香港浸會大學理學院
HKBU Faculty of Science

for favour of posting

DEPARTMENT OF COMPUTER SCIENCE

PhD Degree Oral Presentation

PhD Candidate:	Mr You LI
Supervisor:	Dr Xiaowen CHU
External Examiner:	Dr Qiong LUO Dr Koon Kau CHOI (Proxy for Dr Meng-Qiu DONG)
Time:	14 Aug 2015 (Friday) 2:30 pm – 4:30 pm (35 mins presentation and 15 mins Q & A)
Venue:	RRS 732, Sir Run Run Shaw Bldg., HSH Campus

"High Throughput Mass Spectrometry Based Peptide Identification Search Engine by GPUs"

Abstract

Mass spectrometry (MS) based protein and peptide identification has become a solid method in proteomics. In the high throughput proteomics research, the 'shotgun' method has been widely applied. The main process contains following steps: mixed protein samples are proteolytically digested into peptides, which are subsequently separated by liquid chromatography, and then mass spectrometers scan all peptides (i.e., precursor ions) within an elution time to generate MS; in MS, the selected intense peaks of precursor ions are fragmented, and the fragment ions are scanned to generate tandem mass spectra (MS/MS spectra); the MS/MS spectra and their precursor masses are used to identify peptides, which are then inferred to corresponding proteins. Among all the tandem mass spectrometry-based protein identification in shotgun proteomics, database searching is currently the main method. The software of this method is a fundamental tool in proteomics related analysis. The most widely used traditional search engine searches spectra against the identified protein sequence database. The search engine is evaluated in two respects: 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 are facing serious challenges.

Firstly, the size of protein sequence database is ever increasing. Besides, along with the evolution and even revolution of genome sequencing technologies, proteogenomics research hopes to use genome translated protein sequences for protein identification. Secondly, the increasing demand of search against semi- or non-specific peptides results in a search space of ~ 10 or 100 times larger. At last, post-translational modifications (PTMs) produce exponentially more modified peptides. We analyze the whole identification workflow and find out: first, most search engines spend 50% ~ 90% of their total time on the scoring module, and that the spectrum dot product (SDP) based scoring module is the most widely used. Second, nearly half of the scoring operations are redundant, which cost more time but increase no effectiveness. Third, more than half of the spectra cannot be identified through database search only, but the identified spectra have a connection with the unidentified ones, which could be clustered by their distances.

Based on the above observation, we design and implement a new search engine for protein and peptide identification, including three key modules: first, index system organizes the protein database and the spectra, with no redundant data and low searching computation complexity. Second, the GPU-based SDP module adopts GPU to accelerate the most time consuming step in the whole process. Third, k-means based spectrum clustering module classifies the unidentified spectra to the identified ones for further analysis. As a general purpose and high performance parallel hardware, graphics processing units (GPUs) are promising platforms for speeding up database searches in the protein identification process. We designed the index system, which accelerated the whole identification process 2 ~ 5 times, with no loss of effectiveness. We also designed and implemented a parallel SDP-based scoring module on GPUs that exploits the efficient use of GPU registers and shared memory. Compared with the CPU-based version, we achieved a 30 to 60 times speedup using a single GPU. We also implemented our algorithm on a GPU cluster and achieved an approximately linear speedup. In addition, k-means based spectrum clustering module by GPUs can basically classify the unidentified spectra to the identified ones, with a 20 times faster than the normal k-means spectrum clustering algorithm.

***** ALL INTERESTED ARE WELCOME *****