

A Comprehensive Comparison of the *de novo* Sequencing Accuracies of PEAKS, BioAnalyst and PLGS.

Bin Ma¹; Amanda Doherty-Kirby¹; Aaron Booy²; Bob Olafson²; Gilles Lajoie¹

¹University of Western Ontario, London, ON, Canada

²UVic Genome BC Proteomics Centre, Victoria, BC, Canada

1. Introduction

To identify proteins, a *de novo* sequencing algorithm computes the peptide sequences from MS/MS data without the need of a protein database. When proteins are heavily modified or from an organism whose genome is not sequenced, *de novo* sequencing is the only reliable approach to identify the proteins in a sample. *De novo* sequencing typically requires higher quality data than those required by a database search method. Therefore, a hybrid quadrupole time-of-flight (Q-TOF) instrument is most often used for measuring the MS/MS data. There are three commercial *de novo* sequencing software packages commonly used for the analysis of Q-TOF MS/MS data: BioAnalyst for the MDS Sciex/ABI QSTAR, PLGS for Micromass/Waters Q-TOFs, and PEAKS¹ for both. In this poster we compare the accuracies of the three packages.

2. Method

MS/MS spectra measured with a Micromass Q-TOF GLOBAL were analyzed by PLGS 2.0. Similarly, MS/MS spectra measured with a SCIEX API QSTAR Pulsar were analyzed by BioAnalyst (Analyst QS 1.11.). PEAKS 2.0 was used to analyse both datasets and the *de novo* sequencing results of PEAKS were compared with PLGS and BioAnalyst, respectively. In the analyses, each software outputs more than one sequence for each spectrum, but only the sequence with the highest score is used in this comparison. Three criteria were considered to evaluate the accuracy of each software:

- (1) correct amino acids
- (2) completely correct sequences,
- (3) partially correct sequences with five or more contiguous correct amino acids.

3. Experimental Results

Q-TOF GLOBAL was used to measure the MS/MS spectra for BSA_BOVIN and ADH_YEAST for the comparison of PEAKS and PLGS. A low filter (i.e. 10 cts/sec above background for the precursor ions) was used in the data collection and therefore a large number of spectra (265) were collected as the raw MS/MS data set. We then manually extracted all the spectra that have at least three strong y-ion matches with some peptides of the two proteins. The other spectra were discarded because they generally were of poor quality and we were not able to determine their peptides even knowing the protein sequences. Sixty-one spectra remained after this selection, and there are in total 764 amino acids in their sequences. Then both PLGS 2.0 and PEAKS were employed to compute the sequences *de novo*. Table 1 compares the performance of PEAKS and PLGS.

	PEAKS	PLGS
correct amino acids:	456	232
completely correct sequences:	13	7
partially correct sequences:	38	24

Table 1 PEAKS found more correct amino acids and sequences than PLGS

It is worth noting that because of our selection criteria, many of the 61 spectra are of lower quality than needed by *de novo* sequencing. The numbers shown in Table 1 are valid for the comparison of the two programs. But the low success rate cannot be interpreted as the low quality of either software. It is also interesting that PEAKS and PLGS are complementary to each other, reflecting different methods employed in the two programs. Table 2 shows the sequences that at least one of the two programs computed correctly. The complete results are not included here due to space limitations, but can be found at <http://www.csd.uwo.ca/~bma/peaks/asms04.html>.

m/z	z	correct	PEAKS	PLGS
484.7	2	EALDFFAR	EALDFFAR	EALDFmAR
618.7	2	DGGEGKEELFR	DGGEGKEELFR	DGGEGKEELmR
809.9	2	VLGIDGGEGKEELFR	VLGLDGGEGQEELFR	VLGLDGGEGQEELmR
464.3	2	YLYEIAR	YLYELAR	YLYELVK
418.7	2	IGDYAGIK	LG DYAGLK	seq not found
507.8	2	QTALVELLK	QTALVELLK	TKALVELLK
582.3	2	LVNELTEFAK	LVNELTEFAK	LVNELTVFTK
653.4	2	HLVDEPQNLIK	HLVDEPKNLLK	HLV ^{Pm} PKNLLK
740.4	2	LGEYGFQNALIVR	LGEYGFQNALIVR	LSVYGFKNALLVR
756.5	2	VPQVSTPTLVEVSR	VPQVSTPTLVEVSR	VPKVSTLRAAKVSR
540.2	2	STLPEIYEK	STLPELYEK	STLPEEFEK
567.2	2	VSEAAIEASTR	VSEAALEASTR	VSEAALEGSDR
567.3	2	VSEAAIEASTR	VSEAALEASTR	VSEAPSEASTR
496.7	2	TLPEIYEK	DVPELYEK	TLPELYEK
526.2	2	SIVGSYVGNR	LSVGSYNRR	SLVGSYVGNR
602.3	1	PETQK	EPTKK	PETQK
703.8	2	GIDGGEGKEELFR	GLDGGEGQEGANFR	GLDGGEGQEELFR

Table 2 The sequences that at least one of PEAKS and PLGS found correctly. The wrong amino acids were in black colour and struck through.

For the comparison of PEAKS and BioAnalyst, a SCIEX API QSTAR Pulsar was used to measure the MS/MS spectra for BSA_BOVIN and CYC_HORSE. Only the 6 most intense peaks of BSA_BOVIN and 7 most intense peaks of CYC_HORSE were selected for fragmentation. Therefore, only 13 spectra of good quality were collected. There are 150 amino acids in these sequences. Table 3 compares the performance of PEAKS and BioAnalyst. Table 4 lists the results of the two programs on the 13 spectra, where lower case "c" indicates a carboxyamidomethylcysteine.

	PEAKS	BioAnalyst
correct amino acids:	117	88
completely correct sequences:	8	2
partially correct sequences:	12	7

Table 3 PEAKS found more correct amino acids

m/z	z	Correct	PEAKS	BioAnalyst
464.2	2	YLYEIAR	YLYELAR	YLYELAR
482.7	2	EDLIAYLK	EDLLAYLK	EDLLAYLK
582.3	2	LVNELTEFAK	LVNELTEFAK	LVGGELTEFAK
450.2	2	LcVLHEK	LcVLHEK	LPESVGAk
570.7	2	ccTESLVNR	ccTESLVNR	ccTESLVGGR
512.2	3	LKEccDKPLLEK	LKEccDKPLLEK	LAG ^{Ecc} DAGPLLEK
722.8	2	YIcDNQDTISSK	YLcDNQDTLSSK	YLcDGGGADTLSSK
740.4	2	LGEYGFQNALIVR	LGEYGFQNALIVR	LGEYGF ^{GAGGPS} LVLR
728.8	2	TGQAPGFSYTDANK	TGQAPGAGASFGPPNK	TGGAAPGFHLTDAGGK
545.2	3	IFVQKCAQCHTVEK	CAQELACAKCHTVEK	TSSVTTGGVAGVGGAGVEK
528.9	3	KTGQAPGFSYTDANK	TGAGAGAPGFSYTDANK	GTGAAGAPGAYAGPGPAGGK
478.9	3	GEREDLIAYLKK	KMYVLNHA AFLK	QMAGDPDLLAYLK
1005.5	2	GITWGEETLMEYLENPK	AVTWGEETMFLTGGGDNPK	LGVSTGEETMMETEGTLPK

Table 4 The results of PEAKS and BioAnalyst.

Reference:

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