

摘要

磷酸化是最重要的蛋白质翻译后修饰之一，蛋白质磷酸化和去磷酸化为真核细胞提供了调节机制。随着高通量鉴定磷酸化蛋白质技术的发展，尤其是质谱技术在蛋白质组学中的应用，磷酸化修饰数据不断积累，从现有数据中挖掘规律从而对未知蛋白质进行磷酸化修饰位点预测的条件日益成熟。将计算方法引入磷酸化蛋白质组学的研究中，将有利于发现新的磷酸化修饰规律并为生物学实验提供验证信息，从而推动磷酸化蛋白质组学的发展。

计算智能领域的方法可以很好地应用于位点预测问题。但对于生物信息学来说，除了给出较为准确的预测结果外，还需要给出对判断结果易于理解的解释才能够增加预测方法的可信度。规则抽取不但可以提供合理的解释来指导生物学实验，而且可以从现有数据中发现新的具有生物学意义的磷酸化修饰规律为磷酸化蛋白质的进一步研究提供有价值的参考信息。

本文深入分析了磷酸化修饰位点数据的特点，采用支持向量机分类方法试验和比较了多种特征构造提取、特征选择和分类方法的有效性；提出用 AdaBoost 方法对筛选后的氨基酸性质和邻近序列位置进行特征选择并进行分类器训练，形成了新的磷酸化位点预测算法 AproPhos，该算法在特异性高于已有预测算法（约 2 个百分点）的基础上，大大提高了预测的灵敏度（约 10 个百分点）。同时设计了一种新的基于 AdaBoost 方法的规则抽取方法，可以给出可理解的修饰位点邻近序列上氨基酸性质分布规律，并对分类结果进行解释。AproPhos 及其规则抽取算法扩展了磷酸化位点预测方法在实际中的应用范围，既可以用于提供充分信息的位点预测，又可以用来提高磷酸化蛋白质质谱鉴定效率。

最后本文提出了一种利用串联质谱同位素信息进行分子式预测的算法和系统 FFP(Fragment ion Formula Prediction)，无论从计算效率上还是预测精度上较以前的方法都有了很大的提高。使分子式预测可以广泛用于质谱的预处理和蛋白质（包括磷酸化蛋白质）的鉴定，提高鉴定效率。

关键词：磷酸化，位点预测，规则抽取，SVM，AdaBoost

Research on Protein Phosphorylation Sites Prediction and Rules Extraction

Cai Jinjin (Computer Software and Theory)

Directed by Prof. Zhao Jieyu

Protein phosphorylation is one of the most important reversible post-translational modifications (PTMs). Phosphorylation and dephosphorylation provides a regulatory mechanism in eukaryotic cells. High-throughput methods for the identification of PTMs are being developed, in particular the application of mass spectrometry to the fields of proteomics. With the recent increase in protein phosphorylated sites identified by mass spectrometry, in silico prediction of potential phosphorylation sites may facilitate the identification of phosphorylated protein. It is indeed advantageous to provide validation for biological experiments and discover new rules of phosphorylation by integrating computational approaches into phosphorylated proteins research.

Computational intelligence is a good choice for high performance phosphorylated sites prediction. Furthermore, explaining how a prediction is made is the key to its credibility, especially for applications to bioinformatics. Not only are the extracted rules reasonable interpretations that are useful to guide the biological experiments, but also are helpful to integrate computational technology for advanced deduction.

In this thesis, after comprehensive comparisons among the different features of phosphorylated sites, we select physicochemical and biological properties of amino acids around the sites through the primary structure of protein for the feature extraction. We design a new phosphorylated sites prediction method named AproPhos with AdaBoost as feature selection and classification. Different from other prediction methods with lower sensitivity, our method shows about 10% higher sensitivity as well as about 2% higher specificity. In order to provide the understandable explanation of the prediction, we design a novel approach to extract rules from AdaBoost classification. AproPhos and the rules extraction method expand the application field of the phosphorylated sites prediction. They can give the distribution formulas of amino acids properties around the sites at the same time perform the good prediction, as well as can enhance the efficiency of phosphorylated protein identification with tandem mass spectra.

In this thesis, we also develop a new method FFP (Fragment ion Formula Prediction) which can predict the best formulas of fragment ions more accurately through the minimization of the distance between theoretical and observed isotope patterns within less time. It can help to preprocess the mass spectrum data and improve the reliability of the identification of protein (including phosphorylated proteins) with tandem mass spectra.

Keywords: phosphorylation, prediction, rules extraction, SVM, AdaBoost