

分类号 TP1 TP3 Q5 065

密级 _____

UDC _____

编号 _____

中国科学院研究生院 硕士学位论文

蛋白质鉴定搜索引擎的平台架构和并行加速研究

王乐珩

指导教师 贺思敏 研究员

中国科学院计算技术研究所

申请学位级别 工学硕士 学科专业名称 计算机应用技术

论文提交日期 2010年6月 论文答辩日期 2010年6月

培养单位 中国科学院计算技术研究所

学位授予单位 中国科学院计算技术研究所

答辩委员会主席 _____

声 明

我声明本论文是我本人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，本论文中不包含其他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

作者签名：

日期：

论文版权使用授权书

本人授权中国科学院计算技术研究所可以保留并向国家有关部门或机构送交本论文的复印件和电子文档，允许本论文被查阅和借阅，可以将本论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编本论文。

（保密论文在解密后适用本授权书。）

作者签名：

导师签名：

日期：

摘要

基于串联质谱的蛋白质鉴定是蛋白质组学研究中的关键技术。蛋白质序列库搜索引擎是串联质谱数据鉴定的主要工具之一。pFind 是国内第一个具有自主知识产权的蛋白质鉴定搜索引擎。pFind 引擎设计中，如何应用现代软件工程理论和设计模式，设计适应变化的体系架构，划分“低耦合、高内聚”的接口和模块，建立高效灵活的鉴定流程，将研究成果整合为工业级的软件系统，就成为一个关键课题。

本文首先对蛋白质序列库搜索进行系统分析，在此基础上，对 pFind 引擎展开面向对象设计，针对 pFind 自身特点进行架构，逐步明确模块接口和搜索流程，广泛应用面向对象设计模式和惯用法。在软件实现阶段，通过综合运用迭代开发、版本管理、缺陷管理、单元测试、双人编程、代码审核等多种现代软件工程实践，开发团队高效完成了超过 10 万行代码的 pFind 2.x 引擎及其周边工具的编写、测试和集成工作。

随着蛋白质组学研究的不断深入发展，各种因素导致蛋白质序列库搜索方法的计算量急剧增加，速度问题已经成为蛋白质组学的瓶颈之一。为了应对高通量蛋白质组学带来的大规模计算挑战，对蛋白质鉴定进行加速变得十分重要。在单机版架构的基础上，本文进一步开展了分布式并行加速方面的研究。通过对 pFind 引擎内部各个模块的热点进行分析，建立了针对 pFind 引擎内核的运行时间预测模型，并由此设计出动态和静态两种调度算法，在多个公共数据集和集群环境下，均实现了高处理器核数情况下的近似线性加速比。在一个实验中，鉴定含磷酸化肽的 100 个 Raw 格式文件，在 100 个处理器核的并行环境下，获得了超过 83 倍的加速比。在另一个更大规模的实验中，对含有磷酸化肽的 1,366,471 张质谱的进行鉴定，在 320 个处理器核的并行环境下，并行版 pFind 获得了 258 倍的加速比，加速效率超过 80%。

通过充分应用软件工程、并行计算、工业设计领域的方法，本文实现了 pFind Studio 系列的单机和并行搜索引擎及周边配套工具。这套软件已经投入生化科研一线实用并取得重要成果，为后续的计算蛋白质组学的研究提供了完整的科研平台，也为 pFind 系统今后的快速演进构建了机制保证。

关键词：蛋白质组学，蛋白质鉴定，设计模式，软件工程，并行计算

The architecture and acceleration of protein identification search engine

Leheng Wang (Computer Application Technology)
Supervised By Si-Min He

Tandem mass spectrometry (MS/MS) has become one of the most popular analytical techniques for protein identification in proteomics. Protein database searching is an important tool for tandem mass spectra identification. In our earlier work, we have developed the software pFind, which is employed as a platform in this paper. How to implement efficient and flexible software of pFind search engine becomes the key issue. To address this problem, the system analysis of protein sequence database search has been carried out. A framework in which every module accords to the object-oriented design patterns is built. Furthermore, a lot of modern software engineering practices, such as version control, bug management, pair programming are introduced to develop, test and release more than 100,000 lines of codes.

The efficiency of a protein identification search engine has become a bottleneck in high-throughput proteomics. As the computational demand increases, parallel computing has become an important technique for accelerating proteomic data analysis. To address the speedup problem, we develop a model to estimate running time. Based on this model, two effective load balancing methods, on-line scheduling and off-line scheduling, are developed. An experiment on a public dataset from PhosphoPep consisting of 100 RAW files of phosphopeptides showed that the speedup on 100 processors is about 83. The parallel version of pFind can complete the identification task within 9 minutes, while a stand-alone process on a single PC takes more than 10 hours. On another bigger dataset consisting of 1,366,471 tandem mass spectra, the speedup on 320 processors was about 258 and the parallelization efficiency was greater than 80%.

A lot of practical software tools, pFind Studio, have been developed. This software package provides a full supporting for proteomics. It has become a research and development platform of new algorithms.

Keywords: Proteomics; Protein Identification; Design Pattern; Software Engineering; Parallel Computing