·专论与综述·

基于蛋白质序列的泛素化位点预测研究进展*

何 冰 宋晓峰△

(南京航空航天大学生物医学工程系 江苏 南京 210016)

摘要:泛素化是目前广受关注的一种翻译后修饰过程,对蛋白质降解、DNA 修复等多种细胞过程都具有重要的调控作用。本文根据国内外蛋白质泛素化位点预测的研究,分析了预测泛素化位点的特征属性,总结了对这些特征进行优化的特征选择方法,并对预测过程中所使用的各种机器学习分类器进行了概述。

关键词:泛素化位点;特征选择;机器学习分类器

中图分类号:Q71,TP181 **文献标识码**:A **文章编号**:1673-6273(2012)18-3573-04

Progress in Prediction of Protein Ubiquitination Sites Based on the Protein Sequence*

HE Bing, SONG Xiao-feng△

(Department of Biomedical Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, 210016, China)

ABSTRACT: Ubiquitination, a popular process of post-translational modification, plays an important regulatory role in numerous cellular process such as protein degradation, DNA repair and so on. The essay analyzed the features influencing the protein ubiquitination sites prediction and feature selection approaches that can extract the informative features improving the performance of classifier from the total feature set. Meanwhile, several machine learning classifiers of predicting the protein ubiquitination sites were introduced based on the various relevant studies of home and abroad.

Key words: Ubiquitination sites; Feature selection; Machine learning classifiers

Chinese Library Classification (CLC): Q71, TP181 Document code: A

Article ID: 1673-6273(2012)18-3573-04

前言

自从 2004 年 Aaron Ciechanover 等发现了泛素调节蛋白质降解过程之后,泛素化底物蛋白及位点预测已成为生命科学的研究热点。泛素是一种存在于大多数真核细胞中的高度保守的蛋白质,由 76 个氨基酸残基组成,主要参与短寿命蛋白质的快速降解。泛素化对许多细胞过程都有重要的调控作用,人类许多疾病的形成也与泛素化通路中的某些缺陷有关。目前由实验测得的酵母泛素化位点较完整,可以直接在数据库中获得,所以大多数泛素化位点的预测都是针对酵母展开。本文就基于酵母蛋白序列的泛素化位点预测的研究做一综述,并对人类泛素化位点的预测提出展望。

1 概述

泛素化是一种重要的翻译后修饰过程,即泛素分子共价的结合于靶蛋白的特定赖氨酸残基上,形成肽键从而促进蛋白质降解。该过程也需要泛素激活酶 E1、泛素结合酶 E2、泛素连接酶 E3 的共同参与,具有高度的保守性和可逆性。泛素化在真核细胞中作为一种调控机制,调控着多种细胞过程如蛋白质降解、信号传导、DNA 修复、细胞分裂 / 分化和胞内定位等[1]。

传统的方法主要是在实验基础上从蛋白质序列中鉴定泛素化位点。由于经泛素化修饰的蛋白质大多是稳定性较差的短寿命蛋白,所以通过实验来鉴定泛素化位点需要较高的成本、消耗时间较长,鉴定出的泛素化位点也有限。因此,通过生物信息学的方法来研究更多未知靶蛋白的泛素化位点是明智和可行的。目前国内外很多学者在泛素化位点预测方面均取得了一定的成果,这里我们只列举其中的部分成果,如表 1 所示。

2008 年 Chun-Wei Tung 和 Shinn-Ying H 从 UbiProt 数据库中提取了包含 157 个泛素化位点和 3676 个非泛素化位点的 105 个蛋白质,建立一个准确率达到 72.9 %的预测器。之后他们又提出了一种有效的理化特性挖掘算法(IPMA),进行特征选择后的准确率提高到 84.44 %[2],并在此基础上建立了用于泛素化位点预测的 UbiPred 网站。2010 年 Radivojac P 等从三种途径(一个文献搜索、两项基于蛋白酶体研究和自己实验测得的泛素化位点)中整合了 272 个非冗余的泛素化位点,之后从 124 个线粒体基质蛋白中提取 4651 个非泛素化片段,开发了泛素化预测器 UbPred,预测准确率达到 77.2 %[3]。Yudong Cai 等也在 2010 年从 SysPTM[4]中提取出 273 个蛋白质序列,其中

* 基金项目:国家自然科学基金资助项目(61171191);江苏省自然科学基金资助项目(BK2010500)

作者简介:何冰(1986-),女,硕士研究生,研究方向:生物信息学。E-mail:hebing052@163.com

△通讯作者:宋晓峰,Email: xfsong@nuaa.edu.cn

(收稿日期:2011-10-19 接受日期:2011-11-15)

包含 378 个泛素化位点和 1359 个非泛素化位点，对预测方法进行适当改进后，预测结果的马氏相关系数（MCC）是 0.142，明显高于现存的泛素化位点预测器 UbiPred(0.135)和 UbPred (0.117)[5]。Tzong-Yi Lee 等在 2011 年从 UbiProt 和 UniProtKB/Swiss-Prot 中提取出来自 350 个蛋白质中的 442 个泛素化位点和 16 084 个非泛素化位点，建立的预测器的敏感性、特异性和准确率分别为 65.5 %、74.8 %和 74.5 %[6]。

表 1 几种泛素化位点预测研究成果

Table 1 Studies about prediction of protein ubiquitination sites

Author	Ubiquiti-nation sites	Non-ubiqu-itination sites	Proteins	Accuracy rate	MCC	Classifiers	Feature selection method
Chun-Wei Tung et al	157	3676	105	72.9 % (84.4 %)	0.135	SVM	IPMA
Radivojac P et al	272	4651	124	77.2 %	0.117	Random forest	T-test
Yudong Cai et al	378	1359	273	NA	0.142	Nearest Neighbor Algorithm	mRMR
Tzong-Yi Lee et al	442	16084	350	74.5 %	NA	Radial Basis Function Networks	F value

注：NA 表示在相应文献中没有提到。

Note: NA means no relevant data mentioned in the reference.

从上述研究中我们可以看出，由于采用的特征属性不同、建立预测模型的方法各异，预测的准确率就略有不同，但总体而言预测能力处于不断提高的趋势。同时随着实验的不断进行，越来越多的泛素化位点被鉴定，这对我们进行泛素化位点预测提供了有力的数据支持，也更有利于进行泛素化位点预测的研究。

2 蛋白质泛素化位点预测的相关特征分析

有效的特征属性和合适的分类器对建立一个高性能预测器的重要作用是不言而喻的，在过去的研究中已经提出了大量基于序列的区别蛋白质或氨基酸功能的特征。表 1 中提到的多位研究人员所采用的特征各不相同，但主要是从蛋白质序列中提取四种有用的特征进行评估，即氨基酸组成、进化信息、理化特性、结构特性。

2.1 蛋白质的氨基酸组成

氨基酸是构成蛋白质的基本单位，氨基酸通过肽键共价连接形成聚合物，进一步形成蛋白质。研究蛋白质包含的基本信息，最简单最直观的方法是计算各氨基酸的组成，即 20 种氨基酸在一条蛋白质中的出现频率。如今氨基酸组成对蛋白质的影响已经成功应用在预测蛋白质的各种位点及相互作用[7]等研究中。

由泛素化的定义可知，泛素分子共价的结合于靶蛋白的特定赖氨酸残基（K）上，因此研究泛素化位点 K 周围的各氨基酸组成可能有助于泛素化位点预测。Tzong-Yi Lee 等在预测泛素化位点时，将单氨基酸组成和二联氨基酸组成作为建立分类器模型的主要特征[6]。Radivojac P 等在建立泛素化预测器 UbPred 时，同样也考虑了以赖氨酸 K 为中心的对称窗口长度内的氨基酸组成[3]。其他许多研究人员在预测泛素化位点时也都考虑了氨基酸组成这一基本特征，结果显示单独使用氨基酸组成并不能准确预测泛素化位点，因此还需要考虑其他特征对泛素化位点预测的影响。

2.2 蛋白质的进化信息

蛋白质氨基酸序列的进化保守性通常暗示着重要的生理生化功能，如果一个蛋白质特殊位点的氨基酸保守，则该位点可能位于这个蛋白质的重要功能区域内[8]。通常使用蛋白质序列的位置特异性得分矩阵（PSSM）研究氨基酸的进化信息，PSSM 能测量给定窗口长度的残基保守性。蛋白质的 PSSM 矩阵可以从 PSI-BLAST 网站中测得[9]，该矩阵使用 20 维向量代表 20 种不同氨基酸发生的频率，行数由给定蛋白质序列的残基数决定。

许多研究人员在研究蛋白质相关生物功能时发现，与单独使用氨基酸组成的预测相比，使用 PSSM 能显著的提高预测蛋白质结构特性的准确性[10]。这主要是由于 PSSM 源自多重序列比对的相关位置的频率向量，这种编码方式使用了序列的同源信息，所以可以提高预测准确率。Tzong-Yi Lee[6]和 Radivojac P [3]等人在泛素化位点预测研究中，都通过 PSI-BLAST 中非冗余 GenBank 数据库得到 PSSM 矩阵，进而分析蛋白质片段的进化信息。

尽管 PSSM 被广泛作为研究蛋白质进化信息的重要特征，但是目前大多采用原始 PSSM 矩阵，即对任意一个长度为 的蛋白质片段，都有一个 的矩阵。这就产生了一个高维的特征空间，对之后预测模型的参数优化、建模及预测过程都造成了不便。考虑到上述问题，Rafał Adamczak 提出将特征提取和特征选择综合应用于 PSSM 矩阵，对 PSSM 矩阵降维处理，实现对特征空间降维的同时提高预测的准确性[11]。

2.3 蛋白质的理化特性

蛋白质的理化特性对于研究蛋白质的结构和功能具有重要作用，研究人员通过大量的实验和理论研究已确定了一些氨基酸的理化特性。每一种氨基酸的理化特性都能用 20 个数值组成的数据集表示，这 20 个数值称为氨基酸指数。AAIndex 是 Shuichi Kawashima 等开发的一种包含众多氨基酸理化特性指数的数据库，最新的版本 AAIndex 9.1 中包含 544 个理化特性，去除部分指数不完全的特征，最后可得到 531 个理化特性[12]。在对蛋白质序列的生物信息研究中广泛使用 AAIndex 进行理化特性的分析，如蛋白质 SUMO 修饰位点的预测[13]，蛋白质亚细胞定位预测[14]等。

在蛋白质泛素化位点预测的研究中，可能有些理化特性与预测不相关，甚至影响预测的准确性，研究人员就选取了不同的理化特性进行预测。Chun-Wei Tung 等先对 531 个理化特性进行综合分析，之后又根据每个理化特性对预测性能的影响程度对其排序，最后提取出预测准确率最高时的 31 个理化特性[2]。Radivojac P 等则主要选取了静电荷、总电荷、芳香族含量、疏水性和序列复杂度这 5 种理化特性进行预测[3]。

William R. Atchley 等基于因子分析法的思想，使用多元统计分析用 5 种属性模式替代这 531 个理化特性，分别是极性、二级结构、分子体积、密码子多样性和静电荷，同时还计算了每种模式的氨基酸因子得分[15]。这种方法在极大程度上缩减了理化特性的特征空间，更有利于解决蛋白质序列矩阵问题和不同的生物问题。Tao Huang 等在预测 SNPs 时就使用了这 5 种氨基酸因子研究其理化特性[8]。Yudong Cai 在预测泛素化位点时也采用了这种方法编码赖氨酸片段的每个氨基酸，结果显示旁侧序列中氨基酸的理化特性对泛素化过程有重要作用[5]。

2.4 蛋白质的结构特征

参与翻译后修饰的氨基酸侧链优先接近蛋白质表面[16]，因此研究泛素化位点片段表面区域的溶剂可及性[17]就成为泛素化位点预测的一个特征。Tzong-Yi Lee 等使用 RVP- Net[18]计算了蛋白质中每个氨基酸的溶剂可及性比例值（ASA），结果显示大部分泛素化和非泛素化赖氨酸处于可及表面区域，但泛素化位点片段的溶剂可及性均值远远高于非泛素化位点片段[6]。

为了更好的预测泛素化位点，有些研究人员也考虑了蛋白质的二级结构特征。André Cat- ic 等已经发现泛素化位点片段更倾向于无规则卷曲[19]。Radivojac P 等使用 BLAST 搜索蛋白质结构信息[3]，Tzong-Yi Lee 等则使用 PSIPRED[20]计算蛋白质序列中赖氨酸片段的二级结构，结果都反映泛素化连接酶易出现在无规则卷曲或是螺旋结构区域，非泛素化位点没有明显的二级结构倾向性[6]。

无序蛋白是一种在最原始状态下没有形成稳定三维结构的蛋白质，研究显示与原核基因相比，真核基因中 30 %的蛋白质都处于无序状态，蛋白质中无序区的出现与基因调控和信号传导等功能有密切的关系。Yvonne JK Edwards 等研究发现无序蛋白比有序蛋白更能预测泛素化位点[21]。Yudong Cai 等通过 VSL2 计算出蛋白质序列中赖氨酸片段的每个氨基酸的无序得分，也显示了无序蛋白与泛素化有密切的关系[5]。

除了上述特征，一些基于序列的蛋白质属性预测器如磷酸化[22]、高温度因子[23]、柔性[24]等预测器用以辅助泛素化位点预测。

3 蛋白质泛素化位点预测的特征选择

特征选择能使分类器的误差最小化，提高分类器的性能，产生高效率低成本的分类器。蛋白质泛素化位点预测所构造的特征空间往往是高维的，这种高维的数据直接在预测器中处理，预测时间冗长且有些特征并不能为泛素化位点预测提供依据，甚至会干扰泛素化位点的预测。尽管诸如 SVM 和神经网络等分类器能通过改变参数调整冗余特征的预测性能，去除一些无效的特征从而极大增强预测性能[25]。

F 值是一种常用的能有效区分两个样本集的测量方法，在蛋白质预测中通常被用来评价特征属性[26]。具体公式[27]如下：

$$F(xi)=\left|\frac{\mu_i^+-\mu_i^-}{\sigma_i^++\sigma_i^-}\right| \quad (1)$$

$x_i, i=1,2,...,N$

μ_i^+、μ_i^- 分别表示每个特征的正样本和负样本的平均值；σ_i^+、σ_i^- 分别表示每个特征的正样本和负样本的标准方差。F 值越高则说明该特征区分正样本和负样本的能力越强。

T 检验也是一种常用于分类器中的特征选择方法，主要应用统计学的相关原理。最后得到的 p 值是将观测结果认为具有总体代表性的犯错概率，它是结果可信程度的一个递减指标。与 F 值相反，p 值越小则说明该特征的统计显著性越强，对预测性能的提高越有利。

最大冗余最小相关算法 [28](mRMR) 是一种基于交互信息（MI）的特征选择方法，基本思想是根据目标变量的相关性和特征间的冗余性对特征进行排序，若一个特征与目标变量具有最大相关且在特征间有最小冗余，则该特征的区分能力最强。

前向选择序列属性的特征选择方法（PSFS）也已经使用在预测 SUMO 修饰位点[13]等研究中。该方法首先建立一个空特征集，当一个特征能提高预测的准确率时，将该特征添加到特征集中。对所有的特征进行判断，最后建立起一个选择的特征集。

除了上述特征选择之外，Chun-Wei Tung 等通过改进双目标基因算法（GA）[29]提出了一种有效地理化特性挖掘算法（IP-MA），这种算法的优势在于用最少的特征获得最大的预测准确率[2]。Yudong Cai 等在应用 mRMR 特征选择的同时，又用增量特征选择进一步确定最优的特征[5]。

4 蛋白质泛素化位点预测的机器学习分类器

在蛋白质泛素化位点预测中可使用多种机器学习分类器，如最近邻算法、贝叶斯法、支持向量机（SVM）、随机森林、邻近归并法等。

SVM 被广泛应用在数据分类、生物科学等领域中，它的基本思想是将数据映射到一个高维的特征空间中，寻找一个具有最大间隔的分割超平面使得学习器得到全局最优化[30]。Chih-Jen Lin 等开发了一个易于操作、快速有效的 SVM 软件包 LIBSVM，可供用户解决各种分类问题[31]。在预测蛋白质泛素化位点时，Chun-Wei Tung 等分别用最近邻算法、贝叶斯法和 SVM 这 3 种分类器对氨基酸特性、进化信息和理化特性等特征进行评估，结果显示使用 SVM 预测这 3 种特征的准确率最好，分别是 68.0 %、66.67 %和 72.85 %[2]。

随机森林也是一种广泛应用的机器学习算法，由多个决策树构成，每个决策树都依赖独立采样的向量值，在随机森林中所有的决策树都具有相同的分布。随机森林可以处理大量的输入变量、学习快速，对于不平衡的分类数据集，随机森林还可以平衡误差，很多情况下都能够产生高准确度的分类器。Radivojac P 等在构建 UbPred 时证明，与逻辑回归、SVM 和神经网络相比随机森林有较好的性能，能获得 72.0 %的准确率[3]。

然而 Dudoit 等则认为随机森林在分类数据不平衡时分类能力较差，虽然可以通过调整分类权重参数提高预测率，但却增加了预测的复杂性[32]。Yudong Cai 等考虑到 SVM 和随机森

林算法容易引入过学习问题的缺点，采用简单且运算速度较快的邻近归并算法[5]。Tzong-Yi Lee 等则使用一种有效的径向基函数网络识别蛋白质泛素化位点[6]。

5 问题与展望

蛋白质的泛素化与蛋白质序列中赖氨酸周围氨基酸的组成、进化信息、理化特性、结构特性等特征存在着一定的关系。研究泛素化位点的关联特征，可以帮助我们有效的识别泛素化修饰的靶蛋白，对蛋白质的降解有更深入的了解。目前泛素化位点预测的研究已经取得了一定的成果，通过不断改进方法预测的准确率已经获得了逐渐提高。不足之处在于多数研究都是针对酵母展开，这些研究对人类蛋白质可能并不适合。随着实验鉴定出的人类泛素化位点不断增加，我们将更全面的了解与人类泛素化位点相关的特征因素，从而准确的预测人类蛋白质泛素化位点。

参 考 文 献 (References)

[1] Jannie M. R. Danielsen, Kathrine B. Sylvestersen, Simon Bekker-Jensen, et al. Mass spectrometric analysis of lysine ubiquitylation reveals promiscuity at site level [J]. Molecular & Cellular Proteomics, 2010, 10(11):1-12

[2] Chun-Wei Tung, Shinn-Ying Ho. Computational identification of ubiquitylation sites from protein sequences [J]. BMC Bioinformatics, 2008, 9: 310-319

[3] Radivojac P, Vacic V, Haynes C, et al. Identification, analysis, and prediction of protein ubiquitination sites [J]. Proteins, 2010,78 (2): 365-380

[4] Li H, Xing X, Ding G, et al. A systematic resource for proteomic research on post translational modifications [J]. Mol Cell Proteomics, 2009, 8(8): 1839-1849

[5] Yudong Cai, Tao Huang, Lele Hu, et al. Prediction of lysine ubiquitination with mRMR feature selection and analysis [J]. Amino Acids, 2012, 42(4):1387-1395

[6] Lee T-Y, Chen S-A, Hung H-Y, et al. Incorporating Distant Sequence Features and Radial Basis Function Net- works to Identify Ubiquitin Conjugation Sites[J]. PLoS ONE, 2011,6(3): 17331

[7] Roy S, Martinez D, Platero H, et al. Exploiting Amino Acid Composition for Predicting Protein-Protein Interactions[J]. PloS ONE, 2009,4 (11): 7813

[8] Huang T, Wang P, Ye Z-Q, et al. Prediction of Deleterious Non-Synonymous SNPs Based on Protein Interacti- on Network and Hybrid Properties[J]. PLoS ONE, 2010,5(7): 11900

[9] Altschul SF, Madden TL, Schaffer AA, et al. Gapped BLAST and PSI-BL- AST: a new gene- ration of protein database search programs [J]. Nucleic Acids, 1997, 25(17): 3389-3402

[10] Shandar Ahmad, Akinori Sarai. PSSM-based prediction of DNA binding sites in proteins[J]. BMC Bioinformatics, 2005,6(33):1-6

[11] Rafał Adamczak. Dimensionality reduction of PSSM matrix and its influence on secondary structure and relative solvent accessibility predictions[J]. World Academy of Science, Engineering and Technology, 2009, 58: 310-316

[12] Shuichi Kawashima, Piotr Pokarowski, Maria Pokarowska, et al. AAindex: amino acid index database, progress report 2008[J]. Nucleic Acids Research, 2008(36):202-205

[13] Liu B., Li S., Wang Y., et al. Predicting the protein SUMO modification sit- es based on properties sequential forward selection (PSFS)[J]. Biochemical and Biophysical Research Com- munications, 2007, 358: 136-139

[14] Sarda D., Chua G.H., Li K-B., et al. pSLIP:SVM based protein subcellular localization prediction using multiple physicochemical properties[J]. BMC Bioinformatics, 2005,6(152):1-12

[15] Atchley WR, Zhao J, Fernandes AD, et al. Solving the protein sequence metric problem [J]. Proc Natl Acad Sci USA, 2003, 102(18): 6395-6400

[16] Pang CN, Hayen A, Wilkins MR. Surface accessibility of protein posttranslational modifications [J]. Proteome Res, 2007, 6 (5): 1833-1845

[17] Hu Z, Fu Y, Halees AS, et al. SeqVISTA:a new module of integrated computational tools for studying transcriptional regulation[J]. Nucleic Acids Res, 2004, 32:235-241

[18] Ahmad S, Gromiha MM, Sarai A. RVP-net: online prediction of real valued accessible surface area of proteins from single sequences [J]. Bioinformatics, 2003, 19:1849-1851

[19] Catic A, Collins C, Church GM, et al. Preferred in vivo ubiquitination sites[J]. Bioinformatics, 2004, 20: 3302-3307

[20] McGuffin LJ, Bryson K, Jones DT. The PSIPRED protein structure prediction server[J]. Bioinformatics, 2000, 16: 404-405

[21] Yvonne JK Edwards, Anna E Lobley, Melissa M Pentony, et al. Insights into the regulation of intrinsically di- sordered proteins in the human proteome by analyzing sequence and gene expression data[J]. Genome Biology, 2009, 10:50

[22] Iakoucheva LM, Radivojac P, Brown CJ, et al. The importance of intrinsic disorder for protein phosphorylation [J]. Nucleic Acids Res, 2004,32(3):1037-1049

[23] Radivojac P, Obradovic Z, Smith DK, et al. Protein flexibility and intrinsic disorder[J]. Protein Sci, 2004, 13(1): 71-80

[24] Vihinen M, Torkkila E, Riikonen P. Accuracy of protein flexibility predictions[J]. Protein Science, 1994,13:141-149

[25] Guyon, I., Elisseeff, A. An introduction to variable and feature selection[J]. Journal of Machine Learning Rese- arch, 2003, 3: 1157-1182

[26] Gideon Dror, Rotem Sorekand, Ron Shamir. Accurate identification of alternatively spliced exons using support vector machineVol [J]. Bioinformatics, 2005, 21(7):897-901

[27] Golub T., Slomin D., Tamayo P., et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring[J]. Science, 1999, 286: 531-537

[28] Peng H, Long F, Ding C. Feature selection based on mutual information: criteria of max-dependency, max- relevance, and min-redundancy [J]. IEEE Computer Society, 2005, 27(8):1226-1238

[29] Ho SY, Chen JH, Huang MH. Inheritable genetic algorithm for biobjective 0/1 combinatorial optimization problems and its applications [J]. IEEE Trans Syst Man Cybern B Cybern, 2004, 34(1):609-620

[30] S. Li, B. Liu, R. Zeng, et al. Predicting O-glycosylation sites in mammalian proteins by using SVMs [J]. Comput. Biol. Chem, 2006(30): 203-208

[31] Chih-Chung Chang, Chih-Jen Lin. LIBSVM: a library for support vector machines [J]. ACM Transactions on Intelligent Systems and Technology, 2011, 2(27):1-27

[32] Dudoit S., Fridlyand J. Statistical Analysis of Gene Expression Microarray Data[M]. Boca Raton, FL: Chapman & Hall/CRC Press