Gene expression

**HiXCorr**: a portable high-speed $X_{\text{Corr}}$ engine for high-resolution tandem mass spectrometry

Hyunwoo Kim$^1$, Hosung Jo$^1$, Heejin Park$^2$*, and Eunok Paek$^2$

$^1$Department of Electronics and Computer Engineering and $^2$Department of Computer Science and Engineering, Hanyang University, Seoul, Korea

*To whom correspondence should be addressed.

Associate Editor: Ziv Bar-Joseph

Received on March 26, 2015; revised on August 12, 2015; accepted on August 14, 2015

Abstract

Summary: Peptide identification is an important problem in proteomics. One of the most popular scoring schemes for peptide identification is $X_{\text{Corr}}$ (cross-correlation). Since calculating $X_{\text{Corr}}$ is computationally intensive, a lot of efforts have been made to develop fast $X_{\text{Corr}}$ engines. However, the existing $X_{\text{Corr}}$ engines are not suitable for high-resolution MS/MS spectrometry because they are either slow or require a specific type of CPU. We present a portable high-speed $X_{\text{Corr}}$ engine for high-resolution tandem mass spectrometry by developing a novel algorithm for calculating $X_{\text{Corr}}$. The algorithm enables $X_{\text{Corr}}$ calculation 1.25–49 times faster than previous algorithms for 0.01 Da fragment tolerance. Furthermore, our engine is easily portable to any machine with different types of CPU because it is developed in C language. Hence, our $X_{\text{Corr}}$ engine will expedite peptide identification by high-resolution tandem mass spectrometry.


Contact: hjpark@hanyang.ac.kr

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

Proteomics (Wilkins et al., 1997) is the study of proteins, particularly expression, structures, functions and interactions of proteins. Because proteins play important roles in a human body, correct protein (sequence) identification (Steen et al., 2004) is very important. High-throughput protein identification is generally done by cleaving a protein into peptides, getting tandem mass (MS/MS) spectra of the peptides and analyzing the spectra to identify peptide sequences.

SEQUEST (Eng et al., 1994) is one of the most widely used computer programs for peptide identification from MS/MS spectrum analysis. It compares an experimental spectrum with theoretical spectra computationally created from sequences in peptide database, and finds the theoretical spectrum most similar to the experimental spectrum. To measure the similarity between the theoretical and experimental spectra, SEQUEST uses a sophisticated scoring scheme $X_{\text{Corr}}$ (cross-correlation).

However, calculating $X_{\text{Corr}}$ can be very slow and consumes most of the running time of SEQUEST. Thus, a lot of efforts have been made to overcome this speed issue. The original SEQUEST used fast Fourier transform algorithm (Cormen et al., 2001) to make the $X_{\text{Corr}}$ calculation faster. Later, Crux (Eng et al., 2008) improved the calculation speed of $X_{\text{Corr}}$ by using a precomputation table, which is also used in modern SEQUEST and TurboSEQUEST. Faster $X_{\text{Corr}}$ calculation is performed by Tide (Diament and Noble, 2011). It was optimized for x86 machine by including the x86 assembly code. Later, a portable Tide was developed in C language with exact P-value computation capability. (Hobert and Noble, 2014). To distinguish these two Tide versions, we will call the earlier version with x86 assembly code Tide-x86 and the later portable version Tide-C. Modern processors have multicores and support multithreading. Comet (Eng et al., 2013), an open-source MS/MS search tool by $X_{\text{Corr}}$, supported multithreading for $X_{\text{Corr}}$ calculation. Thus, the more processors and cores a machine has, the faster the Comet runs. Nowadays, more and more spectra are being acquired by high-resolution mass spectrometers. For example, Q-Exactive Orbitrap hybrid mass spectrometers (Thermo Scientific, Bremen, Germany)
HiXCorr: a portable high-speed Xcorr engine

2 Results

In this article, we present a portable hi-speed XCorr engine, which does not create a mass bin array altogether, instead, calculates XCorr directly from the peak list. Thus, it runs in O(p) time where p is the number of peaks in a spectrum, while all the previous engines are based on XCorr algorithms running in O(mf) time where m is the precursor mass and f is the fragment tolerance (pseudocodes are available in the Supplementary Data).

We compared our XCorr engine with previous engines on a machine with an Intel Core i7-3770K CPU (3.50 GHz) and 32 GB RAM under the CentOS 6.6 operating system and the GNU C compiler 4.4.7. First, we implemented our XCorr engine into Tide-C and named it Tide-Hi. We compared Tide-Hi, with Tide-C, and Tide-x86. Since Tide-x86 does not calculate the exact P-value, we compared them without exact P-value calculation. Figure 1a and Supplementary Table S1 show that Tide-Hi is 49 times faster than Tide-C’s XCorr engine, 20 times faster than Comet-Sparse’s and 11 times faster than Tide-x86’s XCorr engine (Fig. 1 and Supplementary Data). Furthermore, our engine is easily portable to almost every machine because it is developed in C. Optimizing our engine for x86 machines by embedding an x86 machine code can be a future research topic. Since XCorr score is widely used in peptide identification, this article may be useful for the community. Finally, we did not trade correctness for efficiency. Our XCorr engine calculates the same XCorr score as Tide and Comet do (Supplementary Theorem 2).

3 Conclusion

We present a portable high-speed XCorr engine for high-resolution tandem mass spectrometry by developing a novel algorithm, which enables XCorr calculation 2.45–49 times faster than before for 0.01 Da fragment tolerance. When the fragment tolerance is 0.001 Da, our engine runs 1000 times faster than Tide-C’s XCorr engine, 20 times faster than Comet-Sparse’s and 11 times faster than Tide-x86’s XCorr engine (Fig. 1 and Supplementary Data). Furthermore, our engine is easily portable to almost every machine because it is developed in C. Optimizing our engine for x86 machines by embedding an x86 machine code can be a future research topic. Since XCorr score is widely used in peptide identification, this article may be useful for the community. Finally, we did not trade correctness for efficiency. Our XCorr engine calculates the same XCorr score as Tide and Comet do (Supplementary Theorem 2).

Funding

This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2012-0006999) and also by the National Research Foundation of Korea [NRF-2012M3A9B9036676, NRF-2014R1A2A1A11054147, NRF-2012M3A9D1054452].

Conflict of Interest: none declared.

References
