A Free MS/MS *de novo* Sequencing and Protein Identification Online Server

Mingjie Xie[1]; Weiming Zhang[1]; Weijie Yang[1]; Weiwu Chen[1]; Gilles Lajoie[2]; Bin Ma[2]

[1]Bioinformatics Solutions, Inc, Waterloo, CANADA;[2]University of Western Ontario, London, CANADA

Overview

By distributing the computation to multiple computers, *de novo* sequencing and database search throughput are increased remarkably. We describe a free server for high-throughput MS data interpretation supporting both *de novo* sequencing and database search approaches.

Introduction

Mass spectrometry has become an essential tool for protein identification in proteomics. Two complementary approaches, *de novo* sequencing and protein database search, exist for the data analysis. Both commercial and free software have been developed for each of these tasks. However up to now, no free software could perform both approaches in a single package. Based on the PEAKS algorithms, we developed the PEAKS Online software and launched a free online service for both *de novo* sequencing and database search, among other functions. Users can submit data and retrieve results through a user-friendly interface using their web browsers. Since its launch in September 2007, several hundred researchers worldwide have used the service freely and thousands of searches have been performed.

Method

To enable parallel processing, Java J2EE technology was employed. PEAKS Online uses the PEAKS algorithms to perform the data analyses. It first divides the search into multiple smaller sub-tasks. A sophisticated algorithm is used to distribute those sub-tasks to the cluster nodes. The cluster nodes finish the sub-tasks using the same *de novo* sequencing and protein identification technology as the acclaimed PEAKS Studio. Once all sub-tasks have been completed, PEAKS Online provides a complete report with the combined results from each search. The results are identical to the commercial software PEAKS, but can be viewed and saved through the web interface. Here we present its performance in different cluster environment.

Result

A dataset containing 2810 MS/MS spectrums was used as test data. The performance was evaluated in three different server configurations, containing one, two and three nodes, respectively. Each node is a dual-core desktop computer and runs two processing threads. The Swiss-Prot database was used for the protein identification. In the single node configuration, the *de novo* sequencing step required 1101 seconds while the database search took another 193 seconds. In the two-node configuration, the first step needed 783 seconds and the protein identification took an additional 125 seconds. In the three-node configuration, the first step was reduced to 323 seconds and the second to 101 seconds. Thus, as expected, the throughput was increased significantly by adding more nodes as shown in figure 1.

PEAKS Online has a highly customizable yet very user friendly web interface. It supports a wide range of instruments and data formats. Users can define their own enzymes, PTMs and even upload their own FASTA databases. A task queue is available to users showing their searches' status. Upon search completion, a notification email will be sent to the user. All search results are stored in the users' profiles for later viewing.

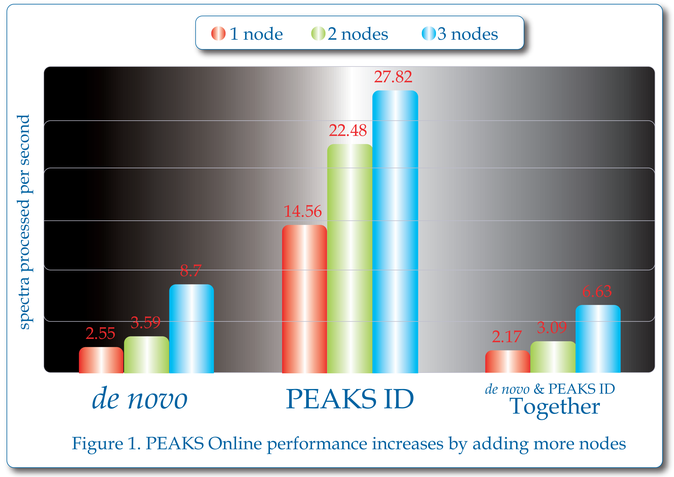
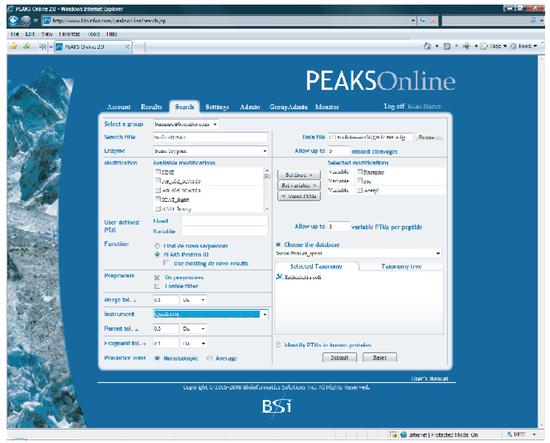
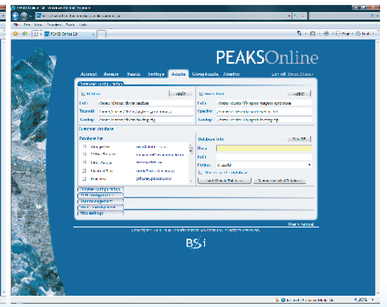


Figure 1. PEAKS Online performance increases by adding more nodes



PEAKS Online 2.0 Search Pane

PEAKS Online 2.0 Result Pane



PEAKS Online 2.0 Admin Pane

PEAKS Client 4.5 SP2
(PEAKS Online Desktop Extension)

The public PEAKS Online server is freely accessible at www.bioinfor.com/peaksonline