

# 摘要

质谱技术是当前蛋白质鉴定研究中使用最广泛的技术。而基于串联质谱鉴定肽序列进而鉴定蛋白质序列的数据库搜索引擎是最常使用的工具之一。本文针对数据库搜索引擎应用背景，以构建高通量、高精度的蛋白质鉴定系统为目标，系统地研究了引擎的效率优化问题和肽鉴定结果验证问题，提出了若干关键技术来提高搜索引擎的高通量数据处理能力以及最后鉴定结果的可靠性。

首先，本文提出了一种为蛋白质鉴定服务的数据库索引技术 **IndexToolkit**，用以解决直接检索 **FASTA** 格式数据库时的低效问题。该技术针对数据库引擎中的候选肽检索问题，以质量值为索引键值，通过表格、倒排文件分块技术等组织方法，兼顾质量值序列和肽序列在数据库引擎系统性能中的不同作用，为提升候选肽查询速度并提供了一个新框架。该框架下，给定一个质量值和质量误差阈值后，可以快速得到落入该质量窗口内的所有候选肽序列。在我们自主开发的搜索引擎 **pFind** 上的应用实践表明，该技术能有效提高候选肽序列检索效率，索引前后处理速度提升约一个量级（10 倍）。此外，针对单机计算平台日渐多 CPU 多核的趋势，本文突破以往搜索引擎常采用的串行体系结构，提出了一个整合索引查询、多线程计算和最新设计的批量数据处理流程模型“发车模式”的新架构。该架构下，本文充分发掘软件的并行性，并把并行的理念融入到系统的设计和实现中，使搜索引擎处理性能在索引加速的基础上还可提升约 10 倍。

基于搜索引擎 **pFind** 的正确/错误结果分布的统计分析，本文提出了一种针对 **pFind** 肽鉴定结果中的 **e-value** 进行自动化验证的方法。根据观察到的目标-诱饵反转序列数据库上查询结果的分布情况，本文提出采用高斯混合模型 (**GMM**) 和期望最大化 (**EM**) 算法计算肽鉴定结果的正确性概率和假阳性率的概率统计模型。该肽鉴定验证模型先利用 **GMM** 方法计算出正确结果（正向序列）中的两个正态分布（分别对应于高可信度序列和低可信度序列）各自的参数，拟合出正确结果分布曲线。然后，采用期望最大化算法学习得到错误结果（反向序列）分布的参数和曲线。最后根据这三个分布计算出每条序列的可能正确的概率以及假阳性率。本文模型的一个优势就是对于数据库中正、反向序列长度不一致的情况也能自适应地调整分布参数，灵活准确地评估每条鉴定序列的正确性概率。因此本文模型能够解决 **pFind** 系统中一直难以解决的小数据库上假阳性率计算问题。实验表明，该模型为解决小数据库上的肽鉴定结果验证以及假阳性率计算提供了一条有效途径。

针对现有搜索引擎的不足，本文给出新的自主研发的基于串联质谱数据的数

数据库搜索鉴定蛋白质序列软件 pFind。pFind 搜索引擎的突出特点是模块工具化和具备良好用户扩充性。本文给出了 pFind 的核心模块研究工作，同时详细剖析 pFind 各个版本（本地版、网络版、集群并行计算版本等）的结构框架和实现技术。通过在 LTQ 和 QTOF 质谱数据集上的实验表明，在相同的假阳性率条件下，pFind 软件的肽鉴定准确度和处理速度均超过目前流行的商业软件 SEQUEST。

**关键词：**生物信息学；搜索引擎；检索效率；高通量性能；肽鉴定结果验证；蛋白质序列数据库索引