

## DOCTORAL THESIS

# High throughput mass spectrometry based peptide identification search engine by GPUs

Li, You

*Date of Award:*  
2015

[Link to publication](#)

### General rights

Copyright and intellectual property rights for the publications made accessible in HKBU Scholars are retained by the authors and/or other copyright owners. In addition to the restrictions prescribed by the Copyright Ordinance of Hong Kong, all users and readers must also observe the following terms of use:

- Users may download and print one copy of any publication from HKBU Scholars for the purpose of private study or research
- Users cannot further distribute the material or use it for any profit-making activity or commercial gain
- To share publications in HKBU Scholars with others, users are welcome to freely distribute the permanent URL assigned to the publication

## **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

Abstract .....	ii
Acknowledgement.....	iv
Table of Contents .....	v
List of Figures .....	vii
List of Tables.....	ix
1. Introduction .....	1
1.1 Background of mass spectrometry–based protein identification.....	1
1.1.1 Biology background .....	1
1.1.2 Mass spectra .....	4
1.2 Mass spectrometry–based protein identification.....	8
1.2.1 Identification algorithms.....	9
1.2.2 Challenges.....	11
1.3 GPU computing .....	14
1.4 Literature review.....	16
1.5 The contribution.....	19
2. Index system .....	22
2.1 Overview .....	22
2.2 Protein Index.....	24
2.3 Peptide inverted index .....	26
2.3.1 Index structure and construction .....	28
2.3.2 Index query and experiment.....	31
2.4 Spectrum inverted index .....	35
2.4.1 Speed up peptide-precursor matching .....	35
2.4.2 Speed up fragment ion-peak matching .....	37
2.5 Parallel index on the CPU+GPU cluster.....	39
2.5.1 Parallel index construction .....	39
2.5.2 Parallel index query .....	41
2.5.3 Optimization of index module on GPU .....	42
2.5.4 Experiments .....	43
2.6 Conclusion .....	45
3. Accelerating scoring module by GPUs.....	46
3.1 Overview .....	47
3.2 Experiment.....	48
3.2.1 SDP on a Single GPU .....	49
3.2.2 SDP on the GPU cluster.....	52
3.3 Spectrum dot product on single GPU .....	54
3.3.1 SDP on the single GPU .....	55

3.3.2	Spectrum dot product on GPU clusters .....	59
3.4	Conclusion .....	61
4.	Spectrum clustering by k-Means on GPUs .....	62
4.1	Overview .....	62
4.1.1	Spectrum clustering .....	62
4.1.2	k-Means .....	63
4.2	k-Means on GPU .....	65
4.2.1	Finding closest centroid .....	67
4.2.2	Computing new centroids.....	71
4.2.3	Dealing with large dataset .....	73
4.3	Experiment on the simulation data.....	74
4.3.1	On low-dimensional data sets .....	74
4.3.2	On high-dimensional data sets .....	77
4.3.3	On large data sets .....	79
4.4	Applying k-Means to Spectrum clustering .....	80
4.4.1	Method .....	80
4.4.2	Experiments .....	81
4.5	Conclusion .....	82
5.	Conclusions and future work .....	85
5.1	Conclusions.....	85
5.1.1	Workflow optimization by indexing system .....	85
5.1.2	Paralleling the scoring system by GPUs .....	86
5.1.3	Analysis of the unidentified spectrum by spectrum-clustering adopting k-means on GPUs.....	88
5.2	Future work.....	89
5.2.1	System architecture .....	90
	Bibliography.....	93
	CURRICULUM VITAE .....	99
	Publications .....	100