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

Bin Ma[1]; Amanda Doherty-Kirby[1]; Aaron Booy[2]; Bob Olafson[2]; Gilles Lajoie[1]
1. University of Western Ontario, London, ON, Canada 2.UVic Genome BC Proteomics Centre, Victoria, BC, Canada

Overview:

We compared three commonly used *de novo* sequencing programs, PEAKS, BioAnalyst and PLGS. The result showed that PEAKS has the best accuracy.

Methods

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:

- number of correct amino acids,
- number of completely correct sequences,
- number of partially correct sequences with five or more contiguous correct amino acids.

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 MicromassWaters Q-TOFs, and PEAKS [1] for both. In this poster we compare the accuracies of the three packages.

Experimental Result

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.

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 de novo sequencing results of both programs. The sequences are in general sorted by the spectrum quality. We regard an amino acid computed by the software is correct if the mass is approximately equal to the mass of the amino acid at the corresponding position of the correct sequence. For example, a letter Q is regarded as correct if it corresponds to a letter K in the correct sequence.

Experimental Result:

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.

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.

m/z	z	correct	PEAKS	PLGS
464.3	2	YLYEIAR	YLYELAR	YLYELVK
507.8	2	QTALVELLK	QTALVELLK	TKALVELLK
540.2	2	STLPEIYEK	STLPELYEK	STLPEEFEK
582.3	2	LVNELTEFAK	LVNELTEFAK	LVNELTVFTK
653.4	2	HLVDEPQNLIK	HLVDEPKNLLK	HLVPmPKNLLK
740.4	2	LGEYGFQNALIVR	LGEYGFQNALLVR	LSVYGFKNALLVR
756.5	2	VPQVSTPTLVEVSR	VPQVSTPTLVEVSR	VPKVSTLRAAKVSR
418.7	2	IGDYAGIK	LGDYAGLK	
567.2	2	VSEAAIEASTR	VSEAALEASTR	VSEAALEGSDR
567.3	2	VSEAAIEASTR	VSEAALEASTR	VSEAPSEASTR
618.7	2	DGGEGKEELFR	DGGEGKEELFR	DGGEGKEELmR
809.9	2	VLGIDGGEGKEELFR	VLGLDGGEGQEELFR	VLGLDGGEGQEELmR
484.7	2	EALDFFAR	EALDFFAR	EALDFmAR
703.8	2	GIDGGEGKEELFR	GLDGGEGQEGANFR	GLDGGEGQEELFR
496.7	2	TLPEIYEK	DVPELYEK	TLPELYEK
526.2	2	SIVGSYVGNR	LSVGSYNRR	SLVGSYVGNR
602.3	1	PETQK	EPTKK	PETQK
547.3	3	KVPQVSTPTLVEVSR	NMPQVLGPTLVEVSR	VKPKVSTPTLKKASR
626.3	2	SISIVGSYVGNR	LSSLVGSYVGNR	SLSLVGSFDGNR
501.3	2	ALKAWSVAR	SPKAWSVAR	
693.8	2	YICDNQDTISSK	YESDNQDTLSSK	YLmAPYPTLSSK
461.8	2	AEFVEVTK	AEFVEAEK	TVFKAKTK
675.8	3	KVPQVSTPTLVEVSRSLGK	QVPQVSTPTLVEPGGLAFGK	
656.8	2	SIGGEVFIDFTK	LSGGEVFDYPTK	
681.8	2	GAAGGLGSLAVQYAK	AGAGGLGSLAVYAGAK	TPDLGSSPVYAGAK
693.8	2	ANGTTVLVGMPAGAK	ELTTVLVGMPAGAK	
693.9	2	ANGTTVLVGMPAGAK	SQQTVLVGMPAGAK	
706.3	2	ADTREALDFFAR	WTVGEALDFFAR	
760.3	2	LGIDGGEGKEELFR	LGLDVGSGQEELFR	LGLDGEGGAGEYPER
771.3	3	ATDGGAHGVINVSVSEAAIEASTR	KNGDNPVHVMSVSEAAGQGASTR	LEDYLSDEDVVPCSALEASEK
507.2	2	ANELLINVK	AGGELLLNVK	
784.4	2	DAFLGSFLYEYSR	WFLGSFLDGAGGANR	SVFLGSGSLPFLSTR
820.5	2	KVPQVSTPTLVEVSR	QVPQVSTPNKAEWR	RAPKVSTMLRLLVR
824.8	3	QNCDQFEKLGEYGFQNALIVR	ETNGFGMKQLSVYGFKNALLVR	
841.2	3	LSQKFPKAEFVEVTKLVTDLTK	LSQKFPLSKFVEVTKLTVDLTK	
515.8	4	YTRKVPQVSTPTLVEVSR	YVVPGTALVTSAAKLVEVSR	
571.9	2	KQTALVELLK	AGAGTALVELLK	QKTVGKKLLK
450.5	3	IDGGEGKEELFR	SPVSTGKEELFR	
582.8	2	ISIVGSYVGNR	LSLVGSYGAAAR	SLLVGAANYTR
631.1	4	LSQKFPKAEFVEVTKLVTDLTK	LDKALGPVSLTVVGAAAPKGVTDLTK	
522.3	2	TVLVGMPAGAK	AELVGMAPGAK	
489.9	3	STLPEIYEKMEK	RAGNELYEAGMEK	
681.9	2	SLHTLFGDELCK	EAHTLFDGESEK	SLHTLSHAPGKSK
625.0	3	SPIKVVGLSTLPEIYEK	TGVLTAGPPDSVVAGMGSEK	SLSHNSATAPEPELYEK
447.2	2	DIPVPKPK	NGGPVPAGPK	MPPVPAGPK
596.8	2	LSTLPEIYEK	LSTLNPAGYEK	
536.3	2	EKDIVGAVLK	VGTDLRAVLK	
417.2	3	FKDLGEEHFK	HGAAGAGAPVNAEK	
438.5	4	LSQKFPKAEFVEVTK	FSVPGGPAGAGGVVPPGVTK	
450.8	2	PTLVEVSR	VVLDLVSR	
465.8	2	LKAWSVAR	LKADATGVR	
584.4	3	LSQKFPKAEFVEVTK	LPPAPKSKATSVLGGVTK	
642.4	2	HPEYAVSVLLR	LADVHSEVSAQK	
700.4	2	TVMENFVAFVDK	EALAGTRGSTHGDK	
747.0	2	FVEVTKLVTDLTK	FVTQGNGLAFPTLK	
767.7	3	NYQEAKDAFLGSFLYEYSR	SSVDPGPNLAGNAGSGGSGLGSGMVR	
434.2	3	TKEKDIVGAVLK	EQLDDGGVTAAPK	
483.3	3	FTKEKDIVGAVLK	MFNLQGGGGVARLK	
518.2	2	SDVFNQVVK	ASFVAAGSVVK	
550.0	4	AMGYRVLGIDGGEGKEELFR	TLRHAGGDTDAGGGGSGSGGRPTR	NAEGKDKYVQQGWEGAAFAK
582.8	2	ISIVGSYVGNR	LAEVTYGANVK	

Table 2. PEAKS and PLGS results on 61 Micromass QTOF spectra. Red fonts indicate the amino acids are correct. Orange area means the software found length ^{3}5 sequence tags *and* performed better than or equal to the other.

Table 3: Summary of Table 2. Peaks found more correct amino acids and sequences

	PEAKS	PLGS
Correct amino acids	456	232
Correct sequences	13	7
Sequences with length>4 tags	45	28

m/z	z	correct	PEAKS	BioAnalyst
482.7	2	EDLIAYLK	EDLLAYLK	EDLLAYLK
464.2	2	YLYEIAR	YLYELAR	YLYELAR
582.3	2	LVNELTEFAK	LVNELTEFAK	LVGGELTEFAK
450.2	2	LcVLHEK	LcVLHEK	LPESVVGAK
570.7	2	ccTESLVNR	ccTESLVNR	ccTESLVGGR
512.2	3	LKEccDKPLLEK	LKEccDKPLLEK	LAGEccDAGPLLEK
722.8	2	YIcDNQDTISSK	YLcDNQDTLSSK	YLcDGGGADTLSSK
740.4	2	LGEYGFQNALIVR	LGEYGFQNALLVR	LGEYGFGAGGPSLVR
728.8	2	TGQAPGFSYTDANK	TGQAPGAGASFGPPNK	TGGAAPGFHLTDAGGK
545.2	3	IFVQKCAQCHTVEK	CAQELACAKCHTVEK	TSSVTTGGVAGVGGAGVEK
528.9	3	KTGQAPGFSYTDANK	TGAGAGAPGFSYTDANK	GTGAAGAPGAYAGPGPAGGK
1005.5	2	GITWGEETLMEYLENPK	AVTWGEETMFLTGGGDNPK	LGVSTGEETMMETEGTLPK
478.9	3	GEREDLIAYLKK	KMYVLNHAAFLK	QMAGDPDLLAYLK

Table 3. PEAKS and BioAnalyst results on 13 MDS Sciex/ABI QSTAR spectra. Red fonts indicate the amino acids are correct. Orange area means the software found length ^{3}5 sequence tags *and* performed better than or equal to the other

	PEAKS	BioAnalyst
Correct amino acids	117	88
Correct sequences	8	2
Sequences with length>4 tags	12	7

Table 4: Summary of Table 4. Peaks found more correct amino acids and sequences