MRUniNovo: an efficient tool for de novo peptide sequencing utilizing the hadoop distributed computing framework

Chuang Li1,†, Tao Chen2,†, Qiang He3, Yunping Zhu2 and Kenli Li1,*

1College of Information Science and Engineering, Hunan University, National Supercomputing Center in Changsha, Changsha 410082, China, 2State Key Laboratory of Proteomics, Beijing Proteome Research Center, National Center for Protein Sciences, Beijing Institute of Radiation Medicine, Beijing 102206, China and 3School of Software and Electrical Engineering, Swinburne University of Technology, Melbourne, Victoria 3122, Australia

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Associate Editor: Alfonso Valencia

Received on August 15, 2016; revised on October 30, 2016; editorial decision November 8, 2016; accepted on November 12, 2016

Abstract

Summary: Tandem mass spectrometry-based de novo peptide sequencing is a complex and time-consuming process. The current algorithms for de novo peptide sequencing cannot rapidly and thoroughly process large mass spectrometry datasets. In this paper, we propose MRUniNovo, a novel tool for parallel de novo peptide sequencing. MRUniNovo parallelizes UniNovo based on the Hadoop compute platform. Our experimental results demonstrate that MRUniNovo significantly reduces the computation time of de novo peptide sequencing without sacrificing the correctness and accuracy of the results, and thus can process very large datasets that UniNovo cannot.

Availability and Implementation: MRUniNovo is an open source software tool implemented in java. The source code and the parameter settings are available at http://bioinfo.hupo.org.cn/MRUniNovo/index.php

Contact: s131020002@hnu.edu.cn; taochen1019@163.com

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

In mass spectrometry-based proteomics, database-search based peptide identification methods suffer from poor performance in processing mass spectra with low signal-to-noise ratio, post-translational modifications and sequence variations. For example, only 6.2 million out of the 13 million spectra in the human colon and rectal cancer data produced by CPTAC (Zhang et al., 2014) can be identified through database search. In such cases, tandem mass spectrometry (MSMS) based de novo peptide sequencing, which is independent of a sequence database, has demonstrated a promising to directly derive peptide sequences from mass spectra that cannot be effectively and efficiently processed by database-search based methods. Thus, de novo peptide sequencing naturally complements database-search methods, especially in the identifying novel peptides and unsequenced organisms as well as confirming the correctness of the results produced by database-search methods (Zhang et al., 2012).

Rapidly developing proteomics studies have generated large volumes of experimental data (Riffle and Eng, 2009). For example, the data generated from the research on human protein expression profiling in 2014 exceeds 1 TB (Kim et al., 2014; Wilhelm et al., 2014). This can also be confirmed by the huge increase in mass spectra data recorded by PRIDE (Supplementary Fig. S1). Existing de novo sequencing tools cannot handle this large amount of spectra, due to their excessive computation time and memory overheads (Supplementary Fig. S2). This has become the major obstacle to the application of de novo sequencing in large-scale scenarios. Therefore, accelerating de novo peptide sequencing has become a critical issue in proteomics research. Fortunately, the recent
emergence of MapReduce, a novel parallel compute paradigm, provides the opportunity to reduce the computation time of many bioinformatics programs through parallelization, including database search based peptide identification programs, e.g. MS/MSPolygraph \cite{Kalyanaraman2011} and MR-Tandem \cite{Pratt2012}.

In this paper, we present MRUniNovo for large-scale \textit{de novo} peptide sequencing, which parallelizes UniNovo—a popular open source \textit{de novo} sequencing tool \cite{Jeong2013}. Based on Hadoop the de facto platform for big data processing in recent years \cite{Dean2008}, MRUniNovo distributes different parts of a mass spectra dataset to different machines to be sequenced concurrently. To our best knowledge, MRUniNovo is the first tool for parallel \textit{de novo} peptide sequencing. Experimental results show that, compared to UniNovo, MRUniNovo can significantly reduce the computation time of \textit{de novo} peptide sequencing without sacrificing the correctness and accuracy of the results.

2 Methods

UniNovo is a universal \textit{de novo} peptide sequencing tool which adopts a scoring algorithm based on a probabilistic model \cite{Jeong2013}. It can automatically filter out low-quality spectra and is applicable for all types of spectra and spectral pairs. We analyzed the workflow of UniNovo (Supplementary Fig. S3) and found that, when sequencing the same type of MS/MS spectra, it processed each spectrum individually. Accordingly, we designed MRUniNovo, a tool that utilizes multiple machines (i.e. Hadoop nodes) to process different parts of a MS/MS spectra dataset in parallel. Following the requirements of Hadoop, the execution of MRUniNovo consists of two phases. The first one is the sequential phase, where MRUniNovo partitions a MS/MS spectra dataset into properly-sized chunks and distributes them across multiple machines based on HDFS (Hadoop Distributed File System). The impact of the chunk size on the performance of MRUniNovo is illustrated in the Supplementary Figure S5. The second phase is the parallel part, where two types of tasks are executed in parallel, the Map tasks and Reduce tasks. A Map task scores candidate peptides and maps the results to key-value pairs in the form of \texttt{<peptide, score>}. A Reduce task summarizes the \texttt{<peptide, score>} pairs and identifies the peptides with the highest scores as the final results. The workflow of MRUniNovo is presented in Supplementary Figure S4. Based on HDFS, MRUniNovo maximizes its overall processing capacity and achieves high fault tolerance. It employs a simple yet efficient data rebalancing scheme to automatically move peptide data from overutilized DataNodes to underutilized DataNodes. MRUniNovo also detects and recovers from runtime faults automatically and quickly. This guarantees the correctness and accuracy of the results.

3 Result

We ran UniNovo on a single machine and MRUniNovo on a cluster of 6 machines running Hadoop 2.2.0 to process Fetal_Brain_brp_Elite, a MS/MS spectra dataset with high accuracy and high resolution \cite{Kim2013} at the National Supercomputing Center in Changsha to compare their performance measured by the computation time. Each machine has a dual-core 3.2 GHz Intel Xeon processor, 2 GB RAM and 500 GB Hard Disk Drive.

In the experiments, we increased the size of the dataset. Figure 1(a) demonstrates the performance of MRUniNovo versus UniNovo as the dataset size increases from 61 MB (13 327 spectra) to 976 MB (213 232 spectra). MRUniNovo took only 26.63 min to completely sequencing the 976 MB dataset, i.e. 133.45 spectra per second, while UniNovo took 290.92 min, i.e. 12.22 spectra per second. It also shows that MRUniNovo scales linearly with the dataset size while the computation time of UniNovo rockets when the dataset size exceeds 793 MB. The results from larger-scale experiments are not included in Figure 1(a) because UniNovo was not able to process datasets larger than 976 MB due to its dependence of the size of the RAM. The performance of MRUniNovo are further demonstrated in Figure 1(b), where the dataset size increases from 1 GB (234 962 spectra) to 12 GB (281 9544 spectra). As demonstrated, MRUniNovo can handle extremely large datasets with a linear increase in computation time with the dataset size. In the largest-case scenario with a 12 GB dataset, MRUniNovo took only 271 min, which is acceptable in most, if not all, cases. The impact of the number of machines in the Hadoop cluster on the performance of MRUniNovo is illustrated in Supplementary Figure S6.

We verified that the sequencing results obtained by MRUniNovo were consistent with those obtained by UniNovo. This validates that MRUniNovo achieves much higher performance than UniNovo without sacrificing the correctness and accuracy of the results.

4 Conclusions

As the size of the mass spectrum data increases rapidly, the low efficiency of \textit{de novo} peptide sequencing has become a critical issue. This paper presents MRUniNovo, a tool for parallel \textit{de novo} peptide sequencing based on Hadoop. Our experimental results demonstrate that MRUniNovo significantly reduces the time needed for \textit{de novo} peptide sequencing without sacrificing correctness and accuracy of the results.

Funding

This research was supported by the Key Program of National Natural Science Foundation of China (Grant No. 61432005), the National Outstanding Youth Science Program of National Natural Science Foundation of China (Grant No. 61625202), the International Science & Technology Cooperation Program of China (Grant Nos. 2013DFA11240, 2014DFB30010), the National High-tech R&D Program of China (Grant No. 2015AA015305), the Ministry of Science and Technology of China (Grant Nos. 2013CB910801, 2014DFB30010, 2015AA020108), and the National Natural Science Foundation of China (Grant Nos. 61303073).

Conflict of Interest: none declared.

References


