

## 摘 要

利用质谱数据搜索蛋白质序列数据库是当前最常用的规模化蛋白质鉴定方法。近年来，蛋白质序列数据库的规模正在以指数级增长；质谱数据的生成速度不断加快；对非特异性酶切和多种类型翻译后修饰的鉴定需求也在不断增加，这些因素对蛋白质搜索引擎的速度提出了更高的挑战。因此，设计并实现一种高效的数据库搜索引擎成为蛋白质鉴定领域中一项重要的研究课题，其中索引系统是搜索引擎的关键组成部分，是提高检索效率的核心模块。本文从蛋白质序列数据库的索引策略出发，在分析现有索引方案的不足之处和蛋白质数据库检索特点的基础上，通过倒排索引组织蛋白质序列数据库，设计并实现了一种在时间和空间性能上获得显著提高的索引创建和查询方法，对提高蛋白质鉴定速度提供了基础性的技术支持。

本文首先设计了蛋白质索引，提高搜索引擎读取蛋白质信息的速度。蛋白质序列数据库通常以无结构的文本格式 **FASTA** 存放，该格式易于查看，却不利于计算机读取。蛋白质索引将蛋白质信息结构化表示并且分段存放，保证索引文件可以载入内存读取。测试表明，将索引文件载入内存之后，读取蛋白质信息的速度提高了 4 到 10 倍。

本文的主体部分是设计肽段索引，建立了肽段质量到序列的索引和肽段到蛋白质的倒排索引，提供高效的肽段查询接口。肽段索引保存非冗余的肽段，并且按照质量排序，通过肽段质量到序列的索引提高根据谱图母离子质量误差窗口查询肽段的速度，通过肽段到蛋白质的倒排索引提高肽段到蛋白质推断的速度。测试表明，肽段索引可以提高鉴定速度 2 到 5 倍。本文还设计了位向量索引存放非特异性酶切肽段。非特异性酶切肽段规模较大，采用常规的索引结构空间消耗较大，例如，**Swiss-Prot** 数据库的非特异性酶切肽段索引空间消耗约 100GB。位向量索引采用位 (bit) 来标记肽段，与常规结构相比空间消耗显著降低，**Swiss-Prot** 数据库的位向量索引空间消耗约 2GB。

本文精细地实现了以上设计方案，并经过大量的实际数据测试，与常用搜索引擎 **Mascot**、**SEQUEST** 和 **X!Tandem** 进行了性能对比。结果表明，本文实现的索引系统在创建索引的时间和空间性能，以及索引的加速效果等方面，超越了常用软件 **Mascot**、**SEQUEST** 和 **X!Tandem**。本文实现的索引软件工具 **pIndex** 已经成为蛋白质搜索引擎 **pFind** 的核心模块，为蛋白质搜索引擎的加速从索引方向提供了技术方案和实用软件工具的支持。

**关键词：**蛋白质组学，蛋白质鉴定，蛋白质数据库索引，倒排索引