



SEQUENCE **4.0**

de novo peptide sequencing software

**Protein identification using de novo sequencing
of proteolytic peptides by ESI-MS/MS**
Application Note

Protein identification using de novo sequencing of proteolytic peptides by ESI-MS/MS
SEQUIT! APPLICATION NOTE

Rodion Demine

Sequit! - software for *de novo* peptide sequencing by tandem mass spectrometry

© 2005, Proteome Factory AG, Berlin, Germany - www.proteomefactory.com

Based on an Invention by the Charité – Universitätsmedizin Berlin, Germany
Licensed by IPAL, Germany

PURPOSE

The major strategy in protein identification is to match mass spectrometric data of proteolytic protein fragments to protein or nucleotide databases. Purpose of this study is to test Sequit! *de novo* peptide sequencing with subsequent database search as alternative protein identification strategy.

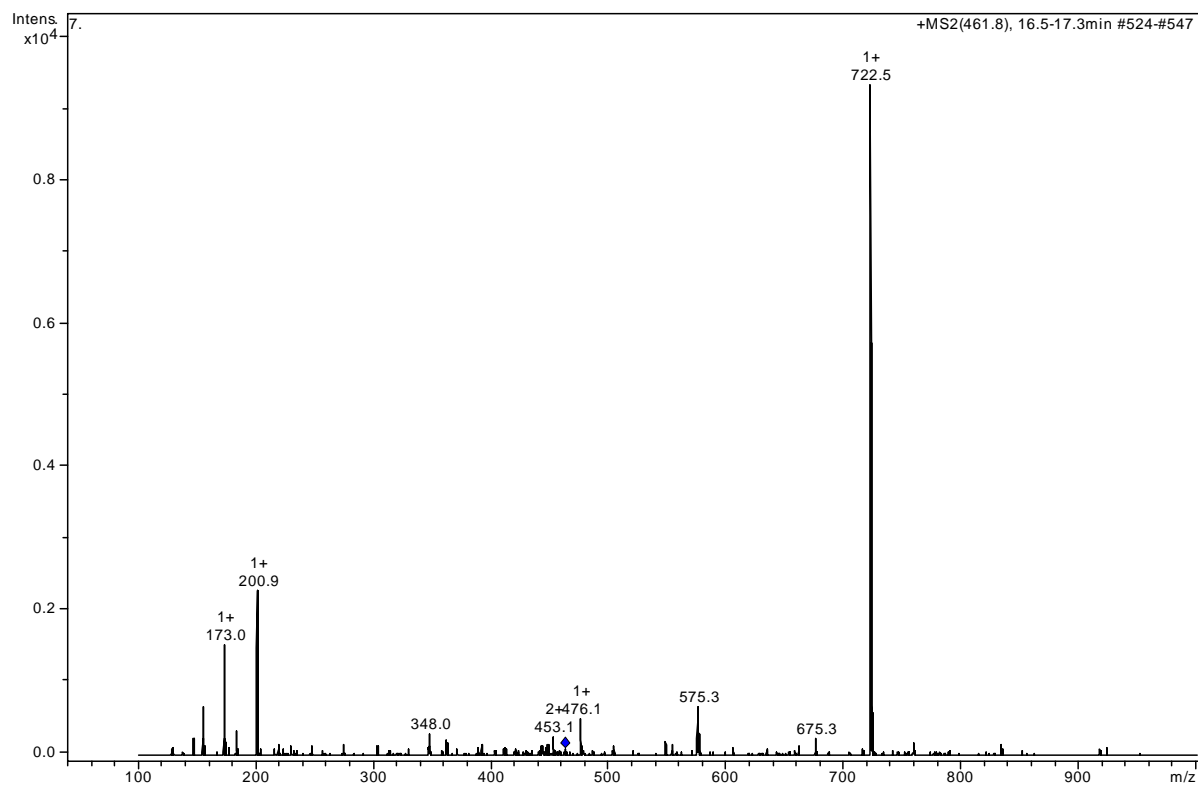
EXPERIMENTAL PROCEDURE

Serum albumin from *Bos Taurus* was digested with trypsin. Resulting mixture of proteolytic peptides was separated by HPLC and measured by ESI-MS/MS. ESI-MS/MS measurements were performed with a Bruker ESI-TRAP Esquire 3000 plus mass spectrometer. The accuracy of peptide mass measurement was 0.3 Da, and that of peptide fragment mass measurement was 0.5 Da. Number of exported most abundant non-deconvoluted ions of each MS/MS spectrum was decreased from default value of 50 to 8 for rapid *de novo* sequencing.

Sequit! output was set to 10 best sequences for each MS/MS spectrum. Data quality was set to "high" which means, that every sequence can have up to 3 gaps in b or y ions series of a size of two amino acids. Swiss-Prot database was used for batch BLAST database searches with *de novo* computed sequences. Database searches were limited to the *Bos Taurus* proteins (1). Sequit! results were compared to Mascot results (2). Mascot was set for analysis of trypsin digests.

70 MS/MS spectra (Cmpd, Compounds) were generated from raw data. If the precursor charge state was not determined, three copies of the same file with charge state 1+, 2+ and 3+ were generated. Spectra of peptides with $[M+H]^+$ below 700 Da were deleted. It resulted in 103 MS/MS spectra.

Example: Cmpd 7.

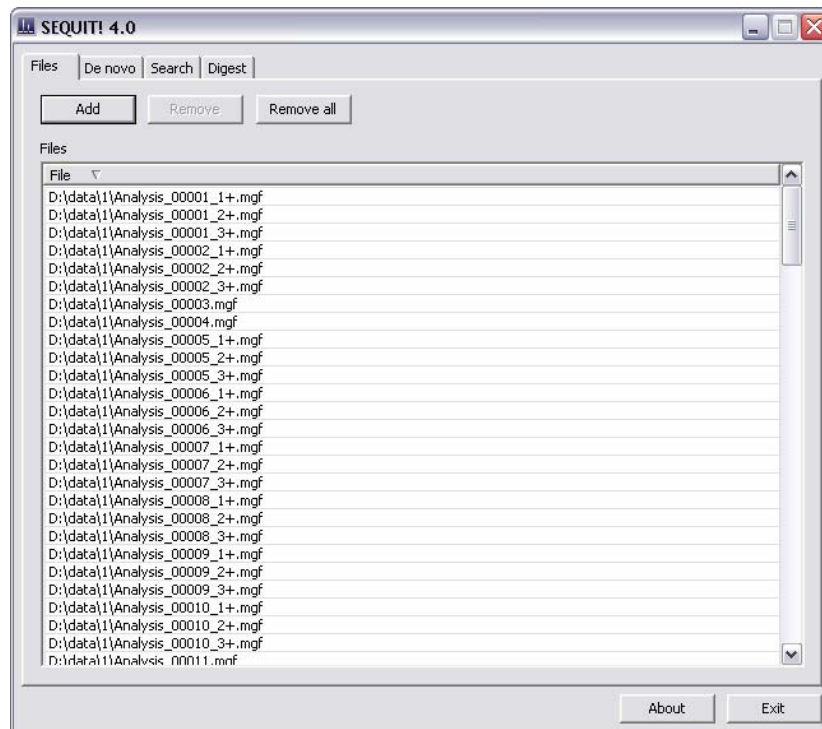


Peak list, which was generated from Cmpd 7.

Analysis_00007.mgf

```
BEGIN IONS
TITLE= Cmpd 7, +MSn(461.8), 16.8 min
PEPMASS=461.72 18065
CHARGE=2+
147.05 229
154.98 787 1+
172.95 1734 1+
183.87 410 1+
200.91 2479 1+
218.92 144
347.96 295
476.17 670 1+
548.13 180
575.33 672
576.37 393 1+
675.33 231
722.45 13125 1+
723.14 357 1+
759.39 159
833.47 141
904.13 432 1+
END IONS
```

All 103 input files were loaded to Sequit! and processed as batch performing successive *de novo* sequencing and database search steps for each file.



RESULTS

Sequit! computed 10 best sequences for each of 103 MS/MS spectra, which were searched in Swiss-Prot database by BLAST. All files were processed in about 3 minutes using Pentium 4 2.4 GHz computer.

Sequit! *de novo* sequencing and BLAST search result of Cmpd 7 are shown here as an example. Eight best sequences with score of 1.50 are congruent in C-terminal sequence tag FVEVTK. However, the correct sequence is unambiguously identified by subsequent batch database search with all proposed sequences for current mass spectrum.

SEQUIT! 4.0 - Analysis_00007.mgf - 922.4327

Files De novo Search Digest

Run Run All Options... Sequence... Autosearch Peptide tol. 0.3000
Fragment tol. 0.5000

Mass	Score	Match	TF Match	Coverage, %	Sequence
147.0500	1.50	10	10	55.55	AEFVEVTK
154.9800	1.50	10	10	55.55	CPFVEVTK
172.9500	1.50	11	10	61.11	EAFVEVTK
183.8700	1.50	10	10	55.55	LSFVEVTK
200.9100	1.50	10	10	55.55	PCFVEVTK
218.9200	1.50	10	10	55.55	SLFVEVTK
347.9600	1.50	10	10	55.55	TVPVEVTK
476.1700	1.50	10	10	55.55	VTFVEVTK
548.1300	1.42	9	8	50.00	AEFKNWK

Theoretical fragment masses

Ion	1	2	3	4	5	6	7	8
a-18								
a-17								
a								
b-18								
b-17		184.0604	331.1288	430.1972	559.2398	658.3082	759.3559	
b		201.0869	348.1553	447.2237	576.2663	675.3347	776.3824	
b+18								794.3930
	A	E	F	V	E	V	T	K
y		851.4510	722.4084	575.3400	476.2716	347.2290	248.1606	147.1129
y-17		834.4245	705.3819	558.3135	459.2451	330.2025	231.1341	
y-18								
immonium	44.0495	102.0550	120.0808	72.0808	102.0550	72.0808	74.0601	101.1074

Exit

SEQUIT! 4.0 - Analysis_00007.mgf - 922.4327

Files De novo Search Digest

Search Options... Organism any

Query	Score	Letters	Identities	Positives	BLAST Score	E
AEFVEVTK	1.50	8	8/8	8/8	28.20	0.02
CPFVEVTK	1.50	8	6/6	6/6	22.30	1.10
EAFVEVTK	1.50	8	6/6	6/6	22.30	1.10
ISFVEVTK	1.50	8	6/6	6/6	22.30	1.10
LSFVEVTK	1.50	8	6/6	6/6	22.30	1.10
PCFVEVTK	1.50	8	6/6	6/6	22.30	1.10
SIFVEVTK	1.50	8	6/6	6/6	22.30	1.10
SLFVEVTK	1.50	8	6/6	6/6	22.30	1.10

Query details

```

AEFVEVTK 1.50 8 8/8 8/8 28.20 0.02
>gi|1351907|sp|P02769|ALBU_BOVIN Serum albumin precursor (Allergen Bos
Length = 607

Score = 28.2 bits (59), Expect = 0.018
Identities = 8/8 (100%), Positives = 8/8 (100%)

Query: 1 AEFVEVTK 8
      AEFVEVTK
Sbjct: 249 AEFVEVTK 256
  
```

Exit

Batch sequencing results file was generated and imported into Microsoft Access template. Predefined query shows only proteins, which were identified by more than tree different *de novo* sequenced peptides. The analyzed sample was identified as *Bovine serum albumin precursor* (GI 135907).

File	Title	Ausdr1	Query	GI
Analysis_00005.mgf	Cmpd 5, +MSn(512.2), 15.6 min	1022.4127	TCTESLVNR	1351907
Analysis_00006_1+.mgf	Cmpd 6, +MSn(722.9), 16.5 min	722.46	FVEVTK	1351907
Analysis_00007.mgf	Cmpd 7, +MSn(461.8), 16.8 min	922.4327	AEFVEVTK	1351907
Analysis_00009.mgf	Cmpd 9, +MSn(681.9), 17.4 min	1362.5127	ETYGDMADFWK	1351907
Analysis_00009.mgf	Cmpd 9, +MSn(681.9), 17.4 min	1362.5127	ETYGDMADFWK	1351907
Analysis_00010.mgf	Cmpd 10, +MSn(682.4), 18.0 min	1362.5927	ETYGDMHWFVK	1351907
Analysis_00010.mgf	Cmpd 10, +MSn(682.4), 18.0 min	1362.5927	ETYGDMHWFVK	1351907
Analysis_00014.mgf	Cmpd 14, +MSn(654.0), 19.4 min	1305.8527	HLVDEPQNLIK	1351907
Analysis_00016.mgf	Cmpd 16, +MSn(653.4), 19.7 min	1305.9927	HLVDEPQNLIK	1351907
Analysis_00017.mgf	Cmpd 17, +MSn(464.6), 19.8 min	927.5127	YLYETAR	1351907
Analysis_00036.mgf	Cmpd 36, +MSn(582.6), 24.0 min	1163.9127	LVNELTEFAK	1351907
Analysis_00038.mgf	Cmpd 38, +MSn(582.9), 25.0 min	1163.6927	LVNELTEFAK	1351907
Analysis_00040.mgf	Cmpd 40, +MSn(582.8), 26.3 min	1163.7527	LVNELTEFAK	1351907
Analysis_00043.mgf	Cmpd 43, +MSn(508.3), 27.2 min	1014.6527	QTALVELLK	1351907
Analysis_00046.mgf	Cmpd 46, +MSn(741.3), 27.7 min	1479.8327	LIWYCFQNALIVR	1351907
Analysis_00046.mgf	Cmpd 46, +MSn(741.3), 27.7 min	1479.8327	LWYCFQNALIVR	1351907
Analysis_00046.mgf	Cmpd 46, +MSn(741.3), 27.7 min	1479.8327	WIYCFQNALIVR	1351907
Analysis_00046.mgf	Cmpd 46, +MSn(741.3), 27.7 min	1479.8327	WLYCFQNALIVR	1351907

Mascot identifies *Bovine serum albumin precursor* with extensive homology as well.

Sequit! output was compared to Mascot results.

Cmpd	RT (min)	[M+H] ⁺	Sequit!		Mascot		Bovine serum albumin
			Score	Sequence	Score	Sequence	
5	15.6	1022.41	2.00	TCTESLVNR	-	-	CCTESLVNR
6	16.5	722.46	1.83	FVEVTK	-	-	FVEVTK
7	16.8	922.43	1.50	AEFVEVTK	35	AEFVEVTK	AEFVEVTK
9	17.4	1362.51	1.45	ETYGDMADFWK	-	-	ETYGDMADCCCK
10	18.0	1362.59	1.20	ETYGDMWFWK	-	-	ETYGDMADCCCK
14	19.4	1305.85	1.72	HLVDEPQNLIK	73	HLVDEPQNLIK	HLVDEPQNLIK
16	19.7	1305.99	1.90	HLVDEPQNLIK	(59)	HLVDEPQNLIK	HLVDEPQNLIK
17	19.8	927.51	1.57	YLYEIAR	44	YLYEIAR	YLYEIAR
35	24.0	1142.91	-	-	9	KQTALVELLK	KQTALVELLK
36	24.0	1163.91	1.80	LVNELTEFAK	72	LVNELTEFAK	LVNELTEFAK
38	25.0	1163.69	1.30	LVNELTEFAK	(53)	LVNELTEFAK	LVNELTEFAK
40	26.3	1163.75	1.30	LVNELTEFAK	(43)	LVNELTEFAK	LVNELTEFAK
43	27.2	1014.65	1.11	QTALVELLK	33	QTALVELLK	QTALVELLK
46	27.7	1479.83	2.16	IWYGFQNALIVR	76	LGEYGFQNALIVR	LGEYGFQNALIVR
61	33.3	1399.73	-	-	14	TVMENFVAFVDK	TVMENFVAFVDK
total peptides				9		8	

Both Sequit! and Mascot identified sequences for compounds 7, 14, 16, 36, 38, 40 and 43 correctly.

The absence of y1, y2 and y3 fragments disallows complete sequence determination from Cmpd 46. However, Sequit! calculates 10 of 13 amino acids correctly, which leads to successful assignment of C-terminal sequence tag YGFQNALIVR to albumin.

Intra-molecular Cys-Cys bridges (indicated with **CC**) hindered the computation of complete sequences from Cmpd 5, 9 and 10. However, non-modified parts of peptides were sequenced correctly and allowed matching to albumin. No matches were found for Cmpd 5, 9 and 10 by Mascot.

Sequit! computed no sequences for compounds 35 and 61. However, compounds 35 and 61 were identified by Mascot with low ion score.

Additionally, Sequit! identifies Cmpd 6 as a non-tryptic albumin peptide FVEVTK.

CONCLUSION

De novo peptide sequencing of tryptic peptides with subsequent database search is a potent protein identification alternative to database searches with MS/MS data. In contrast to database searches with MS/MS, peptides, which deviate from database entries (e.g. modified and non-tryptic peptides), can be identified additionally by Sequit!.

REFERENCES

1. http://www.ncbi.nlm.nih.gov/BLAST/blast_FAQs.shtml#Batch
2. <http://www.matrixscience.com/>