有争议。修饰位点预测的计算方法可分为三大类，包括机器学习（Machine learning）类算法、位置特异性打分矩阵（Position-specific scoring matrix, PSSM）类算法和基于修饰肽段相似性类算法。下面简要介绍 3 类算法的原理与发展历程。

1.1 机器学习类算法

1999 年，丹麦生物序列分析中心（Center for Biological Sequence Analysis, CBS）的 Nikolaj Blom 等人使用当时已知的、实验证实的 210 个酪氨酸（Tyrosine）、584 个丝氨酸（Serine）和 108 个苏氨酸（Threonine）磷酸化位点作为训练集，利用神经网络（Neural network）算法首次实现了非特异性的（Non-specific）蛋白质磷酸化位点的预测[23]。2004 年，韩国学者 Jong Hun Kim 等[26]首次利用支持向量机（Support vector machine, SVM）算法设计了激酶特异性磷酸化位点预测方法。这一算法也被中国学者应用到了蛋白质甲基化（Methylation）、棕榈酰化、SUMO 化、乙酰化、泛素化和磷酸化等位点预测的工作中。2005 年，Li 等[27]利用 *k*-近邻（*k*-Nearest Neighbor, k-NN）算法设计了激酶特异性（Kinase-specific）磷酸化位点预测方法。2014 年，Wu 等[28]基于 *k*-近邻算法设计了磷酸酯酶识别位点预测方法。此外，2007 年 Tang 等[29]利用遗传算法整合神经网络（Genetic algorithm integrated neural network, GANN）设计了非特异性磷酸化位点预测方法。同年，李亦学、蔡煜东两个研究组合作，利用

SVMs 和偏最小二乘法（Partial least squares, PLS）分别构建了 N-糖基化（N-glycosylation）位点预测方法，并获得较高的准确性[30]。

在早期机器学习类算法的应用当中，一般将计算模型考虑成“黑盒子”，不关心具体的序列或氨基酸残基特征（Feature），也不做特征的提取和选择。2007 年，Liu 等[31]首次引入特征提取和选择的概念，使用有序前向选择（Sequential forward selection，SFS）算法提取有效的氨基酸性质作为训练特征，利用 SVMs 训练模型并设计了 SUMO 化位点预测方法。该研究组还使用最大关联最小冗余（Maximum relevance minimum redundancy, mRMR）[32]、渐进特征选择（Incremental feature selection, IFS）[32]和特征前向选择（Feature forward selection, FFS）[33]等算法进行特征的选择。该研究组还最早使用随机森林（Random Forest, RF）算法预测蛋白质的酶切位点（Cleavage site）[34]。蛋白质磷酸化过程受到多种因素影响，磷酸化位点周围的氨基酸序列并不能完全决定磷酸化是否发生。在氨基酸序列基础上，Li 等[35]将蛋白-蛋白相互作用、亚细胞定位、蛋白质结构域等特征引入到激酶识别位点的预测中，使预测效果有了很大提升。此外，Chen 等[36]于 2008 年提出 *k*-空位氨基酸对组成（Composition of *k*-spaced amino acid pairs, CKSAAP）算法用于特征的提取和选择，结合 SVMs 设计了 O-糖基化（O-glycosylation）位点预测方法。2013 年，Zou 等[37]设计了单体谱组成（Composition of monomer spectrum, CMS）算法，应用于氨基酸的编码和特征的提取。2014 年，Hou 等[38]还使用逻辑回归分类器（Logistic regression classifiers）算法设计了乙酰化位点预测方法。整体来说，在采用机器学习算法预测蛋白质翻译后修饰时，并不深入研究序列或者结构特征的具体使用和权重，具体的参数和模型往往难以对应到生物学意义上来。而机器学习算法的深度优化，往往容易带来过训练的问题，需要特别引起注意。

1.2 位置特异性打分矩阵类算法

该类算法是翻译后修饰早期计算研究的方法之一，其基本假设为修饰位点两侧短肽段的特定位置上存在氨基酸残基的偏好，因此对于阳性和阴性数据可分别构建氨基酸出现频率的矩阵，对于给定序列可分别计算在阳性和阴性数据中出现的几率值，两者相除即得分值。2001 年，Yaffe 等[39]引入位置特异性打分矩阵算法，并首次实现激酶特异性磷酸化位点预测。该类算法存在的主要问题有：（1）基本假设有理论缺陷。应用该类算法的基本假设是蛋白质序列上各个氨基酸位点具有独立性，之间没有相互作用或影响，否则概率不能连乘。但事实上邻近的氨基酸相互之间存在影响，例如修饰位点两侧的氨基酸决定修饰的特异性。因此这个基本假设从理论上来说并不够合理；（2）计算模型过于简单，只考虑氨基酸出现的频率，忽略了氨基酸其他的生化性质等。因此，该类算法在我国应用不多，或一般不独立使用。2010 年，蔡煜东研究组将位置特异性打分矩阵算法得出的分值作为可选择的特征之一，并先后设计了羟基脯氨酸（Hydroxyproline）[40]、羟基赖氨酸（Hydroxylysine）[40]、泛素化和酰胺化（Amidation）[41]等位点预测方法。

1.3 基于修饰肽段相似性类算法

在翻译后修饰位点预测方法没有推广之前，实验学家一般将潜在的修饰肽段（由修饰位点及两侧短肽段组成的序列片段）与已知的、实验证实的修饰位点相比较，根据相似肽段具有相似功能的隐含假设，凭经验人为判断给定的位点是否发生修饰，并指导其进一步实验。由此衍生出来的基于修饰肽段相似性类算法的计算模型简单，但具有明显的生物学意义。另一方面，调控底物的上游酶也可能发生修饰，从而改变其构象并导致底物识别特异性的变化，因此一种修饰的底物位点可能分别符合多个模体。基于这两个假设，Zhou 等[42]于 2004 年提出基于分组的磷酸化位点预测和打分（Group-based phosphorylation site predicting and scoring）算法 GPS 1.0，应用于激酶特异性磷酸化位点的预测。该工作定义了磷酸化位点肽段（Phosphorylation site peptide, PSP）的概念[42]，将磷酸化位点与左侧和右侧 3 个氨基酸组成的 PSP(3,3)考虑成一个整体，与一条已知的磷酸化位点肽段相比较，利用氨基酸替代矩阵 BLOSUM62 计算每一对氨基酸对的相似性分值，加和即得两条肽段的分值。为了消除显著不相似肽段造成的干扰，两条肽段加和分值小于零的置为零。给定的磷酸化位点肽段将与训练集中每一个已知的磷酸化位点肽段相比较并计算分值，平均值即为最终的打分分值[42]。2008 年，Xue 等[43]大幅度优化算法，设计 GPS 2.0 算法，在保留已有打分策略不变的基础上，引入矩阵突变（Matrix mutation, MaM）的概念，通过随机改变初始氨基酸替代矩阵中的分值来提高预测性能，结果表明在训练时间足够长的情况下，不同初始矩阵经过优化后的结果收敛成相似或相同的矩阵。由于 GPS 1.0 考虑 PSP(3,3)[42]，而 GPS 2.0 考虑 PSP(7,7)[43]，究竟磷酸化位点肽段选择多长才是最优组合这一问题没有解决。因此，Xue 等[44]设计了 GPS 2.1 算法，提出了模体长度选择（Motif length selection, MLS）方法，能够自动搜索性能最优的肽段组合。此外，还使用了 *k*-均值聚类（*k*-means clustering）方法对已知修饰位点进行分组，设计权重训练（Weight training, WT）来优化不同位置氨基酸偏好的重要性，结合已有的矩阵突变和模体长度选择方法，设计了 GPS 3.0 算法[45]。近期，针对 SUMO 化位点预测，Zhao 等[46]在 GPS 3.0 算法的基础上还引入粒子群优化（Particle swarm optimization, PSO）算法，用来缩短权重训练和矩阵突变的训练时间。

在接受相似肽段具有潜在相似功能的假设条件下，计算模型也可以通过机器学习类算法进行训练。例如，2005 年，Li 等[27]即在该假设的基础上，利用 *k*-近邻算法训练模型。此外，贝叶斯类算法如贝叶斯决策理论（Bayesian decision theory, BDT）[47]、贝叶斯判别方法（Bayesian Discriminant Method, BDM）[48]和朴素贝叶斯算法（Naive Bayes algorithm, NBA）[49]也分别被用于激酶特异性磷酸化位点、赖氨酸乙酰化位点和棕榈酰化位点的预测。2008 年，Li 等[50]提取了不同激酶底物的序列特征，采用打分矩阵的方法预测激酶特异性磷酸化位点。2010 年，Li 等[51]首次将之前用于 DNA 微阵列（DNA microarray）

数据分析、发现功能关联基因的基因集富集分析（Gene Set Enrichment Analysis, GSEA）方法，经过改进后设计了用于乙酰转移酶特异性位点预测的基于乙酰化集富集（Acetylation Set Enrichment-Based, ASEB）算法。计算结果预测 MBD1、MTA1、DNA polymerase β 和 DDB1 可能被乙酰转移酶家族 CBP/p300 或 GCN5/PCAF 修饰，后续实验证实 MBD1 和 MTA1 更倾向于被 CBP/p300 调控，而 DNA polymerase β 和 DDB1 则更倾向于被 GCN5/PCAF 调控[51]。2013 年，他们将 ASEB 方法应用在去乙酰化酶 Class I HDAC 上，也取得了较高的准确性[52]。2014 年，邱建丁研究组在 ASEB 算法的基础上，设计了预测激酶特异性磷酸化位点的方法磷酸化集富集分析（Phosphorylation Set Enrichment Analysis, PSEA），并获得较高的准确性[53]。同年，该研究组将修饰肽段之间的相似性以及氨基酸对组成作为有效特征，引入离散小波变换（Discrete wavelet transform, DWT）来提取和优化特征，设计了细胞亚定位特异性磷酸化位点方法[54]。

2 翻译后修饰相关计算工具的设计

我们收集了本领域 233 个相关的计算预测与分析工具，其链接与简介见列表（http://www.biocuckoo.org/link.php），其中 160 个是修饰底物及位点的预测工具，其他则为质谱数据处理、组学数据分析和文献挖掘等工具。本综述重点讨论翻译后修饰底物及位点预测的计算工具开发情况。虽然本领域的计算工具可以为生物信息学者提供有用的技术平台，但最主要的使用者应当是没有算法和编程基础、对数学和统计学了解不多的相关实验学家。因此，计算工具应当力求输入和输出简单明了，不需要繁琐操作，以及本地化软件能够有图形化界面或实现在线访问的网站等。计算工具的开发和推广，能够有效地为相关实验研究提供重要的参考信息。

在过去的 10 年中，中国学者共开发构建了 40 多个翻译后修饰相关的计算工具，约占本领域总数的 27%（表 2）。其中，磷酸化底物及位点预测工具 9 个，包括 GPS[42~44]、PPSP[47]、PhoScan[50]、iGPS[55]、PKIS[37]、phos_pred[56]、PSEA[53]、GPS-POLO[57]和 SubPhos[54]。对于激酶特异性磷酸化位点预测工具 GPS 系列[42~44]，修饰肽段相似性是其算法的打分策略。Zhou 等[58]于 2005 年发布第一个在线版本，能够预测 71 个激酶组共 216 个激酶的特异性修饰位点，并在此基础上提出了设计预测翻译后修饰位点预测工具的规范。2008 年，Xue 等[43]发布 2.0 版本，能够分层次地预测 408 个人类激酶特异性的位点，并将在线工具实现为本地化的软件。2012 年，Song 等[55]在 GPS 算法的基础上，进一步考虑激酶与底物之间的相互作用并以之作为限制条件，设计了 iGPS（*In vivo* GPS）算法及相应的本地化软件，能够明显降低预测结果的假阳性，进一步提高预测的准确性。2014 年，Fan 等[56]在基于修饰肽段相似性类算法的基础上，进一步考虑了基因本体论（Gene Ontology）注释和蛋白质相互作用信息，设计了激酶

特异性磷酸化位点预测工具 phos_pred。此外，GPS-POLO 基于 GPS 算法，能够预测激酶 POLO-like kinase（Plk）家族的磷酸化位点和磷酸化结合位点[57]。而邱建丁研究组通过结合离散小波变换算法和支持向量机的方法设计的 SubPhos 是第一个预测细胞亚定位特异性磷酸化位点的在线工具[54]。最近，Huang 等[59]系统研究了不同的翻译后修饰之间的相互影响（PTM crosstalk），并开发了相应的预测工具 PTM-X。

在翻译后修饰底物和位点预测工具的开发方面，中国学者的许多工作在领域内有一定的领先优势。例如，SSP 为领域内第一个 SUMO 化底物及位点预测软件[24]，CSS-Palm 为第一个棕榈酰化位点预测工具[25]，GPS-PUP 为第一个原核类泛素化（Pupylation）位点预测工具[60]，PAIL 为第一个乙酰化位点预测工具[48]，GPS-SNO 为第一个亚硝基化位点预测工具[45]，GPS-YNO2 为第一个硝基化位点预测工具[61]。此外，Jiang 等[62]设计了第一个丙酮酰化位点预测工具，而 Sun 等人[63]设计的 SGDB 也是第一个谷胱甘肽化（S-glutathionylation）位点预测工具。李婷婷研究组开发的 PTM-X 则是第一个预测修饰之间相互影响的工具[59]。因此，无论从数量还是质量来讲，中国学者在翻译后修饰底物及位点预测的计算工具开发方面，并不落后于国外学者。开发的这些工具都提供了简便易用的在线预测网站或者可本地运行的软件。在线预测的网页都以 FASTA 格式的蛋白质序列以及预测相关参数为输入，预测结果列表为输出，均符合先前提出的规范[58]。可本地运行的软件则差异性较大，比如 CKSAAP_OGlySite 提供 Perl 源代码[36]，phos_pred 提供 Matlab 代码[56]，WAP-Palm 提供 Windows 下命令行方式调用的可执行软件[64]，GPS 系列软件则基于 Java 语言提供了 Windows、Linux、Mac OS3 种操作系统下可下载安装的有图形用户界面的软件[43~46,60,61,65]。随着蛋白质组学技术的发展，翻译后修饰相关的数据正在急速增长。对于一个预测工具而言，除了在线预测网站以外，可本地运行的软件版本将越来越重要。

除了上述提到的位点预测工具之外，中国学者也开发了一些其他工具。例如，Wang 等[66]利用自然语言处理技术（Natural language processing, NLP），设计了 PPTM 在线工具，能够从文献中挖掘磷酸化信息，包括底物、激酶及磷酸化位点数据等。此外，Suo 等[67]开发了第一个预测氨基酸变异影响赖氨酸乙酰化修饰的在线工具 AcetylAAVs。

表 2　国内学者构建的 43 个翻译后修饰相关计算工具

工具名称	预测内容	网站链接
GPS[42~44,68]	激酶特异性磷酸化位点	http://gps.biocuckoo.org
PPSP[47]	激酶特异性磷酸化位点	http://ppsp.biocuckoo.org
PhoScan[50]	激酶特异性磷酸化位点	http://bioinfo.au.tsinghua.edu.cn/phoscan
iGPS[55]	激酶特异性磷酸化位点	http://igps.biocuckoo.org
PKIS[37]	激酶特异性磷酸化位点	http://bioinformatics.ustc.edu.cn/pkis
phos_pred[56]	激酶特异性磷酸化位点	http://bioinformatics.ustc.edu.cn/phos_pred
PSEA[53]	激酶特异性磷酸化位点	http://bioinfo.ncu.edu.cn/PKPred_Home.aspx
GPS-POLO[57]	Plk 磷酸化位点和结合位点	http://polo.biocuckoo.org

SubPhos[54]	细胞亚定位特异性磷酸化位点	http://bioinfo.ncu.edu.cn/SubPhos.aspx
SSP[24]	SUMO 化底物及位点	http://ssp.biocuckoo.org
SUMOsp[69]	SUMO 化位点	http://sumosp.biocuckoo.org
GPS-SUMO[46]	SUMO 化位点和结合位点	http://sumosp.biocuckoo.org
CSS-Palm[25,70]	棕榈酰化位点	http://csspalm.biocuckoo.org
NBA-Palm[49]	棕榈酰化位点	http://nbapalm.biocuckoo.org
WAP-Palm[64]	棕榈酰化位点	http://bioinfo.ncu.edu.cn/WAP-Palm.aspx
CKSAAP-Palm[71]	棕榈酰化位点	http://doc.aporc.org/wiki/CKSAAP-Palm
PPWMs[72]	棕榈酰化位点	http://math.cau.edu.cn/PPWMs.html
CKSAAP_UbSite[73]	泛素化位点	http://protein.cau.edu.cn/cksaap_ubsite
UbiProber[74]	泛素化位点	http://bioinfo.ncu.edu.cn/UbiProber.aspx
hCKSAAP_UbSite[75]	人类泛素化位点	http://protein.cau.edu.cn/cksaap_ubsite
GPS-PUP[60]	原核类泛素化位点	http://pup.biocuckoo.org
PupPred[76]	原核类泛素化位点	http://bioinfo.ncu.edu.cn/PupPred.aspx
PAIL[48]	乙酰化位点	http://pail.biocuckoo.org
LysAcet[77]	乙酰化位点	http://www.biosino.org/LysAcet
LAceP[38]	乙酰化位点	http://www.scbit.org/iPTM
SSPKA[78]	物种特异性乙酰化位点	http://www.structbioinfor.org/Lab/SSPKA
ASEB[51,79]	酶特异性乙酰化位点	http://bioinfo.bjmu.edu.cn/huac/
PLMLA[80]	甲基化和乙酰化位点	http://bioinfo.ncu.edu.cn/inquiries_PLMLA.aspx
MeMo[81]	甲基化位点	http://www.bioinfo.tsinghua.edu.cn/~tigerchen/memo.html
PMeS[82]	甲基化位点	http://bioinfo.ncu.edu.cn/inquiries_PMeS.aspx
Oglyc[83]	糖基化位点	http://www.biosino.org/Oglyc
CKSAAP_OGlySite[36]	糖基化位点	http://bioinformatics.cau.edu.cn/zzd_lab/CKSAAP_OGlySite
GPS-CCD[65]	Calpain 酶切位点	http://ccd.biocuckoo.org
LabCaS[84]	Calpain 酶切位点	http://www.csbio.sjtu.edu.cn/bioinf/LabCaS
Cascleave 2.0[85]	Caspase 酶切位点	http://www.structbioinfor.org/cascleave2
pyrupred[62]	丙酮酰化位点	http://www.zhni.net/pyrupred/index.html
SGDB[63]	谷胱甘肽化位点	http://csb.shu.edu.cn/SGDB
PredSulSite[86]	硫化位点	http://bioinfo.ncu.edu.cn/inquiries_PredSulSite.aspx
GPS-TSP[87]	硫化位点	http://tsp.biocuckoo.org
GPS-SNO[45]	亚硝基化位点	http://sno.biocuckoo.org
GPS-YNO2[61]	硝基化位点	http://yno2.biocuckoo.org
pptm[66]	磷酸化位点文本挖掘	http://bioinformatics.ustc.edu.cn/pptm
AcetylAAVs[67]	影响乙酰化的序列变异	http://bioinfo.ncu.edu.cn/AcetylAAVs_Home.aspx
PTM-X[59]	修饰之间的相互影响	http://bioinfo.bjmu.edu.cn/ptm-x/

3 翻译后修饰相关数据库的构建

翻译后修饰调控的分子机制非常复杂。以磷酸化为例，催化磷酸化反应的激酶被称为“录入器”（Writer），执行去磷酸化的磷酸酶被称为“擦除器”（Eraser），而识别特定磷酸化位点并与之结合的磷酸化结合功能域（Phospho-binding domain, PBD）则被称为“读取器”。类似录入器-擦除器-读取器的机

制至少在泛素化和乙酰化中也存在，例如 E1-E2-E3 级联调控泛素化修饰，DUB 执行去泛素化修饰，而泛素结合域/泛素相互作用模体（Ubiquitin-binding domain/ubiquitin-interacting motif）识别特定泛素化位点；HAT 催化乙酰化反应，HDAC 执行去乙酰化，而包含 Bromodomain 的蛋白质则可以识别并结合特定乙酰化位点。因此，修饰酶和去修饰酶共同发挥作用，动态调控底物的修饰水平，并通过包含修饰结合功能域的蛋白质识别修饰位点，向通路的下游传递信号。因此，修饰酶、去修饰酶和包含修饰结合功能域的蛋白质的系统发现与分类，是进一步深入研究翻译后修饰调控机制的基础。此外，近年来随着高通量质谱学等技术的迅猛发展，成千上万的修饰位点被实验鉴定。收集、整理、整合以及深入注释这些数据，能够为进一步的计算分析和实验研究提供重要的数据资源。

鉴于蛋白质翻译后修饰的复杂性和相关数据的迅速积累，构建翻译后修饰相关数据库的意义重大。但相对于修饰位点预测方法设计与工具开发来说，已建立并良好维护的相关数据库并不多，目前世界上总共仅有 53 个可访问、可检索、可使用的公共数据库。我国在数据库开发方面尤为薄弱，仅建立 8 个数据库（表 3），仅占总数的 15%。2009 年，李亦学、谢鹭等研究组构建了中国第一个翻译后修饰数据库 SysPTM[88]，通过整合国外其他数据库的信息以及文献检索的方式，共收集了约 50 种修饰的 33 421 个底物及 117 349 个位点。该数据库在发表时是当时数据量最大的修饰数据库之一。2014 年，SysPTM 发布 2.0 版本[89]，包括 2031 个物种中 50 多种修饰的 53 235 个底物和 471 109 个位点，而国外数据量最大的同类数据库 PhosphoSitePlus 仅包含 19 939 个修饰底物和 349 603 个位点[90]。2010 年，Ren 等[91]首次定义了磷酸化相关的单核苷酸多态性（Phosphorylation-related SNP, PhosSNP），计算系统鉴定了 64 035 个潜在影响磷酸化修饰的 SNP，按照影响方式分为 5 类，并构建了相应的数据库 PhosSNP。2011 年，Du 等[92]首次构建人类泛素化数据库 hUbiquitome，通过文献检索收集了 1 个 E1 酶、12 个 E2 酶、138 个 E3 酶、279 个泛素化底物和 17 个 DUB。在此基础上，Gao 等[8]通过文献检索进一步收集了 26 个已知的 E1、105 个 E2、1,003 个 E3 和 148 个 DUB，按序列相似性和功能分成 30 个家族后，分别构建隐马尔科夫模型（Hidden Markov model, HMM）谱，系统鉴定了 70 个真核生物的 738 个 E1、2937 个 E2、46 631 个 E3 和 6647 个 DUB，并构建了相应的数据库 UUCD，是目前最全面的泛素化相关酶的数据库。利用类似的策略，Wang 等[7]还系统鉴定了 84 个真核生物的 50 433 个激酶和 11 296 个磷酸酶，并构建相应的数据库 EKPD，是目前最全面的激酶与磷酸酶数据库。2011 年，Liu 等[9]通过文献检索的方式收集了已知的 3311 个乙酰化底物的 7151 个位点，构建了第一个乙酰化数据库 CPLA。2014 年，该数据库经历了较大幅度的完善，更名为 CPLM[93]，成为第一个专注于赖氨酸修饰的数据库，包括 122 个物种中 12 种赖氨酸修饰如乙酰化、泛素化和 SUMO 化等信息，共有 45 748 个底物的 189 919 个位点。Chen 等[94]构建的同类数据库 mUbiSiDa 包括了 5 个物种中 35 494 个实验证实的泛素化底物和

110 976 个位点。这些数据库从不同的角度为翻译后修饰相关研究提供了重要的数据，且提供了简便易用的查询和下载方式。由于数据的收集和整理非常耗时费力，因此目前没有一个数据库能包含所有的翻译后修饰数据。但是，通过学者之间的交流和合作，将来翻译后修饰相关数据库可以考虑进行交叉和整合，互相之间提供数据接口，方便进行统一的查询和使用。

表 3 国内学者构建的 8 个翻译后修饰相关数据库

数据库名称	网址	修饰类型	数据内容
SysPTM[88,89]	http://lifecenter.sgst.cn/SysPTM/	多种	多种修饰的底物位点
PhosSNP[91]	http://phossnp.biocuckoo.org	磷酸化	影响磷酸化的 SNP
hUbiquitome[92]	http://bioinfo.bjmu.edu.cn/hubi/	泛素化	人类中泛素化相关酶和底物位点
CPLA[9]	http://cpla.biocuckoo.org	乙酰化	赖氨酸乙酰化底物位点
UUCD[8]	http://uucd.biocuckoo.org	泛素化	70 种真核生物中泛素与类泛素化相关酶
CPLM[93]	http://cplm.biocuckoo.org	赖氨酸修饰	12 种赖氨酸修饰的底物位点
EKPD[7]	http://ekpd.biocuckoo.org	磷酸化	84 种真核生物中蛋白激酶和磷酸酶
mUbiSiDa[94]	http://reprod.njmu.edu.cn/mUbiSiDa	泛素化	哺乳动物泛素化底物位点

4 基于修饰蛋白质组数据的生物信息学分析

近年来，随着蛋白质组学技术的发展，基于蛋白质组尤其是修饰蛋白质组学数据的生物信息学研究可以对翻译后修饰提供整体和更深层次的认识。在这方面中国学者目前开展的研究有限，主要涉及 4 个方面，包括系统分析修饰的序列特征与功能调控的偏好、磷酸化网络模拟与分析以及翻译后修饰的进化等。

2005 年，周丰丰与薛宇合作对人和小鼠中保守的潜在 SUMO 化蛋白质进行了系统的计算分析，结果表明 SUMO 化在转录调节、发育、信号转导、磷酸化、细胞生长与细胞周期等生命活动中起着重要的作用[24]。2013 年，Chen 等[76]对原核类泛素化的序列、结构和进化等方面的特征进行了系统的分析，并将原核类泛素化位点和真核生物中的泛素化位点进行了细致的比较。2014 年，该研究组对人类磷酸化组的亚细胞特异性进行了细致的分析[54]，提出磷酸化位点在不同的亚细胞区域有着不同的序列模体。2013 年，Zhou 等[95]系统地研究了人类泛素化修饰位点的结构特征，对溶剂可及性、残基网络中心性以及局部结构微环境进行了细致的分析，指出泛素化位点的局部结构包含了序列层次未能体现的许多信息。2014 年，该研究组综述了泛素化位点预测的进展[96]，详细分析了不同预测工具的算法和特征，并对酵母（*Saccharomyces cerevisiae*）、人（*Homo sapiens*）、小鼠（*Mus musculus*）和拟南芥（*Arabidopsis thaliana*）中的泛素化位点进行了分析，讨论了开展物种特异性泛素化位点预测的必要性。

修饰蛋白质组数据除了可以提供对翻译后修饰底物位点整体的了解以外，还使系统分析修饰的调控、模拟磷酸化网络成为了可能。2013 年，Liu 等[57]通过构建计算预测，系统地分析了酵母、线虫（*Caenorhabditis elegans*）、果蝇（*Drosophila melanogaster*）、爪蟾（*Xenopus laevis*）、小鼠和人中 PLK 激酶的磷酸化底物位点和磷酸化结合位点。统计分析表明，相较于磷酸化底物而言，PLK 激酶的磷酸化结合蛋白质更倾向与有丝分裂相关，而 PLK 激酶的磷酸化调控偏好于分布式模型。对中体、中心体和动点上的蛋白质的大规模分析表明这些亚细胞位置显著富集 PLK 的底物。进一步的体外和体内实验证明动点蛋白质 Mis18B 确实是人源 PLK1 的磷酸化结合蛋白质，而结合位点与计算预测吻合。除了对单个激酶的磷酸化调控进行系统分析以外，基于组学数据还可以构建和模拟磷酸化网络。2012 年，Song 等[55]开发并利用 iGPS 软件系统构建并分析了酵母、线虫、果蝇、小鼠和人的磷酸化网络。在此基础上，利用人类肝脏磷酸化组数据构建了人类肝脏磷酸化网络，并系统分析了该网络中的主要调控激酶。2014 年，Qi 等[97]系统地研究了小鼠睾丸中的磷酸化组和激酶调控网络。通过分析表明，不同富集方法得到的磷酸化组之间虽然重叠有限，但是均能反映样本的生物学状态。对基于不同方法得到的磷酸化组数据进行的激酶活性分析（Kinase activity analysis, KAA）计算结果非常一致地表明 PLK 家族激酶在小鼠睾丸组织中具有高活性。进一步的实验验证了该推测，并验证了 PLK 激酶的活性在小鼠睾丸中的重要性。

此外，磷酸化作为生命活动的重要调控机制，研究其在进化过程中的特征具有重要的意义。国际上在翻译后修饰尤其是磷酸化的进化问题上开展了一系列的研究工作，而国内这方面的研究工作较少。2011 年，Wang 等[98]系统研究了脊椎动物不同功能模块中磷酸化的进化，发现磷酸化位点在脊椎动物特异性功能模块（Vertebrate-specific functional module, VFM）如细胞信号处理、对刺激的响应等比基本功能模块（Basic functional module, BFM）如代谢和遗传过程等更为保守。而且脊椎动物特异性功能模块中的磷酸化位点更多。这些结果表明在脊椎动物进化过程中磷酸化在脊椎动物特异性功能模块更为重要。由于磷酸化是一个较为快速的生化反应，脊椎动物特异性功能模块可能依赖于这种快速的反应来对外界和内部信号做出反应。2014 年，Pan 等[87]系统地分析了酪氨酸上磷酸化、硫化和硝基化三种修饰之间的相互影响（Crosstalk），而长程进化分析表明多修饰酪氨酸与单修饰酪氨酸的保守性无显著差异。

5 结语与展望

2014 年 5 月，本文的大多数作者齐聚武汉华中科技大学，共同组织、召开了一个小型的、非正式的学术研讨会，重点讨论如何在蛋白质组尤其是修饰蛋白质数据处理后进行生物信息学分析和建立分子

系统生物学模型等问题。由于在过去的 10 年里，中国从事翻译后修饰生物信息学研究的学者多半各自为战，相互之间缺少沟通和交流，因此与会学者一致认为在今后的工作和研究中，有必要加强国内同行之间的交流与合作，共同推动中国翻译后修饰生物信息学研究的发展。

现阶段我国在翻译后修饰的生物信息学研究方面有一定特色，主要体现在翻译后修饰底物和位点预测计算方法学的多元化与众多计算工具的设计与维护。在计算方法学的设计和完善方面，我国学者在 3 类主流算法包括机器学习类算法、置特异性打分矩阵类算法和基于修饰肽段相似性类算法等方面都有深入的研究，并不断尝试新的研究策略和方法。由于目前尚没有普适的、能够精确预测翻译后修饰位点的算法，因此计算方法学的多元化是进一步研究的基础。而计算工具的设计与长期维护，能够为实验学家提供有用的技术平台，相应的预测结果能够为进一步实验研究提供有价值的参考信息。

我国在本领域的研究中也存在明显不足，例如：（1）翻译后修饰相关的知识型数据库较少，而数据的通量和质量决定后续生物信息学预测和分析的可靠性。因此翻译后修饰数据的收集、整理、整合与注释，是今后我国学者需要重视的问题。（2）我国学者利用修饰组学数据开展深度分析、发现修饰调控规律的工作不多。近年来的研究趋势表明，单一层面的组学数据分析并不能完全系统地生物体当前的状态及其对外界刺激的反映，而不同组学的数据之间的关联度有限，跨层次、多组学数据的整合与比较分析逐渐受到青睐。例如，近期国外报道的一项针对结肠和直肠癌的工作中，作者将蛋白质组数据与基因组的数据结合起来进行分析[99]，发现存在体细胞突变的异构体（Somatic variants），蛋白质表达量显著低；发现转录后的 mRNA 表达水平与蛋白质表达水平关联不高；发现 DNA 拷贝数异常（Copy number alterations, CAN）主要影响 mRNA 表达，但与蛋白质表达关联不大。而国外另一项比较小鼠正常皮肤与皮肤癌组织的工作中[100]，结合定量蛋白质组与定量磷酸化组数据，通过比较发现 47.3%的蛋白质在蛋白和磷酸化水平都受到调控，而超过一半的蛋白质仅在磷酸化水平受到调控。显然，这些发现都需要多层次、多组学数据的整合和比较。因此，如何整合更多的、其他层次的数据，结合修饰组学数据做出新的、重要的生物学发现，是未来我国本领域学者所面临的重大挑战。（3）有价值的生物信息学预测，必然是生物学功能的预测。在所有翻译后修饰底物和位点预测的工作背后，都存在一个隐含的假设，即蛋白质的修饰必然有生物学功能。因此实验学家与生物信息学家事实上共同关心的是两个问题：第一，是否发生修饰；第二，该修饰有没有功能。从理论上来说这个假设没有问题，但现实中实验学家更关心是在细胞信号通路中发挥重要功能，并且其状态的改变能够显著改变可观测的细胞或生物体表型的修饰。近期的研究表明，这个假设并不正确，因为有相当比例的修饰位点仅对蛋白质结构的维持有一定作用，不参与重要的信号通路，其修饰的缺失或敲除并不会影响细胞表型，因此被学者称为无功能修饰（Non-functional PTM）。有研究估计大约 65%的磷酸化位点可能为无功能的磷

酸化位点[101]。如何从修饰组学数据中正确预测出功能修饰（Functional PTM），也是未来本领域需要解决的重大问题。4）单纯的计算预测因为总会存在假阳性结果，生物信息学的研究范式也与主流生物学研究大不相同，并且复杂的数学和计算模型也令人望而生畏，因此预测结果很难使主流实验学家信服。计算与实验合作，通过实验学的方法验证或至少部分验证预测结果，已成为领域的研究趋势。例如上述研究小鼠皮肤癌发病机制的研究中[100]，作者比较了非恶性刺瘤（Papilloma）和恶性鳞细胞癌（Squamous cell carcinoma）的磷酸化组数据，利用类似于我们提出的激酶活性分析方法预测了皮肤癌中潜在高活性的激酶，并用实验学的方法予以证实。与实验学家加强合作，验证计算结果的准确性，并结合实验做出重要的生物学发现，是国内本领域亟需解决的问题。

综上所述，在过去的 10 年里，中国在翻译后修饰生物信息学研究方面，从零星的、断续的探索开始，逐渐形成有体系、有特色并有希望持续发展壮大的研究方向。虽然在某些方面能够与国外同类工作保持同步或互有领先，但我们更应当学习国外学者们研究的长处，不断提高研究水平。我们相信，中国翻译后修饰生物信息的研究能够有更好的发展，其研究成果一方面可为主流生物信息学领域提供重要的方法学参考，另一方面也可为蛋白质组学领域提供有力的分析技术。

参考文献

[1]. Xue Y, Liu Z, Cao J, Ren J. Computational prediction of post-translational modification sites in proteins. *Systems and Computational Biology - Molecular and Cellular Experimental Systems*, InTech, 2011, 5772(6): 18559.

[2]. Ubersax JA, Ferrell JE, Jr. Mechanisms of specificity in protein phosphorylation. *Nat Rev Mol Cell Biol*, 2007, 8(7):530-541.

[3]. Hershko A. The ubiquitin system for protein degradation and some of its roles in the control of the cell-division cycle (Nobel lecture). *Angew Chem Int Ed Engl*, 2005, 44(37):5932-5943.

[4]. Smith KT, Workman JL. Introducing the acetylome. *Nat Biotechnol*, 2009, 27(10):917-919.

[5]. Riedl SJ, Shi Y. Molecular mechanisms of caspase regulation during apoptosis. *Nat Rev Mol Cell Biol*, 2004, 5(11):897-907.

[6]. Croall DE, Ersfeld K. The calpains: modular designs and functional diversity. *Genome Biol*, 2007, 8(6):218.

[7]. Wang Y, Liu Z, Cheng H, Gao T, Pan Z, Yang Q, Guo A, Xue Y. EKPD: a hierarchical database of eukaryotic protein kinases and protein phosphatases. *Nucleic Acids Res*, 2014, 42(Database issue):D496-502.

[8]. Gao T, Liu Z, Wang Y, Cheng H, Yang Q, Guo A, Ren J, Xue Y. UUCD: a family-based database of ubiquitin and ubiquitin-like conjugation. *Nucleic Acids Res*, 2013, 41(Database issue):D445-451.

[9]. Liu Z, Cao J, Gao X, Zhou Y, Wen L, Yang X, Yao X, Ren J, Xue Y. CPLA 1.0: an integrated database of protein lysine acetylation. *Nucleic Acids Res*, 2011, 39(Database issue):D1029-1034.

[10]. Walsh CT, Garneau-Tsodikova S, Gatto GJ, Jr. Protein posttranslational modifications: the chemistry of proteome diversifications. *Angew Chem Int Ed Engl*, 2005, 44(45):7342-7372.

[11]. Spiro RG. Protein glycosylation: nature, distribution, enzymatic formation, and disease implications of glycopeptide bonds. *Glycobiology*, 2002, 12(4):43R-56R.

[12]. Schopfer FJ, Baker PR, Freeman BA. NO-dependent protein nitration: a cell signaling event or an oxidative inflammatory response? *Trends Biochem Sci*, 2003, 28(12):646-654.

[13]. Geiss-Friedlander R, Melchior F. Concepts in sumoylation: a decade on. *Nat Rev Mol Cell Biol*, 2007, 8(12):947-956.

[14]. el-Husseini Ael D, Bredt DS. Protein palmitoylation: a regulator of neuronal development and function. *Nat Rev Neurosci*, 2002, 3(10):791-802.

[15]. Paik WK, Paik DC, Kim S. Historical review: the field of protein methylation. *Trends Biochem Sci*, 2007, 32(3):146-152.

[16]. Kehoe JW, Bertozzi CR. Tyrosine sulfation: a modulator of extracellular protein-protein interactions. *Chem Biol*, 2000, 7(3):R57-61.

[17]. Hess DT, Matsumoto A, Kim SO, Marshall HE, Stamler JS. Protein S-nitrosylation: purview and parameters. *Nat Rev Mol Cell Biol*, 2005, 6(2):150-166.

[18]. Ren J, Gao X, Liu Z, Cao J, Ma Q, Xue Y. Computational analysis of phosphoproteomics: progresses and perspectives. *Curr Protein Pept Sci*, 2011, 12(7):591-601.
[19]. Xue Y, Gao X, Cao J, Liu Z, Jin C, Wen L, Yao X, Ren J. A summary of computational resources for protein phosphorylation. *Curr Protein Pept Sci*, 2010, 11(6):485-496.
[20]. Liu Z, Wang Y, Xue Y. Phosphoproteomics-based network medicine. *FEBS J*, 2013, 280(22):5696-5704.
[21]. Blom N, Kreegipuu A, Brunak S. PhosphoBase: a database of phosphorylation sites. *Nucleic Acids Res*, 1998, 26(1):382-386.
[22]. Kreegipuu A, Blom N, Brunak S, Jarv J. Statistical analysis of protein kinase specificity determinants. *FEBS Lett*, 1998, 430(1-2):45-50.
[23]. Blom N, Gammeltoft S, Brunak S. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J Mol Biol*, 1999, 294(5):1351-1362.
[24]. Zhou F, Xue Y, Lu H, Chen G, Yao X. A genome-wide analysis of sumoylation-related biological processes and functions in human nucleus. *FEBS Lett*, 2005, 579(16):3369-3375.
[25]. Zhou F, Xue Y, Yao X, Xu Y. CSS-Palm: palmitoylation site prediction with a clustering and scoring strategy (CSS). *Bioinformatics*, 2006, 22(7):894-896.
[26]. Kim JH, Lee J, Oh B, Kimm K, Koh I. Prediction of phosphorylation sites using SVMs. *Bioinformatics*, 2004, 20(17):3179-3184.
[27]. Li A, Wang L, Shi Y, Wang M, Jiang Z, Feng H. Phosphorylation site prediction with a modified k-nearest neighbor algorithm and BLOSUM62 matrix. *Conf Proc IEEE Eng Med Biol Soc*, 2005, 6:6075-6078.
[28]. Wu Z, Lu M, Li T. Prediction of substrate sites for protein phosphatases 1B, SHP-1, and SHP-2 based on sequence features. *Amino Acids*, 2014, 46(8):1919-1928.
[29]. Tang YR, Chen YZ, Canchaya CA, Zhang Z. GANNPhos: a new phosphorylation site predictor based on a genetic algorithm integrated neural network. *Protein Eng Des Sel*, 2007, 20(8):405-412.
[30]. Li S, Liu B, Cai Y, Li Y. Predicting protein N-glycosylation by combining functional domain and secretion information. *J Biomol Struct Dyn*, 2007, 25(1):49-54.
[31]. Liu B, Li S, Wang Y, Lu L, Li Y, Cai Y. Predicting the protein SUMO modification sites based on Properties Sequential Forward Selection (PSFS). *Biochem Biophys Res Commun*, 2007, 358(1):136-139.
[32]. Niu S, Huang T, Feng K, Cai Y, Li Y. Prediction of tyrosine sulfation with mRMR feature selection and analysis. *J Proteome Res*, 2010, 9(12):6490-6497.
[33]. Cai YD, Lu L. Predicting N-terminal acetylation based on feature selection method. *Biochem Biophys Res Commun*, 2008, 372(4):862-865.
[34]. Li BQ, Cai YD, Feng KY, Zhao GJ. Prediction of protein cleavage site with feature selection by random forest. *PLoS One*, 2012, 7(9):e45854.
[35]. Li T, Du P, Xu N. Identifying human kinase-specific protein phosphorylation sites by integrating heterogeneous information from various sources. *PLoS One*, 2010, 5(11):e15411.
[36]. Chen YZ, Tang YR, Sheng ZY, Zhang Z. Prediction of mucin-type O-glycosylation sites in mammalian proteins using the composition of k-spaced amino acid pairs. *BMC Bioinformatics*, 2008, 9:101.
[37]. Zou L, Wang M, Shen Y, Liao J, Li A, Wang M. PKIS: computational identification of protein kinases for experimentally discovered protein phosphorylation sites. *BMC Bioinformatics*, 2013, 14:247.
[38]. Hou T, Zheng G, Zhang P, Jia J, Li J, Xie L, Wei C, Li Y. LAceP: lysine acetylation site prediction using logistic regression classifiers. *PLoS One*, 2014, 9(2):e89575.
[39]. Yaffe MB, Leparc GG, Lai J, Obata T, Volinia S, Cantley LC. A motif-based profile scanning approach for genome-wide prediction of signaling pathways. *Nat Biotechnol*, 2001, 19(4):348-353.
[40]. Hu LL, Niu S, Huang T, Wang K, Shi XH, Cai YD. Prediction and analysis of protein hydroxyproline and hydroxylysine. *PLoS One*, 2010, 5(12):e15917.
[41]. Cui W, Niu S, Zheng L, Hu L, Huang T, Gu L, Feng K, Zhang N, Cai Y, Li Y. Prediction of protein amidation sites by feature selection and analysis. *Mol Genet Genomics*, 2013, 288(9):391-400.
[42]. Zhou FF, Xue Y, Chen GL, Yao X. GPS: a novel group-based phosphorylation predicting and scoring method. *Biochem Biophys Res Commun*, 2004, 325(4):1443-1448.
[43]. Xue Y, Ren J, Gao X, Jin C, Wen L, Yao X. GPS 2.0, a tool to predict kinase-specific phosphorylation sites in hierarchy. *Mol Cell Proteomics*, 2008, 7(9):1598-1608.
[44]. Xue Y, Liu Z, Cao J, Ma Q, Gao X, Wang Q, Jin C, Zhou Y, Wen L, Ren J. GPS 2.1: enhanced prediction of kinase-specific phosphorylation sites with an algorithm of motif length selection. *Protein Eng Des Sel*, 2011, 24(3):255-260.
[45]. Xue Y, Liu Z, Gao X, Jin C, Wen L, Yao X, Ren J. GPS-SNO: computational prediction of protein S-nitrosylation sites with a modified GPS algorithm. *PLoS One*, 2010, 5(6):e11290.

[46]. Zhao Q, Xie Y, Zheng Y, Jiang S, Liu W, Mu W, Liu Z, Zhao Y, Xue Y, Ren J. GPS-SUMO: a tool for the prediction of sumoylation sites and SUMO-interaction motifs. *Nucleic Acids Res*, 2014, 42(Web Server issue):W325-330.

[47]. Xue Y, Li A, Wang L, Feng H, Yao X. PPSP: prediction of PK-specific phosphorylation site with Bayesian decision theory. *BMC Bioinformatics*, 2006, 7:163.

[48]. Li A, Xue Y, Jin C, Wang M, Yao X. Prediction of Nepsilon-acetylation on internal lysines implemented in Bayesian Discriminant Method. *Biochem Biophys Res Commun*, 2006, 350(4):818-824.

[49]. Xue Y, Chen H, Jin C, Sun Z, Yao X. NBA-Palm: prediction of palmitoylation site implemented in Naive Bayes algorithm. *BMC Bioinformatics*, 2006, 7:458.

[50]. Li T, Li F, Zhang X. Prediction of kinase-specific phosphorylation sites with sequence features by a log-odds ratio approach. *Proteins*, 2008, 70(2):404-414.

[51]. Li T, Du Y, Wang L, Huang L, Li W, Lu M, Zhang X, Zhu WG. Characterization and prediction of lysine (K)-acetyl-transferase specific acetylation sites. *Mol Cell Proteomics*, 2012, 11(1):M111 011080.

[52]. Li T, Song B, Wu Z, Lu M, Zhu WG. Systematic identification of Class I HDAC substrates. *Brief Bioinform*, 2014, 15(6):963-972.

[53]. Suo SB, Qiu JD, Shi SP, Chen X, Liang RP. PSEA: Kinase-specific prediction and analysis of human phosphorylation substrates. *Sci Rep*, 2014, 4:4524.

[54]. Chen X, Shi SP, Suo SB, Xu HD, Qiu JD. Proteomic analysis and prediction of human phosphorylation sites in subcellular level reveal subcellular specificity. *Bioinformatics*, 2014.

[55]. Song C, Ye M, Liu Z, Cheng H, Jiang X, Han G, Songyang Z, Tan Y, Wang H, Ren J, Xue Y, Zou H. Systematic analysis of protein phosphorylation networks from phosphoproteomic data. *Mol Cell Proteomics*, 2012, 11(10):1070-1083.

[56]. Fan W, Xu X, Shen Y, Feng H, Li A, Wang M. Prediction of protein kinase-specific phosphorylation sites in hierarchical structure using functional information and random forest. *Amino Acids*, 2014, 46(4):1069-1078.

[57]. Liu Z, Ren J, Cao J, He J, Yao X, Jin C, Xue Y. Systematic analysis of the Plk-mediated phosphoregulation in eukaryotes. *Brief Bioinform*, 2013, 14(3):344-360.

[58]. Zhou F, Xue Y, Yao X, Xu Y. A general user interface for prediction servers of proteins' post-translational modification sites. *Nat Protoc*, 2006, 1(3):1318-1321.

[59]. Huang Y, Xu B, Zhou X, Li Y, Lu M, Jiang R, Li T. Systematic Characterization and Prediction of Post-Translational Modification Cross-talk. *Mol Cell Proteomics*, 2015.

[60]. Liu Z, Ma Q, Cao J, Gao X, Ren J, Xue Y. GPS-PUP: computational prediction of pupylation sites in prokaryotic proteins. *Mol Biosyst*, 2011, 7(10):2737-2740.

[61]. Liu Z, Cao J, Ma Q, Gao X, Ren J, Xue Y. GPS-YNO2: computational prediction of tyrosine nitration sites in proteins. *Mol Biosyst*, 2011, 7(4):1197-1204.

[62]. Jiang Y, Li BQ, Zhang Y, Feng YM, Gao YF, Zhang N, Cai YD. Prediction and Analysis of Post-Translational Pyruvoyl Residue Modification Sites from Internal Serines in Proteins. *PLoS One*, 2013, 8(6):e66678.

[63]. Sun C, Shi ZZ, Zhou X, Chen L, Zhao XM. Prediction of S-glutathionylation sites based on protein sequences. *PLoS One*, 2013, 8(2):e55512.

[64]. Shi SP, Sun XY, Qiu JD, Suo SB, Chen X, Huang SY, Liang RP. The prediction of palmitoylation site locations using a multiple feature extraction method. *J Mol Graph Model*, 2013, 40:125-130.

[65]. Liu Z, Cao J, Gao X, Ma Q, Ren J, Xue Y. GPS-CCD: a novel computational program for the prediction of calpain cleavage sites. *PLoS One*, 2011, 6(4):e19001.

[66]. Wang M, Xia H, Sun D, Chen Z, Wang M, Li A. Literature mining of protein phosphorylation using dependency parse trees. *Methods*, 2014, 67(3):386-393.

[67]. Suo SB, Qiu JD, Shi SP, Chen X, Huang SY, Liang RP. Proteome-wide analysis of amino acid variations that influence protein lysine acetylation. *J Proteome Res*, 2013, 12(2):949-958.

[68]. Xue Y, Zhou F, Zhu M, Ahmed K, Chen G, Yao X. GPS: a comprehensive www server for phosphorylation sites prediction. *Nucleic Acids Res*, 2005, 33(Web Server issue):W184-187.

[69]. Xue Y, Zhou F, Fu C, Xu Y, Yao X. SUMOsp: a web server for sumoylation site prediction. *Nucleic Acids Res*, 2006, 34(Web Server issue):W254-257.

[70]. Ren J, Wen L, Gao X, Jin C, Xue Y, Yao X. CSS-Palm 2.0: an updated software for palmitoylation sites prediction. *Protein Eng Des Sel*, 2008, 21(11):639-644.

[71]. Wang XB, Wu LY, Wang YC, Deng NY. Prediction of palmitoylation sites using the composition of k-spaced amino acid pairs. *Protein Eng Des Sel*, 2009, 22(11):707-712.

[72]. Li YX, Shao YH, Deng NY. Improved prediction of palmitoylation sites using PWMs and SVM. *Protein Pept Lett*, 2011, 18(2):186-193.

[73]. Chen Z, Chen YZ, Wang XF, Wang C, Yan RX, Zhang Z. Prediction of ubiquitination sites by using the composition of k-spaced amino acid pairs. *PLoS One*, 2011, 6(7):e22930.

[74]. Chen X, Qiu JD, Shi SP, Suo SB, Huang SY, Liang RP. Incorporating key position and amino acid residue features to identify general and species-specific Ubiquitin conjugation sites. *Bioinformatics*, 2013, 29(13):1614-1622.

[75]. Chen Z, Zhou Y, Song J, Zhang Z. hCKSAAP_UbSite: improved prediction of human ubiquitination sites by exploiting amino acid pattern and properties. *Biochim Biophys Acta*, 2013, 1834(8):1461-1467.

[76]. Chen X, Qiu JD, Shi SP, Suo SB, Liang RP. Systematic analysis and prediction of pupylation sites in prokaryotic proteins. *PLoS One*, 2013, 8(9):e74002.

[77]. Li S, Li H, Li M, Shyr Y, Xie L, Li Y. Improved prediction of lysine acetylation by support vector machines. *Protein Pept Lett*, 2009, 16(8):977-983.

[78]. Li Y, Wang M, Wang H, Tan H, Zhang Z, Webb GI, Song J. Accurate in silico identification of species-specific acetylation sites by integrating protein sequence-derived and functional features. *Sci Rep*, 2014, 4:5765.

[79]. Wang L, Du Y, Lu M, Li T. ASEB: a web server for KAT-specific acetylation site prediction. *Nucleic Acids Res*, 2012, 40(Web Server issue):W376-379.

[80]. Shi SP, Qiu JD, Sun XY, Suo SB, Huang SY, Liang RP. PLMLA: prediction of lysine methylation and lysine acetylation by combining multiple features. *Mol Biosyst*, 2012, 8(5):1520-1527.

[81]. Chen H, Xue Y, Huang N, Yao X, Sun Z. MeMo: a web tool for prediction of protein methylation modifications. *Nucleic Acids Res*, 2006, 34(Web Server issue):W249-253.

[82]. Shi SP, Qiu JD, Sun XY, Suo SB, Huang SY, Liang RP. PMeS: prediction of methylation sites based on enhanced feature encoding scheme. *PLoS One*, 2012, 7(6):e38772.

[83]. Li S, Liu B, Zeng R, Cai Y, Li Y. Predicting O-glycosylation sites in mammalian proteins by using SVMs. *Comput Biol Chem*, 2006, 30(3):203-208.

[84]. Fan YX, Zhang Y, Shen HB. LabCaS: labeling calpain substrate cleavage sites from amino acid sequence using conditional random fields. *Proteins*, 2013, 81(4):622-634.

[85]. Wang M, Zhao XM, Tan H, Akutsu T, Whisstock JC, Song J. Cascleave 2.0, a new approach for predicting caspase and granzyme cleavage targets. *Bioinformatics*, 2014, 30(1):71-80.

[86]. Huang SY, Shi SP, Qiu JD, Sun XY, Suo SB, Liang RP. PredSulSite: prediction of protein tyrosine sulfation sites with multiple features and analysis. *Anal Biochem*, 2012, 428(1):16-23.

[87]. Pan Z, Liu Z, Cheng H, Wang Y, Gao T, Ullah S, Ren J, Xue Y. Systematic analysis of the in situ crosstalk of tyrosine modifications reveals no additional natural selection on multiply modified residues. *Sci Rep*, 2014, 4:7331.

[88]. Li H, Xing X, Ding G, Li Q, Wang C, Xie L, Zeng R, Li Y. SysPTM: a systematic resource for proteomic research on post-translational modifications. *Mol Cell Proteomics*, 2009, 8(8):1839-1849.

[89]. Li J, Jia J, Li H, Yu J, Sun H, He Y, Lv D, Yang X, Glocker MO, Ma L, Yang J, Li L, Li W, Zhang G, Liu Q, Li Y, Xie L. SysPTM 2.0: an updated systematic resource for post-translational modification. *Database (Oxford)*, 2014, 2014:bau025.

[90]. Hornbeck PV, Zhang B, Murray B, Kornhauser JM, Latham V, Skrzypek E. PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res*, 2014.

[91]. Ren J, Jiang C, Gao X, Liu Z, Yuan Z, Jin C, Wen L, Zhang Z, Xue Y, Yao X. PhosSNP for systematic analysis of genetic polymorphisms that influence protein phosphorylation. *Mol Cell Proteomics*, 2010, 9(4):623-634.

[92]. Du Y, Xu N, Lu M, Li T. hUbiquitome: a database of experimentally verified ubiquitination cascades in humans. *Database (Oxford)*, 2011, 2011:bar055.

[93]. Liu Z, Wang Y, Gao T, Pan Z, Cheng H, Yang Q, Cheng Z, Guo A, Ren J, Xue Y. CPLM: a database of protein lysine modifications. *Nucleic Acids Res*, 2014, 42(Database issue):D531-536.

[94]. Chen T, Zhou T, He B, Yu H, Guo X, Song X, Sha J. mUbiSiDa: a comprehensive database for protein ubiquitination sites in mammals. *PLoS One*, 2014, 9(1):e85744.

[95]. Zhou Y, Liu S, Song J, Zhang Z. Structural propensities of human ubiquitination sites: accessibility, centrality and local conformation. *PLoS One*, 2013, 8(12):e83167.

[96]. Chen Z, Zhou Y, Zhang Z, Song J. Towards more accurate prediction of ubiquitination sites: a comprehensive review of current methods, tools and features. *Brief Bioinform*, 2014.

[97]. Qi L, Liu Z, Wang J, Cui Y, Guo Y, Zhou T, Zhou Z, Guo X, Xue Y, Sha J. Systematic Analysis of the Phosphoproteome and Kinase-substrate Networks in the Mouse Testis. *Mol Cell Proteomics*, 2014, 13(12):3626-3638.

[98]. Wang Z, Ding G, Geistlinger L, Li H, Liu L, Zeng R, Tateno Y, Li Y. Evolution of protein phosphorylation for distinct functional modules in vertebrate genomes. *Mol Biol Evol*, 2011, 28(3):1131-1140.

[99]. Zhang B, Wang J, Wang X, Zhu J, Liu Q, Shi Z, Chambers MC, Zimmerman LJ, Shaddox KF, Kim S, Davies SR, Wang S, Wang P, Kinsinger CR, Rivers RC, Rodriguez H, Townsend RR, Ellis MJ, Carr SA, Tabb DL, Coffey RJ, Slebos RJ, Liebler DC. Proteogenomic characterization of human colon and rectal cancer. *Nature*, 2014, 513(7518):382-387.

[100]. Zanivan S, Meves A, Behrendt K, Schoof EM, Neilson LJ, Cox J, Tang HR, Kalna G, van Ree JH, van Deursen JM, Trempus CS, Machesky LM, Linding R, Wickstrom SA, Fassler R, Mann M. In Vivo SILAC-Based Proteomics Reveals Phosphoproteome Changes during Mouse Skin Carcinogenesis. *Cell Rep*, 2013, 3(2):552-566.

[101]. Landry CR, Levy ED, Michnick SW. Weak functional constraints on phosphoproteomes. *Trends Genet*, 2009, 25(5):193-197.

（责任编委：赵方庆）