A Comparative Study of Peptide Sequencing Software Tools for MS/MS

Chengzhi Liang[1], Jeffrey C. Smith[2,3], Christopher Hendrie[1], Ming Li[4], K. W. Michael Siu[2,3]

[1]Bioinformatics Solutions Inc., Waterloo, ON N2L 3L2 - [2]Centre for Research in Mass Spectrometry, York University, Toronto, ON M3J 1P3
[3]Department of Chemistry, York University, Toronto, ON M3J 1P3 - [4]School of Computer Science, University of Waterloo, Waterloo, Ont. N2L 3G1

Overview

- MS/MS spectra of known and unknown peptides were used to study the performance of several *de novo* sequencing and database search programs.
- Proteins of the unknown peptides are not in databases, thereby necessitating *de novo* sequencing for identification.
- Three *de novo* sequencing programs, PEAKS, BioAnalyst, and Lutefisk were compared in this study. PEAKS provided the most reliable and accurate results for high-quality ESI QqTOF data.

Introduction

A current bottleneck in proteomics is automated and accurate sequencing of enzymatically cleaved peptides. It is estimated that over two thirds of the MS/MS spectra produced by high-end quadrupole-TOF and TOF-TOF instruments in proteomics-research based corporations do not provide useful information [1]. An important contributing factor in this is the lack of high-quality software. The software currently available for MS/MS peptide sequencing mainly falls into two categories: (1) database searching by assigning a peptide sequence based on scoring against a protein (or peptide) database; and (2) *de novo* sequencing by deriving a (partial) sequence directly from an MS/MS spectrum. This study compares several programs representative of these two categories.

Methods

The cilia from the single-celled aquatic model organism *Tetrahymena thermophila* were isolated using dibucaine extraction, and their membranes removed using a detergent. The resulting proteins (comprised mostly of the well-characterized structural proteins α-tubulin and β-tubulin) were separated via 2-dimensional polyacrylamide gel electrophoresis, excised, and digested with trypsin. The resultant peptides were analysed and the α- and β-tubulin spots on the gel were identified. Peptides from these spots were subjected to MS/MS experiments on an MDS Sciex QSTAR QqTOF prototype mass spectrometer (MDS Sciex, Concord, ON) equipped with a nanoelectrospray ionisation source. Other types of peptides were also used, including twelve non-tubulin peptides from *T. thermophila* that were manually sequenced (denoted "unknown" peptides, *vide infra)*, one bovine trypsin autodigest peptide and bradykinin (Sigma). MS/MS spectra of the tubulin peptides were also obtained on an ABI MDS Sciex QSTAR XL QqTOF mass spectrometer (MDS Sciex, Concord, ON) equipped with a MALDI source. *De novo* sequencing of the raw MS/MS data was accomplished using PEAKS (v1.3, http://www.BioinformaticsSolutions.com/), BioAnalyst (v1.1, MDS Sciex, Concord, ON), and Lutefisk (v1.3.2 http://www.immunex.com/researcher/lutefisk.html). Additionally, the database search program Mascot (www.matrixscience.com) was used to identify the peptides and ascertain their sequences. The precursor mass error tolerance employed was 0.2 Da, and carbamidomethylation of cysteine residues was selected.

Results and Discussion

The primary goal of this study is to provide a thorough comparison and evaluation of these widely used *de novo* sequencing software tools with a common set of MS/MS data either generated from known proteins or have been manually analysed (Table 1).

Table 1: Number of Peptides Used in Study

	ESI	MALDI
Known**	24(3*)	24(6*)
Unknown***	12	0

*Peptides with PTMs (oxidation or pyro-Gln).
** Includes á- and â-tubulin and trypsin peptides, as well as bradykinin. This data may be viewed at http://www.bioinformaticssolutions.com/products/peaks/data/
***from *T. thermophila*, as described above.

Additionally, this study will probe the differences between, and the efficacy of, database-searching methods versus *de novo* methods. Database-searching methods have a distinct advantage in giving unambiguous peptide sequences (even I and L are differentiated); however, they are only applicable to proteins with known sequences. The results of this study (Table 2) clearly show that Mascot correctly identified most of the known peptides (36/48).

Table 2: Number of Peptides Correctly Identified by Mascot

	ESI	MALDI
Known	22/24	14/24
Unknown*	1/12	No data

*Match is not based on the full sequence, but only on those residues that are easily identified manually.

As expected, for peptides from unknown proteins, Mascot reported only one that resembles the peptide whose sequence was determined manually. The MS/MS data were also used to test *de novo* sequencing tools, PEAKS, BioAnalyst, and Lutefisk. The results are summarized in Tables 3 and 4. All spectra used had good signal-to-noise-ratios. However, in automated sequencing, some MS/MS spectra were apparently easier to interpret than others with all three programs giving good results, whereas others were more difficult for one or more programs (Table 5).

Conclusions

Based on our results, PEAKS has markedly better performance. It correctly identified more full-length sequences, more accurate sequence tags, and more single residues than the other two *de novo* sequencing software tools. Automated sequencing is easier on ESI data than on MALDI data. However, it is apparent that there remains a great deal of room for improvement in *de novo* sequencing software.

Reference

1. Kearney, P; Thibault, P *J Bioinf & Comp Biology 2003; (1): 183-200*

Table 4: No. of Residues Correctly identified in Unknown Peptides

	Correct residues*	Percentage in correct residues**	Percentage in all residues***
PEAKS1.3	97	92.4%	≥76.4%
BioAnalyst1.1	71	67.6%	≥56.0%
Lutefisk1.3.2	69	65.7%	≥54.3%
Mascot	46	43.8%	≥36.2%

*Verified by manual *de novo* sequencing.
**The number of correct residues unambiguously identified by manual sequencing is 105.
***The total number of residues based on manual sequencing is 127; some residues not included in this counting may also be correct.

Table 3: Number of Peptides/Residues Correctly Identified in the Known Peptides

Ionization method / sequencing tool		Correct + "almost correct" peptides*		Correct residues in "good" spectra***		Correct residues in all peptides**	
ESI	PEAKS1.3	4+6/21	47.6%	166/228	72.8%	171/246	69.5%
	BioAnalyst1.1	2+4/21	28.5%	121/228	53.1%	124/246	50.4%
	Lutefisk1.3.2	2+3/21	23.8%	101/228	44.3%	103/246	41.9%
MALDI	PEAKS1.3	4+3/19	36.8%	109/142	76.8%	121/206	58.7%
	BioAnalyst1.1	3+1/19	21.1%	76/142	53.5%	87/206	42.2%
	Lutefisk1.3.2	0+0/19	0	35/142	25%	38/206	18.5%

*"Almost correct" peptide: there exists at the most a pair of erroneous residues with the same or similar mass, e.g., N=GG, Q=GA, TL=VD, and AD=DV.
**Peptides with oxidation on methionine (3 for ESI and 5 for MALDI) are not used in this comparison. Methionine oxidation is easily identified from the precursor mass and the fragmentation pattern.
***In a "good spectrum", at least 4 consecutive residues were correctly identified. For ESI, the number of good spectra is 20 out of 21; for MALDI, this number is 12 out of 19.

Table 5: Peptides Correctly Identified by at Least One Program.

	Peptides	PEAKS1.3	BioAnalyst1.1	Lutefisk1.3.2
ESI	**DVNASIATIK**	DVNASLATLK	DVNASLADVK	[214.1]NASL[285.1]K
	LAVNLIPFPR	LAVNLLPFPR	LAVNLNCAPPK	[184.1]VNLLPFPR
	INVYYNEATGGR	LNVYYNEATGGR	LNVYYNEATGGR	LNVYYNEAT[142.0]K
	VAEQFTAMFR	VAEGAFTHDPR	VAEQFTAMFR	VAEKFTAMFR
	RPPGFFPSR	RPPGFFPSR	RPPGHPDER	RP[301.1]SPFR
	LSVDYGKK	LSVDYGAGK	LSVDYGAGK	LSVDYGKK
MALDI	QLFHPEQLISGK	QLFHPEQLLSGK	QLFHPEQLLSGK	[290.2]LLK[226.1]H[214.1
	INVYYNEATGGR	LNVYYNEATGGR	VGAVYYGGEATGGR	[354.2]ANEPEATNR
	YLTASALFR	YLTASALFR	YLTASALFR	[276.1]GPSAMFR
	TIQFVDWC*PTGFK	TLKFVDWC*PTGFK	QLTFVDSASTSSVTR	WRFVDWCSSVEK
	FPGQLNSDLR	FPGQLSDNLR	FPGQLNSDLR	[244.0][168.1]FPSDLR

* The residues in red are those correctly identified (using I=L, K=Q).