Protein ID: Comparing De Novo Based and Database Search Methods

I. Rogers, B. Ma.

Bioinformatics Solutions Inc.

Introduction

Using the correct tool for the job is as important in proteomics as it is in any other discipline. When identifying proteins from MS/MS data there are a number of tools to choose from. In the case where the data comes from a well studied organism, the researcher may choose a standard database search tool. In the case where the results from a database search are questionable, some validation is necessary. In a situation where no database program turns up a hit, the researcher must rely on de novo sequencing -- be it manually or using an automatic de novo software.

Peaks is a powerful and intuitive software package, combining remarkably accurate de novo sequencing with a new approach to protein identification. In this poster we prove Peaks' new method is able to identify proteins just as well as standard database search software. In this light, we compare Mascot and Peaks. Further, we show Peaks to be the ideal validation tool for standard database search software. Finally, an perhaps most importantly, we show Peaks to be the best automatic de novo software.

Method

The Mascot approach is a single-stage database search approach, directly comparing spectra with peptides theoretically digested from protein databases. The PEAKS approach first uses *de novo* sequencing to generate peptide candidates. These partial or complete sequences are used in an error-tolerant sequence search against a protein sequence databases. The resulting list of protein candidates is double-checked exhaustively against the spectra, and the resulting proteins are scored and grouped by similarity.

Three separate data-sets were used to compare Mascot with PEAKS. Two were from ESI-QTOF instruments, and the third was from an ion trap (LCQ) instrument. PEAKS Studio 2.2 was compared to the publicly available Mascot web-server.

Next, different MS/MS spectra measured with a Q-TOF instrument by PLGS 2.0. Similarly, MS/MS spectra measured with a QSTAR instrument were analyzed by BioAnalyst (Analyst QS 1.11.). PEAKS 2.4 was used to analyze both data-sets. Peaks' de novo sequencing results were compared with PLGS and BioAnalyst, respectively. In this comparison, only the highest scoring sequence from each software is used. Three criteria were considered to evaluate the accuracy of each software: number of correct amino acids, number of completely correct sequences, and number of partially correct sequences (>5 contiguous correct amino acids).

Run2 - QTof data

Mascot	***Peaks***
MYG_HORSE	CYC_BOVIN
CYC_EQUAS	MYG_HORSE
MYG_RABIT	ALBU_BOVIN
MYG_MACFA	IF2B_YEAST
MYG_ORCOR	ASNA_TREPA
MYG_BOVIN	ALBU_RABIT
MYG_OCHPR	YK23_ARCFU
CYC_APTPA	HT22_ARATH
MYG_KOGSI	
CYC_RANCA	
ALBU_BOVIN	
CYC_SQUSU	
MYG_POPH	
ALBU_FELCA	
CYC_KATPE	
MYG_APTFO	
MYG_TACAC	
CYC7_YEAST	

Proteins listed in RED are correctly identified, and non-redundant.

Run3 - LCQ data

Mascot	***Peaks***
ALBU_BOVIN	ALBU_BOVIN
PHS2_RABIT	OVAL_CHICK
ALBU_PIG	MLRS_RABIT more...
MLE1_RAT	MLRV_HUMAN more...
LACB_BOVIN	PHS2_RABIT more...
OVAL_CHICK	MLE1_RAT more...
ALBU_FELCA	LACB_BOVIN more...
ANT3_HUMAN	ANT3_HUMAN
MLRV_RAT	G3P_SHEEP more...
ALBU_CANFA	2ABA_PIG more...
MLE1_CHICK	PAK2_RABIT
MLRS_RABIT	A2HS_SHEEP more...
LACB_SHEEP	IF41_RABIT more...
2ABA_HUMAN	MLEF_HUMAN more...
IF41_HUMAN	IVD_HUMAN
G3P_PIG	BGAL_ECOLI
ALBU_HUMAN	GELS_PIG more...
ALBU_MUMAN	G3P2_HUMAN
ALBU_MOUSE	LCA_SHEEP more...
A2HS_BOVIN	DPOL_HSVI1
	XDHA_BACSU
	ECH1_RAT
	G3P_CANAL
	TRPE_SALTY more...

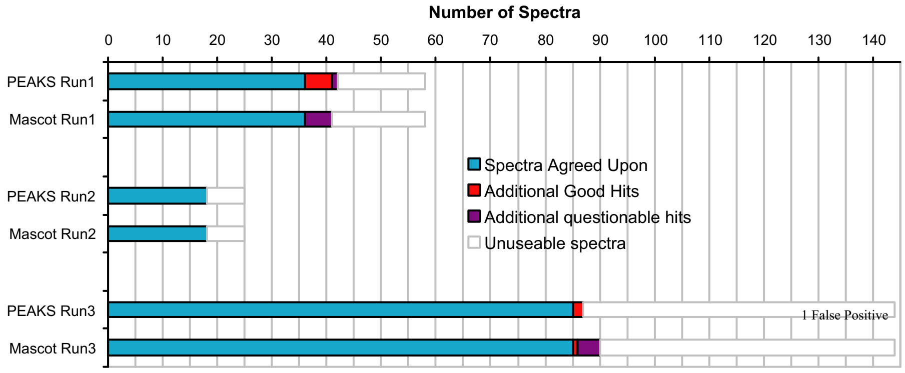
Table 1: Protein ID Results

Results

Table 1 summarizes each program's ability to correctly identify proteins and display them in a useful fashion. In the first run (not shown in Table 1) Mascot and PEAKS both identified the four proteins. In the second run (the in silico mix), PEAKS' top 3 results are the three correct proteins. Mascot, however, doesn't return ALBU_BOVIN until result #11. In the final run (the lcq data), PEAKS gives 19 non-redundant proteins in a reasonable order, followed by 6 incorrect proteins (with lower confidences). Mascot gives 12 good proteins and 8 redundant versions mixed together.PEAKS does a better job of grouping or ignoring redundant proteins, thus it sometimes leaves room for weaker hits in the top tento spectra.

Chart 1 compares the quality of protein identifications by gauging each program's ability to assign peptide sequences to spectra. In the first run, Mascot assigned a few extra plausible sequences where PEAKS did not, however the spectra are weak and it is difficul to confirm the correctness of these assignments. Two of PEAKS's extra hits show a clear pyridoxal phosphate modification, which Mascot was unable to detect. In the second run (the in silico mix), the data is very clean, and both Mascot and PEAKS performed excellently, correctly rejecting 7 uninformative spectra. In the third run (LCQ), PEAKS gave one false positive, Mascot gave three, but in all cases they gave low confidence.

Chart 1: Spectrum By Spectrum Comparison



De Novo Sequencing

The performance comparison between PEAKS and PLGS is summarized in Table 2 and Chart 2. Data collected from the instrument were filtered, and then selected manually for those with at lease three strong y-ion matches with the known protein sequence. Because of the selection criteria, many of the 61 spectra are of lower quality than needed by de novo sequencing. The de novo results 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. Note: for this study correctness is determined by matching masses, thus L is equivalent to I, K=Q, etc. There are 764 amino acids in these sequences.

The performance comparison between PEAKS and Bioanalyst is summarized in Table 3 and Chart 3. Only the 13 most intense peaks in the sample were collected. There are 150 amino acids in these sequences.

Peaks clearly shows a superior result quality as compared to PLGS and Bioanalyst when it comes to *de novo* sequencing.

Chart 2: Comparison with PLGS, % correct.

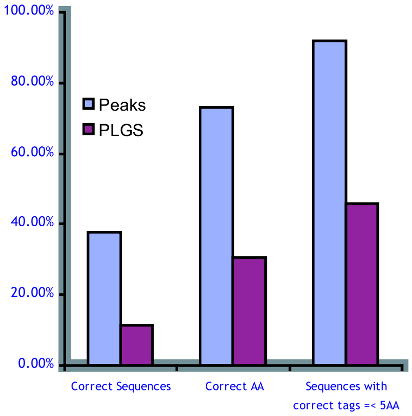
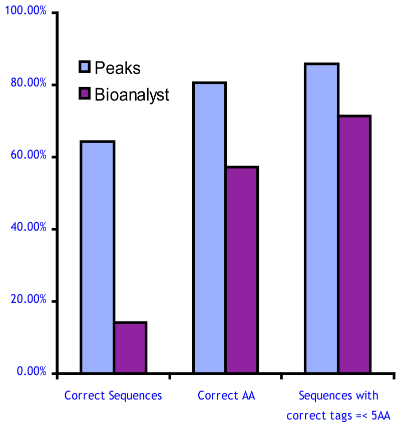


Chart 3: Comparison with Bioanalyst, % correct.



m/z	z	Correct Sequence	PEAKS 2.4	Peaks Score	PLGS
464.3	2	YLYEIAR	YLYELAR	100%	YLYELVK
582.3	2	LVNELTEFAK	LVNELTEFAK	100%	LVNELTVFTK
693.8	2	YICDNQDTISSK	YLCDNQDTLSSK	100%	YLmAPYPTLSSK
418.7	2	IGDYAGIK	LGDYAGLK	100%	.
484.7	2	EALDFFAR	EALDFFAR	100%	EALDFmAR
567.2	2	VSEAAIEASTR	VSEAALEASTR	100%	VSEAALEGSDR
567.3	2	VSEAAIEASTR	VSEAALEASTR	100%	VSEAPSEASTR
618.7	2	DGGEGKEELFR	DGGEGKEELFR	100%	DGGEGKEELmR
571.9	2	KQTALVELLK	KQTALVELLK	99%	QKTVGKKLLK
496.7	2	TLPEIYEK	LTPELYEK	98%	TLPELYEK
706.3	2	ADTREALDFFAR	WKEEALDFFAR	98%	.
461.8	2	AEFVEVTK	AEFVEVTK	97%	TVFKAKTK
540.2	2	STLPEIYEK	STLPELYEK	97%	STLPEEFEK
700.4	2	TVMENFVAFVDK	VTMENFVAFVDK	95%	.
653.4	2	HLVDEPQNLIK	HLVDEPQNLLK	92%	HLVPmPKNLLK
740.4	2	LGEYGFQNALIVR	LGEYGFQNALLVR	91%	LSVYGFKNALLVR
760.3	2	LGIDGGEGKEELFR	LGLDNEGKEELFR	90%	LGLDGEGGAGEYPER
507.2	2	ANELLINVK	ANELLLNVK	89%	.
584.4	3	LSQKFPKAEFVEVTK	LSQKFPKCPFVEVTK	85%	.
582.8	2	ISIVGSYVGNR	LSLVGSYVGNR	82%	SLLVGAANYTR
693.9	2	ANGTTVLVGMPAGAK	ANGTTVLVGMPAGAK	82%	.
507.8	2	QTALVELLK	QTALVELLK	79%	TKALVELLK
809.9	2	VLGIDGGEGKEELFR	VLGLDGGEGKEELFR	74%	VLGLDGGEGQEELmR
465.8	2	LKAWSVAR	LKAWSVAR	69%	.
681.9	2	SLHTLFGDELCK	VTHTLFDGELCK	68%	SLHTLSHAPGKSK
447.2	2	DIPVPKPK	NGGPVPAGPK	67%	MPPVPAGPK
626.3	2	SISIVGSYVGNR	LSSLVGSYVGNR	67%	SLSLVGSFDGNR
518.2	2	SDVFNQVVK	EKFAAVGSVK	59%	.
784.4	2	DAFLGSFLYEYSR	DAFLGSFLYEYSR	56%	SVFLGSGSLPFLSTR
841.2	3	LSQKFPKAEFVEVTKLVTDLTK	LVFPFVHNWLLHTKLTVDLTK	56%	.
450.8	2	PTLVEVSR	VVLVEVSR	54%	.
824.8	3	QNCDQFEKLGEYGFQNALIVR	ELCDQFEKLWYGFKNALLVR	53%	.
656.8	2	SIGGEVFIDFTK	LSGGEVFLDFTK	52%	.
602.3	1	PETQK	EPTQK	50%	PETQK
501.3	2	ALKAWSVAR	LAKAWSVAR	49%	.
820.5	2	KVPQVSTPTLVEVSR	QVPQVSTPNKAEDVK	49%	RAPKVSTMLRLLVR
526.2	2	SIVGSYVGNR	VTVGSYVGNR	46%	SLVGSYVGNR
547.3	3	KVPQVSTPTLVEVSR	RAPKVSTPTLVEVSR	45%	VKPKVSTPTLKKASR
625	3	SPIKVVGLSTLPEIYEK	TLHTPALSSNVAELYEK	45%	SLSHNSATAPEPELYEK
631.1	4	LSQKFPKAEFVEVTKLVTDLTK	MPKTVMLVYLVGNSAKLVTDLTK	43%	.
450.5	3	IDGGEGKEELFR	LWATGKEELFR	43%	.
438.5	4	LSQKFPKAEFVEVTK	EYKPVPDAMFPEVTK	40%	.
536.3	2	EKDIVGAVLK	VGTDLVGAVLK	40%	.
693.8	2	ANGTTVLVGMPAGAK	ANGTTVLVGMPAGAK	40%	.
756.5	2	VPQVSTPTLVEVSR	VPQVSTPTLVEVSR	39%	VPKVSTLRAAKVSR
703.8	2	GIDGGEGKEELFR	LGDGGEGQEELFR	39%	GLDGGEGQEELFR
417.2	3	FKDLGEEHFK	FKDLGEEHFK	38%	.
550	4	AMGYRVLGIDGGEGKEELFR	DVRYGNSPVGADGTDEQNFR	35%	NAEGKDKYVQQGWEGAAFAK
681.8	2	GAAGGLGSLAVQYAK	QAGGLGSLAVQYAK	33%	TPDLGSSPVYAGAK
489.9	3	STLPEIYEKMEK	TSLPELYEQMEK	30%	.
747	2	FVEVTKLVTDLTK	FVERDGLVTDTLK	29%	.
596.8	2	LSTLPEIYEK	LSLTPNQYEK	28%	.
515.8	4	YTRKVPQVSTPTLVEVSR	NNWWTVVTSAKALVEVSR	23%	.
767.7	3	NYQEAKDAFLGSFLYEYSR	MYNKCEPKDAGSFLYEYSR	20%	.
771.3	3	ATDGGAHGVINVSVSEAAIEASTR	ERYVEAAMAASVSEAALEASTR	20%	LEDYLSDEDVVPCSALEASEK
642.4	2	HPEYAVSVLLR	HPEYAVKLGNR	17%	.
675.8	3	KVPQVSTPTLVEVSRSLGK	KVPQVSTPTLVEVKPAFGK	17%	.
434.2	3	TKEKDIVGAVLK	DVAHGQHTNPPK	15%	.
522.3	2	TVLVGMPAGAK	LSLVGMPAGAK	15%	.
582.8	2	ISIVGSYVGNR	LSLVGSCSGDVK	12%	.
483.3	3	FTKEKDIVGAVLK	EKYNGRNVARLK		.

Table 2: Peaks and PLGS sequence results
Correct amino acids are in blue.

m/z	z	real	Peaks 2.4 beta	Peaks Score	Bioanalyst
464.2	2	YLYEIAR	YLYELAR	100%	YLYELAR
570.7	2	ccTESLVNR	ccTESLVNR	100%	ccTESLVGGR
582.3	2	LVNELTEFAK	LVNELTEFAK	100%	LVGGELTEFAK
722.8	2	YIcDNQDTISSK	YLcDNQDTLSSK	100%	YLcDGGGADTLSSK
740.4	2	LGEYGFQNALIVR	LGEYGFQNALLVR	100%	LGEYGFGAGGPSLVR
482.7	2+	EDLIAYLK	EDLLAYLK	100%	EDLLAYLK
728.82	2+	TGQAPGFSYTDANK	TGQAPGFSYTDANK	100%	TGGAAPGFHLTDAGGK
512.2	3	LKEccDKPLLEK	LKEccDKPLLEK	96%	LAGEccDAGPLLEK
1005.47	2+	GITWGEETLMEYLENPK	LGTWGEETLMGTFGGDNPK	95%	LGVSTGEETMMETEGTLPK
450.2	2	LcVLHEK	LcVLHEK	87%	LPESVVGAK
528.91	3+	KTGQAPGFSYTDANK	TGKAGAPGFSYTDANK	43%	GTGAAGAPGAYAGPGPAGGK
478.9	3+	GEREDLIAYLKK	KGTYGAHLLAYLK	42%	QMAGDPDLLAYLK
545.19	3+	IFVQKCAQCHTVEK	CAQDPTCAKCHTVEK	13%	TSSVTTGGVAGVGGAGVEK

Table 3: Peaks and Bioanalyst sequence results
Correct amino acids are in blue.

Conclusion

PEAKS provides a powerful tool in the study of proteins by mass spectrometry. When identifying proteins, PEAKS is closely matched with recognized database search technology. Both PEAKS and mascot identified all the proteins known to be present in the sample data. This demonstrates that a protein identification approach incorporating de novo sequencing can be at least as good as a standard database search method. When it comes to the quality of the protein identification results, we can see that both PEAKS and Mascot delivered, matching a large number of spectra to the proteins. PEAKS assigned some peptides that Mascot missed, and vice versa. Both programs performed well in detecting post-translational modifications. We cannot say from these results that either program is superior. Preference could only be determined subjectively, depending on the value one places on extra hits and false positive hits. Alternatively, using the two together means that you can get more out of your data: more coverage, more confidence. Where these two programs agree, using drastically different approaches, you can be sure that you have the answer.

When the sample in question came from an organism not in the databases, or failed to match a protein in the database, the researcher must turn to PEAKS for de novo sequencing. PEAKS provides 3 times as many correct sequences as PLGS and as many as 5 times that provided by bioanalyst.

Bioinformatics Solutions Inc. 145 Columbia St. West, Suite 2B Waterloo, ON, CANADA
University of Western Ontario, London ON, CANADA
Mascot public webserver at www.matrixscience.com
Peaks software demo available by request at: www.bioinformaticssolutions.com