

摘要

蛋白质鉴定是蛋白质组学研究的基础问题之一。串联质谱技术和数据库搜索已成为自底向上蛋白质鉴定策略的常规技术手段。为了鉴定蛋白质序列，首先需要鉴定由蛋白质酶切产生的肽序列。可以说，肽鉴定模块是蛋白质鉴定软件的核心，而将肽序列与串联质谱关联起来的肽打分函数，则成为肽鉴定的关键。本文研究的重点就在于如何利用统计学习技术来设计肽打分函数以及提高肽打分函数的性能。

为了衡量各种肽谱匹配特征的重要性，找到最具代表性的肽谱匹配品质描述指标，从而基于这些特征构建肽打分函数，本文首先提出了一种基于支持向量机-逐步特征排除算法 (SVM-RFE) 的肽谱匹配特征重要性排序方法。在已有的针对肽鉴定的工作中，肽谱匹配特征主要用于肽鉴定结果的后续评价和验证，对于它们的重要性以及能否直接用于肽打分函数，并没有相关工作进行考查。本文提出的方法利用线性排序支持向量机的权重向量，对肽谱匹配特征的重要性进行排序。同时通过一次去掉一个的特征排除过程，能够观察到某个特征对线性肽打分函数的贡献程度。实验表明，对于肽打分函数来说，实验谱峰强度匹配比例 *IntenRatio*、理论碎片离子匹配比例 *IonRatio* 和匹配碎片离子的连续互补性 *CnscCmpl* 是最为重要的三个肽谱匹配特征，这三个特征足以将训练集中 99% 以上的正确肽序列排在候选肽列表的首位。

基于肽谱匹配特征重要性排序的结果，本文提出了一个利用实验谱峰强度匹配比例 *IntenRatio* 和理论碎片离子匹配比例 *IonRatio* 构造的肽打分函数 i^2Score 。在肽打分函数中尝试了上述两个特征的乘积和加和两种运算形式。通过实验比较，发现乘积形式具有更高的灵敏度和正误匹配区分度。在三个不同类型数据集上的对比实验表明， i^2Score 的肽鉴定性能要显著优于同样基于相似性度量的 SEQUEST 肽打分函数。在 1% 的假发现率条件下，在谱图水平， i^2Score 能多鉴定出 17% 到 78%，而在非冗余肽水平， i^2Score 能多鉴定出 13% 到 45%。同时，还与 Mascot 以及本文课题组先前提出的 KSDP 肽打分函数进行了比较， i^2Score 的性能也表现出一定的优势。

反相高效液相色谱和串联质谱联用是蛋白质鉴定中常用的分析技术。反相色谱用来分离肽混合样品，不同肽序列所具有的不同理化性质，会导致其在色谱柱中的保留时间不同。肽反相保留时间是可以根据其氨基酸序列进行预测的。已有

许多工作致力于预测反相色谱条件下的肽保留时间，并取得了较好的预测效果。实际保留时间和预测保留时间的差异可以用来度量肽序列的可靠性。虽然目前已存在许多种肽打分方法，但是融入保留时间信息的肽打分函数仍是不可用的。为了进一步提高上述打分函数的肽鉴定性能，本文将肽的实际保留时间和预测保留时间的差值 $rDiff$ 融入一个新的肽打分函数 $i^2rScore$ ，这个打分函数是 i^2Score 、 Δi^2 （见3.2.2小节）和 $rDiff$ 的线性组合。线性组合的权重向量通过在特定质谱数据集上训练的线性分类 SVM 模型来动态确定。实验表明，与 i^2Score 肽打分函数相比， $i^2rScore$ 能够提高 10% 以上的谱图鉴定数和非冗余肽鉴定数。

关键词： 计算蛋白质组学； 统计学习； 质谱； 肽鉴定； 打分函数

Research on Methods of Peptide Identification Based on Statistical Learning

Hai-Peng WANG (Computer Application Technology)

Supervised by Prof. Wen GAO

Protein identification is one of the fundamental issues in proteomics research. Tandem mass spectrometry and database searching have become routine techniques for bottom-up protein identification strategies. As the first step in protein identification, peptides have to be sequenced in order to identify the proteins from which they are derived. In this sense, the module of peptide identification is the core part of the protein identification software, and the peptide scoring function, which correlates peptide sequences with tandem mass spectra, is the key to peptide identification. In this thesis, the research focuses on how to design peptide scoring functions and how to improve their peptide identification performance by using statistical learning techniques.

In order to measure the relative importance of peptide-spectrum match (PSM) features, a ranking method based on the support vector machine recursive feature elimination (SVM-RFE) algorithm is proposed to find the most essential and representative features, with which we can construct novel peptide scoring functions. In currently existing work on peptide identification, the PSM features are mostly employed to evaluate and validate peptide identification results. There is no relative work focusing on examining the importance of the PSM features and their usability for peptide scoring functions. The method proposed here can rank the PSM features in terms of their relative importance which is measured by the weight vector of the linear ranking SVM. The degree to which one PSM feature contributes to the linear peptide scoring function can be observed through one-by-one feature elimination. The experiments discovered the three most important PSM features for peptide scoring functions. These features are as follows, (1) *IntenRatio* - the ratio of the total matched peak intensity to the total ion current, (2) *IonRatio* - the ratio of the number of matching fragment ions to the number of all predicted fragments, and (3) *CnscCmpl* - the measure for both the continuity and the complementarity of the matching fragment ions. These features are sufficient to rank more than 99% correct peptides at the first place in the candidate peptide lists in the training set.

According to the feature ranking results, a simple and effective peptide scoring function, $i^2\text{Score}$, is then constructed using the two simplest PSM features, *IntenRatio*

and *IonRatio*. Two types of mathematical operations, multiplication and addition, of the two features are attempted in the scoring function. Comparative experiments show that the multiplication operation is superior to the addition one with respect to the identification sensitivity and the discriminative power for discrimination between correct and random PSM matches. The experiments on three different types of spectral data sets demonstrate that the peptide scoring function, $i^2\text{Score}$, performs significantly better than the one in SEQUEST, which is also based on similarity measures. With a false discovery rate of 1% at the spectrum level, $i^2\text{Score}$ identifies more spectra than SEQUEST by 17%~78%, and more unique peptides by 13%~45%. Through additional experiments, $i^2\text{Score}$ also shows certain advantages over Mascot and KSDP previously proposed by our group.

Reversed-phase high performance liquid chromatography (RP-HPLC) coupled with tandem mass spectrometry is the commonly used analytical technique for protein identification. RP-HPLC can be employed to separate complex peptide mixtures in that different peptides show different retention times in the column due to their different physiochemical properties. The retention time for a peptide can be predicted from its amino acid sequence. Many efforts have been made to predict retention times for peptides under RPLC conditions and achieve good prediction performance. The difference between experimental and predicted retention times can be adopted as a measure of reliability for peptides. There exist many peptide scoring functions, however, a peptide scoring function incorporating peptide retention time is still not available. To further improve the peptide identification performance of the above scoring function, the difference between experimental and predicted retention times, $r\text{Diff}$, is incorporated into a new scoring function, $i^2\text{rScore}$, which is a linear combination of three features, $i^2\text{Score}$, Δi^2 (see section 3.2.2), and $r\text{Diff}$. The weight coefficients in the linear combination are dynamically determined by a linear classification SVM model trained on a specific spectral data set. The experiments show that $i^2\text{rScore}$ can identify more spectra and more unique peptides than $i^2\text{Score}$ by more than 10%.

Keywords: Computational Proteomics; Statistical Learning; Mass Spectrometry; Peptide Identification; Scoring Function