DOCTORAL THESIS

High throughput mass spectrometry based peptide identification search engine by GPUs

Li, You

Date of Award:
2015

Abstract

Mass spectrometry (MS)–based protein and peptide identification has become a solid method in proteomics. In high-throughput proteomics research, the "shotgun" method has been widely applied. Database searching is currently the main method of tandem mass spectrometry–based protein identification in shotgun proteomics. The most widely used traditional search engines search for spectra against a database of identified protein sequences. The search engine is evaluated for its efficiency and effectiveness. With the development of proteomics, both the scale and the complexity of the related data are increasing steadily. As a result, the existing search engines face serious challenges.

First, the sizes of protein sequence databases are ever increasing. From IPI.Human.v3.22 to IPI.Human.v3.49, the number of protein sequences has increased by nearly one third. Second, the increasing demand of searches against semispecific or nonspecific peptides results in a search space that is approximately 10 to 100 times larger. Finally, posttranslational modifications (PTMs) produce exponentially more modified peptides. The Unimod database (http://www.unimod.org) currently includes more than 1000 types of PTMs.

We analyzed the entire identification workflow and discovered three things. First, most search engines spend 50% to 90% of their total time on the scoring module, the most widely used of which is the spectrum dot product (SDP)–based scoring module. Second, nearly half of the scoring operations are redundant, which costs more time but does not increase effectiveness. Third, more than half of the spectra cannot be identified via a database search alone, but the identified spectra have a connection with the unidentified ones, which can be clustered by their distances.

Based on the above observations, we designed and implemented a new search engine for protein and peptide identification that includes three key modules. First, a parallel index system, based on GPU, organizes the protein database and the spectra with no redundant data, low search computation complexity, and no limitation of the protein database scale. Second, the graphics processing unit (GPU)–based SDP module adopts GPUs to accelerate the most time-consuming step in the process. Third, a k-means–based spectrum-clustering module classifies the unidentified spectra to the identified spectra for further analysis. As general-purpose high-performance parallel hardware, GPUs are

promising platforms for the acceleration of database searches in the protein identification process.

We designed a parallel index system that accelerated the entire identification process two to five times with no loss of effectiveness, and achieved around 80% linear speedup effect on the cluster. The index system also can be easily adopted by other search engines. We also designed and implemented a parallel SDP-based scoring module on GPUs that exploits the efficient use of GPU registers and shared memory. A single GPU was 30 to 60 times faster than the central processing unit (CPU)–based version. We also implemented our algorithm on a GPU cluster and achieved approximately linear acceleration. In addition, a k-means–based spectrum-clustering module with GPUs can classify the unidentified spectra to the identified spectra at 20 times the speed of the normal k-means spectrum-clustering algorithm.

Table of Contents