

Automated *de novo* Sequencing Using ToF-ToF MS/MS Data

Jennifer Locke¹, Jason Rogalski¹, Lei Guo², Bin Ma³, Juergen Kast¹, Gilles Lajoie³

¹University of British Columbia ²Bioinformatics Solutions Inc. ³University of Western Ontario

Overview

PEAKS software works well for both *de novo* sequencing (with no protein database) and protein identification (with a protein database) with MS/MS data obtained from a MALDI ToF/ToF instrument.

Introduction

Peptide *de novo* sequencing using tandem mass spectrometry data is a standard approach in proteomics. However, most of the automated *de novo* sequencing software including the software provided by the mass spectrometer manufacturers, is designed for *de novo* sequencing using Q-ToF and ion-trap generated data. Due to the different peptide fragmentation in ToF-ToF instruments, it was uncertain whether or not the *de novo* sequencing software PEAKS, using parameters optimized for other instruments, will work for ToF-ToF data. In this poster we demonstrate that the Q-ToF based internal parameters of the PEAKS software work very well for both *de novo* sequencing and protein identification even on the average quality of ToF-ToF MS/MS data.

Methods

Our proteins were reduced, alkylated and digested in solution with trypsin. They were then mixed with a saturated solution of α -cyano-4-hydroxycinnamic acid and spotted. MALDI analyses were carried out on an Applied Biosystems 4700 Proteomics Analyzer.

The ToF-ToF spectra were exported into text files that include the precursor ion information and the peak list. The PEAKS 2.4 software was then used to centroid and deisotope the spectra, and perform *de novo* sequencing on each spectrum. The default Q-ToF parameters on PEAKS 2.4 were used.

In order to test the *de novo* sequencing performance, only those spectra that could be assigned manually to peptides of the experimental proteins were used. The other spectra were removed due to either low signal to noise ratio or their inability to be assigned to any peptides from the given proteins.

However, when testing the performance of protein identification by database search, all spectra given by the instrument were used.

Result 1. *de novo* Sequencing of BSA Digest

The first experiment was done using bovine serum albumin (BSA) digest. Thirteen peptides were successfully fragmented and manually assigned to BSA peptides, with 6000 shots/spectrum acquired, using either atmosphere or UHP argon for collision gas.

The table lists the expected sequences, the sequences obtained by PEAKS' *de novo* sequencing on spectra obtained with UHP argon as collision gas, and the sequences found from spectra obtained with atmosphere as collision gas, respectively. The red letters indicate the amino acids that PEAKS software correctly computed. The percentage in the table is the score PEAKS software gave for that sequence. The table showed that PEAKS software, with default Q-ToF parameters, can be used for *de novo* sequencing of ToF-ToF spectra without any modification.

	Expected Sequence	Argon	Atmosphere
927.49	YLVEIAR	YLVELAR	YLVELAR
1163.62	LVNELTEFAK	LVNELTEFAK	LVLGWTEFAK
1249.60	FKDLGEEHFK	SSTDLGEEHFK	SSTLDWEEHFK
1305.69	HLVDEPQNLK	HLVDEPQNLK	HLVNEFLMFR
1439.81	RHPEYAVSVLLR	RHPEYAVSPDLR	RHPEYAVSVLLR
1480.77	LGEYGFQNALIVR	LGEYGFQDALLVR	LGEYGFQDALLVR
1567.72	DAFLGSLFYEYSR	DAFLGSLFYEYSR	DAFLGSLFYEYSR
1639.92	KVPQVSTPTLVEVSR	ARQPVSTPTLVEVSR	QVPARSTPTLVEVSR
1724.82	MPcTEDYLSLLNLR	MPcTEDYLSLLNLR	MPcTEDYLSLLNLR
1778.81	SQYLQcPFDEHVK	QSYLQcPFDEHVK	QSYLQcPFDEHVK
2019.97	KLDPNLTLeDEFKADEK	QLVLNLTLeDEFKWR	KLDPNLTLeDEFWcEK
2247.94	EccHGDLLcADDRADLAK	FVMFAYQVcWDRWLAK	ETTTKVVLeWcDKPASDR
2612.21	VHKEccHGDLLcADDRADLAK	ESDPLTFcMDKNLcADDKVRDR	QQcTRNYGDLLcADDKSWPR

Result 2. *de novo* Sequencing of ADH, MYG and CYC Digest

This time we measured alcohol dehydrogenase (ADH, yeast), myoglobin (MYG, horse) and cytochrome C (CYC, horse) digests with and without atmospheric collision gas. In total 23 spectra were successfully manually assigned to the peptides of the three proteins. Thirteen of them were collected with CAD gas turned off, and ten were collected using a CAD gas. PEAKS' *de novo* sequencing results are the following.

	Expected Sequence	CAD gas ON	CAD gas OFF
968.59	EALDFFAR	EALDFFAR	EALDFFAR
1168.76	TGPNLHGLFGR	NPGLHGLFGR	TGPNLHGLFGR
1251.79	SISIVGSYVGNR	LSSLVGSYVGNR	LSSLVGSYVGNR
1312.79	SLGGEVFDFTK	TVGWVLDFTK	LSGGEVLDFTK
1406.83	GIDGGEGKEELFR	no spectrum	AVDVGKEELFR
1470.79	TGQAPGFTYTDANK	TGQAPGFTYTDANK	GGTAAPGFTYTDANK
1502.77	HPGDFGADAQGAAMTK	no spectrum	HPGDFGADAQGAAMTK
1606.98	VEADIAGHGQEVLLR	VEVLAGHGQEVLLR	VEADLAGHGQEVLLR
1618.96	VLGDGGEGKEELFR	AVVLDGWGKEELFR	LVGLDGGEGKEELFR
1672.96	DLHAWHGDWPLPTK	no spectrum	FFAADAHGDWPLGLR
1816.03	GLSDGEWQVNLVWVGK	GLSDWVWQVNLVWVGK	GLSDDAWQVNLVWVGK
1885.14	YLFIQSDAIHVLHSK	KELLVLIQSDAIFLCK	EEFFLDALLHVLHSK
1911.06	HTDLHAWHGDWPLPTK	QQPLHADAHTAWLPTK	QQPLHAWHTAWPLGLR

Result 3. Sequencing by database search

We used the mixed spectra from ADH, MYG, and CYC to identify the proteins by searching in the SwissProt database. PEAKS software listed ADH1 (yeast), ADH2 (yeast), MYG (horse), and CYC (horse) as the top four proteins. The homologous proteins from other species were correctly classified as homologues of the correct proteins, which are still readable by clicking the "more..." button in the user interface.

It is worth noting that our ADH contains both forms of ADH1 and ADH2, whose sequences are highly similar. PEAKS software was able to identify both forms as existing proteins with high confidence, while it correctly classifies the ADH from other species as homologues. This is due to two unique peptides that are in ADH2 but not in ADH1. We also performed a Mascot search using the same data. Mascot could identify ADH1 (yeast), MYG (horse), and CYC (horse) as top proteins, but put ADH2 (yeast) after many other homologous proteins from other species. Therefore, it is hard for a user to tell whether ADH2 is there because it is real or because it is similar to the real protein ADH1.

Both PEAKS and Mascot identified MYG (rabbit). The reason PEAKS did not classify MYG (rabbit) as a homolog of MYG (horse) is because the peptide HPGNFGADAQAAMSK of MYG (horse) was deamidated a peptide of MYG (rabbit). When deamidation was specified as a variable modification, PEAKS correctly classified MYG (rabbit) as the homolog of MYG (horse).

Search results from database p1(0.1 0.1 Trypsin without PTMs)

Accession	Mass	Score	Coverage	Protein
PLDEBYA	34807.66	99%	13.51%	alcohol dehydrogenase (EC 1.1.1.1) 1 - yeast (Saccharomyces)
PLDEBYA	34695.63	97.97%	13.79%	alcohol dehydrogenase (EC 1.1.1.1) 2 - yeast (Saccharomyces)
PLMYGVA	14921.87	95.97%	30.72%	myoglobin [valdeius] - horse
PLCYCNO	11476.125	83.83%	51.92%	cytochrome c [valdeius] - horse
PLMYGVB	17065.96	74.7%	19.61%	myoglobin - rabbit (ventral empusoid)
PL_A65367	64000.218	21.99%	8.52%	transferrin protein (c-raf) - mouse

Unique peptides

Reference:

1. B. Ma, K. Zhang, A. Doherty-Kirby, C. Hendrie, C. Liang, M. Li and G. Lajoie, *Rapid Communication in Mass Spectrometry* 17(20): 2337-2342. 2003.