Peptide Sequence Tag Identification Using the Cell BE

Robert Peace, Hanan Mahmoud, and James R. Green
Systems and Computer Engineering, Carleton University, Ottawa, Canada
{rpeace@connect.carleton.ca, hamahmoud@connect.carleton.ca, jrgreen@sce.carleton.ca}

Tandem mass spectrometry (MS/MS) is an analytical technique of great importance in the area of proteomics and systems biology. When coupled with suitable software systems, it is possible to determine the sequence of a protein from an unknown mixture. Currently, mass spectrometry techniques employ an abundance-driven data collection methodology to select ions for detailed analysis. Because only a limited number of fragments can be identified during an MS/MS experiment, and the system has no way of determining if it is collecting useful or important fragment identities, results from abundance-driven MS/MS can be ambiguous. Furthermore, since the database is searched offline and after the fact, results from abundance-driven MS/MS are not immediately available. Thus, there is room for improvement in both the speed and accuracy of MS/MS. If computational protein identification could be sufficiently accelerated, then real-time identification results may be used to guide data collection within the MS/MS instrument, leading to improved accuracy and efficiency. The Cell BE processor may provide the processing power required to achieve this so-called hypothesis-driven mass spectrometry.

In order to improve the accuracy and speed of MS/MS systems, it is imperative that the process is moved from an abundance-driven methodology to a hypothesis-driven methodology – where fragment identities are processed in real-time once collected and this data is used to guide the MS/MS system through its iterations in order to minimize the likelihood of ambiguous results. The hypothesis-driven methodology moves all of the data processing steps online, removing the delay between the MS/MS process and results. There are three main steps that must be accelerated for hypothesis-driven MS/MS: 1) *de novo* sequencing to generate the sequence tags from the observed spectra; 2) searching the proteomic database in order to obtain a list of potential protein matches given the observed peptide sequence tags to date; and 3) processing the list of potential protein matches in order to determine which peak will be maximally informative and should therefore be investigated by the MS/MS instrument.

This study focuses on step 2: the identification of all proteins containing a specific peptide sequence tag. As string matching plays a critical role in peptide sequence tag identification, in this study we have implemented, optimized, and evaluated several leading string matching algorithms on the Cell BE in order to determine the feasibility of using the Cell BE for hypothesis-driven mass spectrometry. The characteristics of the present challenge, such as small processing code size, predictable memory access patterns, and parallel computations, make it highly suitable for implementation on the Cell BE.

Several string matching algorithms were explored, including the shifting substring, Boyer Moore, Rabin Karp and tree-based algorithms. Performance of each algorithm was measured using the cycle-accurate Cell BE simulator. Furthermore, dynamic profiling was used to evaluate the effectiveness of tailoring each algorithm implementation to the unique architecture of the Cell BE. Optimizations such as parallelization across SPEs, SIMDization, double buffering, branch hinting, and loop unrolling provided impressive speedup for many string matching algorithms. Conversely in some cases, algorithms such as Boyer Moore proved very difficult to

parallelize and little speedup was observed. This study also examines the properties of MS/MS derived peptide sequence tags, wherein queries are only a few characters long and commonly contain wildcards, and protein databases, which contain thousands of proteins composed of several hundred characters each in some cases. The impact of these characteristics on the optimization of each algorithm is also discussed.

Among the string searching techniques that this project explores is the Parabix approach, an innovative technique that is currently under development at Simon Fraser University primarily for XML stream parsing. This study represents the first application of the Parabix approach to a bioinformatics application. An extension to the Parabix approach, where an orthogonal bit encoding scheme is used for each of the 20 amino acids, proved to be the most efficient string matching algorithm.

In tests which have been run using simulated hardware decrementers, the Parabix approach as implemented provides a speedup of six times over an optimized Rabin Karp implementation, and a speedup of nine times over an optimized shifting substring implementation. The orthogonal extension to the Parabix approach, when parallelized across 8 SPEs, searches a 1MB sample database in 213μs. The speed achieved by the orthogonal Parabix implementation is sufficient in order to meet the real-time deadlines of hypothesis-driven mass spectrometry.