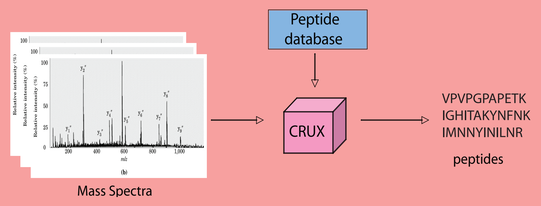
Improved candidate peptide generation using a precomputed peptide database.

Christopher Park, Aaron Klammer, William Stafford Noble
Department of Genome Sciences & Computer Science, University of Washington, Seattle

Abstract

We have implemented a software toolkit called Crux that uses a pre-computed peptide database that yields fast peptide searches for large proteomes when searching tandem MS spectra data against a sequence database.



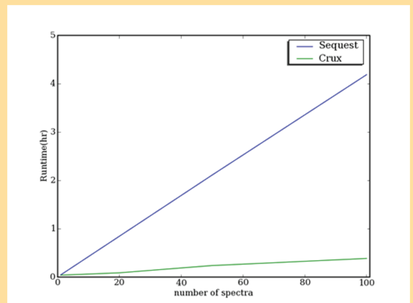
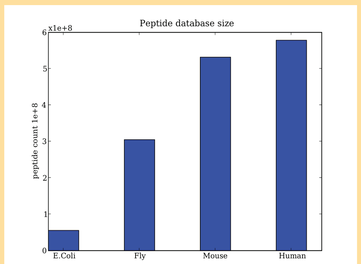


Figure 1. Crux vs Sequest (number spectra)

Introduction

In algorithms that are used to search tandem mass spectra data against a sequence database, a major bottleneck is the identification of candidate peptides within a given mass window. Unfortunately, the search space for a given proteome is quite large, as shown in figure 2.



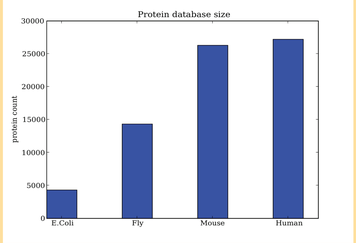


Figure 2. Protein, peptide database size

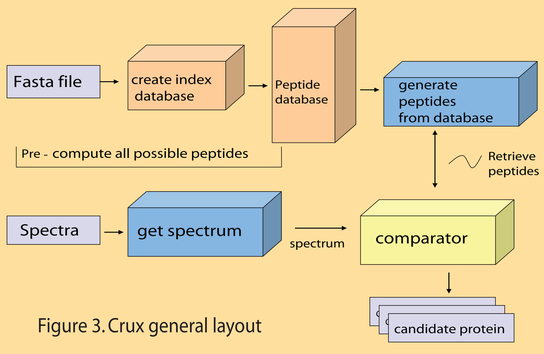


Figure 3. Crux general layout

We present an efficient method for identifying peptides in a particular mass window by using a pre-computed peptide database. Our method reduces the runtime dramatically for large sequence database searches compared to Sequest (Eng et al. 1995). Other methods that avoid re-generating peptides for each search require large amounts of disk space (e.g. TurboSequest).

Methods

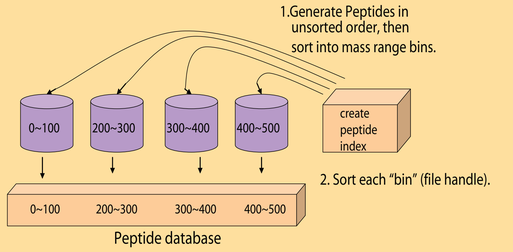


Figure 4. Creating peptide database

>gi|6381999|ref|NP_009311.2|
MLMKVPVPGPAPETKDIKNLYESIMNNYINILNKYTI
NINKDNINKLKFLDNYTEEEKGYYLSGLFEGDGNIYL
ANKIGHITAKYNFNKELTAVKWNIMKKKEQEVFMNY

protein index, (peptide start index: 4), (peptide length:11), mass
protein index, (peptide start index: 65), (peptide length: 12), mass
...

Figure 5 . Peptide database serialization

Searching with Crux requires two phases.

1. A user creates a peptide database on disk for the sequence database of interest. Instead of storing the sequence of individual peptides, as in TurboSequest, Crux only stores the protein index, the start index and the peptide length (Figure 5). This allows quick retrieval of peptide sequences from the database, while maintaining a manageable peptide database size.

2. Once the peptide database is on disk, retrieving all peptides within a given mass window can be accomplished efficiently. This use of the peptide database eliminates the need to re-scan the sequence file for each search.

Results

We test Crux compared to Sequest.

First, using Crux, we precomputed a peptide database for the tryptic peptides in the nr-db (06/02/2007, 3,292,818 proteins). This procedure required 19hrs and 14GB of disk space. (Figure 6)

We then tested our method by comparing its runtime to Sequest For both methods, we performed searches with varying number of spectra, using a mass tolerance of 3. (Figure 1)

We also tested Crux using a smaller mass tolerance window, such as might be used with high-resolution mass spectrometry instrumentation. (Figure 7)

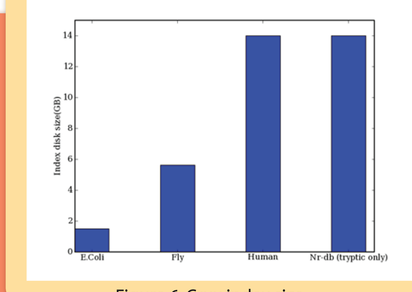


Figure 6. Crux index size

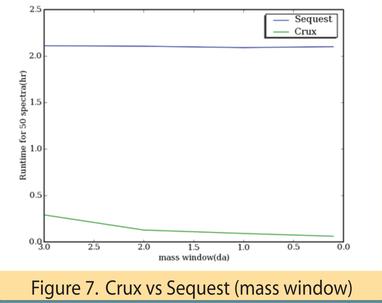


Figure 7. Crux vs Sequest (mass window)

Conclusion

When compared to current searching methods, our method uses a compact pre-computed peptide database that yields fast peptide searches for large proteomes. We are currently developing new scoring methods for peptide, spectrum matches that can utilize the new Crux infrastructure.

Reference
Jimmy K. Eng, Ashley L. McCormack and John R. Yates, III An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database JASMS, Volume 5, Issue 11, November 1994, Pages 976-989