PEAKS: Powerful Software for Peptide *De Novo* Sequencing by MS/MS

Bin Ma[1*], Kaizhong Zhang[1], Christopher Hendrie[2], Chengzhi Liang[2], Ming Li[3],
Amanda Doherty-Kirby[4], Gilles Lajoie[4]

[1] Department of Computer Science, University of Western Ontario, London, ON, Canada N6A 5B7
[2] Bioinformatics Solutions Inc., Waterloo, ON, Canada N2L 3L2
[3] Department of Computer Science, University of Waterloo, Waterloo, ON, Canada N2L 3G1
[4] Department of Biochemistry, University of Western Ontario, London, ON, Canada N6A 5C1

Abstract

A number of different approaches have been described to identify proteins from tandem mass spectrometry (MS/MS) data. The most common approaches rely on the available databases to match experimental MS/MS data. These methods suffer from several drawbacks and cannot be used for the identification of proteins from unknown genomes. In this communication, we describe a new *de novo* sequencing software package, PEAKS, to extract amino acid sequence information without the use of databases. PEAKS uses a new model and a new algorithm to efficiently compute the best peptide sequence whose fragment ions can best interpret the peaks in the MS/MS spectrum. The output of the software gives amino acid sequences with confidence scores for the entire sequence as well as an additional novel positional scoring scheme for portions of the sequence. The performance of PEAKS is compared with Lutefisk, a well known *de novo* sequencing software, using quadrupole-time-of-flight (Q-TOF) data obtained for several tryptic peptides from standard proteins.

1. Introduction

Tandem mass spectrometry (MS/MS) is emerging as the most reliable tool to identify proteins. There are now several configurations of mass spectrometers that provide MS/MS data with sufficient mass accuracy to deduce peptide sequences of enzymatically digested proteins from low energy CID MS/MS spectrum. However, deducing peptide sequences from raw MS/MS data is slow and tedious when done manually. Instead, the most popular approach is to search database of known genomes with the un-interpreted experimental MS/MS data. A number of such approaches have been described, the most popular being Mascot[1] and Sequest.[2] These methods are effective but often give false positives or incorrect identifications. Searching databases with masses and partial sequences (sequence tags) derived from MS/MS data give more reliable results.[3] For unknown genomes, *de novo* sequencing must be done in order to get sequences or partial sequences. Full sequences can then be obtained by cloning the gene of interest.

The deduction of amino sequences from MS/MS spectra is dependent on the quality of the data and further complicated by poor fragmentation and inaccuracy due to mass shift caused by temperature and other instrumental parameters. To aid the assignment of sequences a number of

[*] *Correspondence to*: Bin Ma, Department of Computer Science, University of Western Ontario, London, ON, Canada N6A 5B7. Email: bma@csd.uwo.ca. Phone: (519) 661-2111 x. 86890. Fax:(519) 661-3515.

chemical techniques have developed to favor the formation of more stable "y" or "b" ions.[4,5] Isotopic labeling introduced in the tryptic digestion step can also be used to identify "y" ions.[6]

A number of algorithms and software packages have been reported for the deduction of protein sequences from MS/MS data.[7-15] Several instruments manufacturers have develop their own but these are in many cases unsatisfactory. One software package developed independently, Lutefisk, has gained a lot of attention.[10,11] Most of these software packages, including Lutefisk, use a graph theory approach. The spectrum is first translated into a "spectrum graph" where nodes in the graph correspond to peaks in the spectrum and two nodes are connected by an edge if the mass difference between the two corresponding peaks is equal to the mass of an amino acid. The software then attempts to find a path that connects the N and C termini, and to connect all the nodes corresponding to the y-ions (or b-ions). In this paper we describe another approach with a new mathematical model and software, called PEAKS for *de novo* sequencing of peptides from MS/MS data.

PEAKS performs *de novo* sequencing directly from the MS/MS data and therefore does not rely on a protein database. It computes the best possible sequence among all possible amino acid combinations. Analogous approaches have been described, but were computationally inefficient and abandoned.[13-15] Instead, PEAKS relies on a sophisticated dynamic programming algorithm to perform the computation efficiently. The mathematical model that PEAKS uses is also different than the graph theory approach. In our approach, PEAKS computes peptides whose ions corresponds to as many high abundance peaks in the spectrum as possible. We describe below the basic concepts behind this new PEAKS software and compare its performance with experimental MS/MS data with Lutefisk, another available software tool for *de novo* sequencing.

2. Method

The approach taken in PEAKS can be summarized into four steps: (1) preprocessing, (2) candidate computation, (3) refined scoring, and (4) global and positional confidence scoring. The first step consists of preprocessing of the raw MS/MS data. This involves a new method for noise filtering, peak centering, as well as deconvolution of the doubly and triply charged species to singly charged ions. This step is very important for the interpretation of MS/MS data by PEAKS. In fact we found a much higher success rate using raw data instead of using data preprocessed by various manufacturers' software. This indicates that optimal preprocessing of data is an important step for the *de novo* sequencing by MS/MS.

The second step, candidate computation, is the critical step in which the 10000 best sequences of all possible combinations of amino acids for a given precursor ion mass are computed. For this computation, the a, b, c, x, y and b/y-17/18 ions are considered. The basic assumption of our model is that the more high abundance peaks are matched by those ions of a sequence, the more likely the predicted sequence is the correct sequence. For each mass value m, this new algorithm first computes the reward/penalty that a y (or b) ion has mass m. If there is a peak close to m, the reward is equal to the logarithmic abundance of the peak, multiplied by a factor reflecting the mass error between m and mass value of the peak, and multiplied by a factor reflecting the co-existence of the x, y-H_2O, y-NH_3 (or a, c, b-H_2O, b-NH_3) ions. If there is no peak close to m, the reward is a negative constant value. The problem is then reduced to finding a sequence, such that its y and b ions maximize the total rewards at their mass values.

The initial mathematical formula to compute the reward at mass m was purely empirical but has been refined. Because the PEAKS algorithm is very modular, the modification or change of the formula for reward computation is relatively easy. In fact several formulas have been evaluated but we found the following formula to be satisfactory for Q-TOF MS/MS and therefore used in PEAKS 1.3.

$$f(\frac{h_1}{h})\times f(\frac{h_2}{h})\times f(\frac{h_3}{h})\times \exp\left(-(\frac{m'-m}{d})^2\right)\times \log h \qquad (1)$$

In Formula (1), m is the mass of a y-ion, m' is the mass of the observed peak for that y-ion, and d is the mass error tolerance of the spectrometer. Thus, the exponential factor in Formula (1) is designed to represent the mass error. h, h_1, h_2, h_3 denote the relative abundances of the observed y-ion peak and the corresponding x, y-H_2O, y-NH_3 peaks ($h_i = 0$ if the corresponding peak is not present). Thus, the logarithmic factor is designed to represent the relative abundance, and $f(\frac{h_i}{h})$ are designed to represent the presence of the x, y-H_2O, y-NH_3 peaks (supporting peaks). The choice of the function $f(x)$ was fairly arbitrary. Because we expect that the supporting peaks will have comparable abundance with the y-ion peak, in PEAKS 1.3, we chose $f(x)$ that has the curve shown in Figure 1. Thus, the supporting factor is no less than 1, but is greater than 1 when h_i is comparable with h.

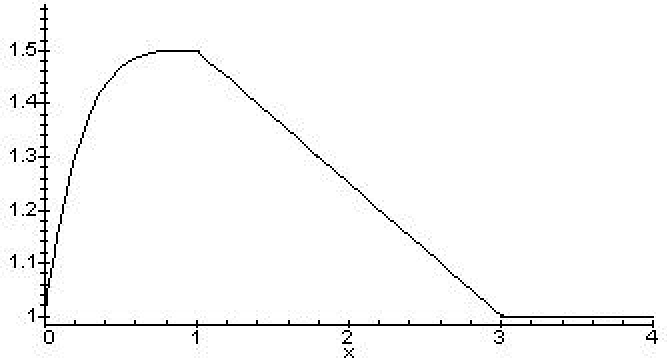

Figure 1. The curve of the supporting function $y = f(x)$ in PEAKS 1.3.

The rewards for b-ions are computed in the same way as for y-ions. The only difference is that we now have four supporting factors, a, c, b-H_2O and b-NH_3. Also, because y-ions are usually more abundant than b-ions for tryptic peptides on Q-TOF instruments, we multiply all the b-ion rewards by 0.5 to force the algorithm to use y-ions first to explain the mass of the fragments.

Our approach to tabulate the total reward is very different than the spectrum graph model used by previous *de novo* sequencing software and algorithms. Because the spectrum graph model attempts to find a path connecting the N and C termini, the absence of ions may break such a path and makes the finding of the sequence very difficult. However in our approach, a reward/penalty score is computed for every possible mass value, regardless of the existence of a peak around that mass value. Therefore, the absence of peaks does not cause major problems. Also, the reward/penalty score accounts for many factors like the abundance of the peak, the mass errors and the co-existence of other peaks, all of which significantly improve the accuracy of the *de novo* sequencing results. A modified version of the recently published *de novo* sequencing algorithm using dynamic programming (Ma *et al.* [16]) is used in PEAKS to compute very efficiently the 10000 sequences with the highest scores.

In the third step, each of the 10000 candidates is re-evaluated by a more stringent scoring scheme, and the best candidates (the number can be specified by users) under the new scoring scheme will be output. In this refined rescoring step, ion mass error tolerance is stricter. The rewards of immonium ions as well as internal cleavage ions are now considered. The reward/penalty computation is the same as y and b ions. The immonium and internal cleavage ions are not counted in the second step because their inclusion would be too computationally inefficient to derive the best 10000 candidates. Finally, a recalibration of the data is performed to account for minor deviation in the MS data. This recalibration method is similar to the one performed by Taylor *et al.* [11].

In the last step, PEAKS computes a confidence score for each of the top-scoring peptide sequences. The refined scores can be seen as non-normalized measures of the likelihood of correctness for each peptide, and the distribution of scores gives a measure of the overall probability of successful sequencing. PEAKS first converts the refined score x of each peptide sequence to a raw confidence X by the following formula $X = \exp(cx)$, where c is a parameter that is estimated from the spectrum by PEAKS. Then the raw confidence scores for all the top-scoring peptide sequences are normalized to be the final confidence scores so that they sum up to 1. Finally, the positional confidences for each residue are derived from consensus among the globally top-scoring sequences.

3. PEAKS' Input and Output

PEAKS can read tandem MS spectra in several different formats including Micromass .pkl files, Sequest .dta files, and Mascot Generic Format (.mgf) files. Data from other manufacturers can be input as text files. For each spectrum, PEAKS outputs a list of amino acid sequences that can possibly generate the MS/MS spectrum, from the most to the least likely sequence. The default number of output sequences in the list is five and can be changed by the user. PEAKS also associates each output sequence with a confidence level. The confidence level is a percentage number between 0% to 100%, indicating how likely the complete sequence is correct.

PEAKS also outputs a confidence level for each individual amino acid in the sequence using different colors. In the current version, an amino acid (one letter code) colored red indicates a 95% confidence to be correct, green correspond to 90%-95%, blue 80%-90%, and black less than 80%. (In this manuscript bold fonts are used to indicate the red, green and blue colors.) This unique feature allows a user to get very high confidence sequence tags even in cases where PEAKS cannot find the complete sequence with high confidence level due to poor quality of the experimental data.

4. Experimental Results

The internal parameters of PEAKS were initially adjusted using MS/MS data from known proteins. A blind test was then used to evaluate the performance of PEAKS' *de novo* sequencing. MS/MS data were obtained from a Q-TOF2 and a Q-TOF-Global mass spectrometer (Micromass, UK) with four standard proteins purchased from Sigma and digested in solution with trypsin: alcohol dehydrogenase (yeast), myoglobin (horse), albumin (bovine, BSA), and cytochrome C (horse). The results reported here were obtained with PEAKS version 1.3. The PEAKS software can be used on line free of charges at http://www.bioinformaticssolutions.com. The *de novo* sequencing software Lutefisk[11] was used as a comparison for the same set of data. Lutefisk was graciously provided by one of its authors through email contact.

For each protein, a collated data file of the MS/MS spectra was obtained as follows. For each precursor mass, the corresponding scans were combined automatically using the PeptideAuto.exe function of MassLynx 3.5. The peak list of this summed spectrum was copied using the Edit, Copy Spectrum List function of Masslynx 3.5 into Notepad. The precursor m/z was added at the very beginning of each peak list in the text file using the following format: m/z 0.000 z. The resulting text file then contains several peptide spectra for each protein. If the precursor ion masses of two spectra in the same file differ by no more than 0.05 dalton, then the two spectra were merged into one MS/MS spectrum by putting the two peak lists into one. Next, a simple criterion is applied to remove the poor quality spectra as follows.

For an MS/MS spectrum, we define the *average signal intensity* as $\frac{s}{m}$, where s is the sum of the abundances of the peaks higher than 2 (peaks lower than 2 cannot be distinguished from noise), and m is the peptide mass (which is equal to the precursor ion mass minus the protons). s is divided by m because peptides with higher masses are generally longer and therefore the larger number fragments give more total signal intensity. Thus, for larger peptides, higher total signal intensity is required for the *de novo* sequencing. Visual inspection revealed that the quality of the spectra with average signal intensity lower than 0.6 is generally very poor. Hence, the spectra whose average signal intensity was lower than 0.6 were removed from the raw data files. Figure 2 shows an example of an excluded spectrum with average signal intensity 0.56, and Figure 3 shows an example of a remaining spectrum (the precursor ion 675.72 in the albumin data set) with average signal intensity of 0.7. As given in Table 1, PEAKS computed a correct partial sequence of nine consecutive amino acids for the MS/MS spectrum in Figure 3.

After this initial sorting, the remaining data contain 54 spectra from tryptic digestions (C-terminal is either R or K) and 4 spectra from non-tryptic digestions (C-terminal is not R or K). The average signal intensities for the four spectra from non-tryptic digestions are 21.5, 4.7, 1.0 and 0.9 respectively. PEAKS got partially correct sequences of length 7, 4, 4, and 2, respectively. This suggests that PEAKS requires high quality spectra for *de novo* sequencing of non-tryptic spectra. Lutefisk, however, did not find any sequences for these four spectra.

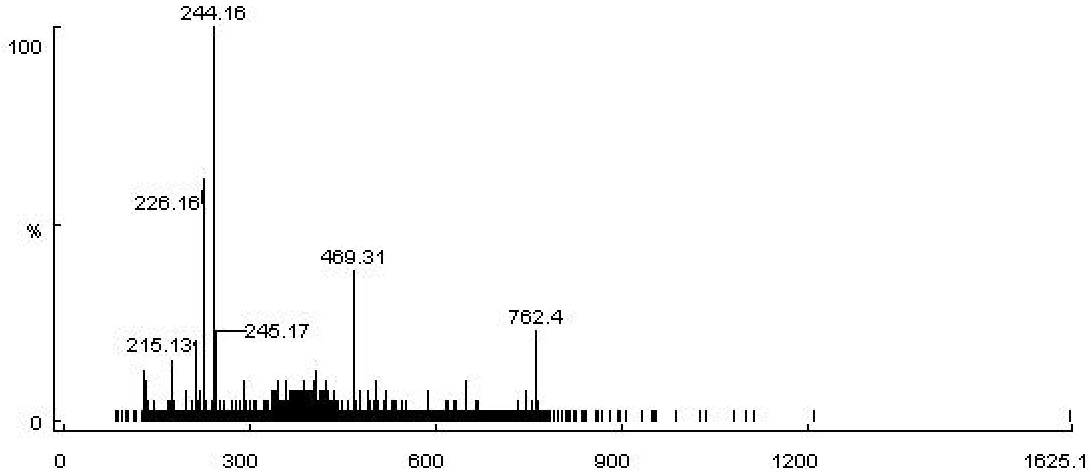

Figure 2. A spectrum of poor quality with average signal intensity of 0.56 not selected for analysis.

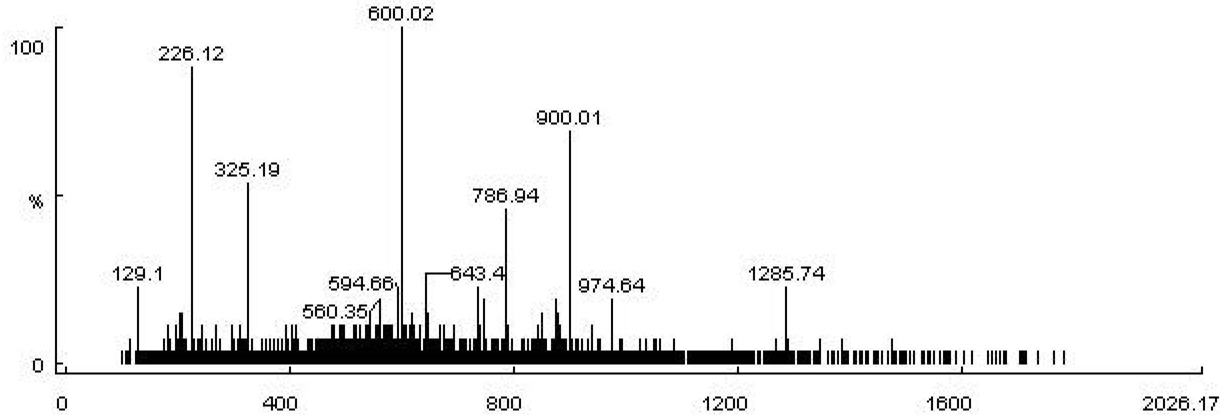

Figure 3. Spectrum from BSA digestion (precursor ion 675.72) of acceptable quality with average signal intensity of 0.7 and selected for analysis.

Both PEAKS and Lutefisk were then used to compute the sequences of the 54 MS/MS spectra *de novo.* Although both software packages output several results for one spectrum, we selected only here the first result (with the highest score) among their outputs. Table 1 summarizes the results obtained by PEAKS and Lutefisk for BSA MS/MS spectra. The underlined amino acids are those correctly computed by PEAKS or Lutefisk (no distinction between the amino acids L with I, and K with Q). The bold amino acids (one letter code) in PEAKS' computation indicate that PEAKS gave confidence scores >= 80% for those amino acids. The computed sequences of the other three proteins are not given here but can be found at http://www.csd.uwo.ca/~bma/peaks/. The 54 MS/MS spectra are also available at the same web site.

m/z	z	correct	PEAKS	Lutefisk	s/m
Albumin					

417.21	3	FKDLGEEHFK	RLCM**GEEHFK**	no quality sequence found	5.1
454.88	3	SLHTLFGDELCK	TV**HTLFGDELCK**	no quality sequence found	4.3
461.72	2	AEFVEVTK	**AEFVEPCK**	[200.08]FVEVTK	61.1
464.24	2	YLYEIAR	**YLYELAR**	YLYELAR	45.8
465.77	2	LKAWSVAR	**LKAWSVAR**	LKAWSVAR	2.1
473.58	3	LKECCDKPLLEK	RTLCCDK**PLLEK**	no quality sequence found	9.9
501.29	2	ALKAWSVAR	LA**KAWSVAR**	[184.12]KAWWAR	2.3
507.79	2	QTALVELLK	GA**TALVELLK**	[229.11]ALVELLK	5.8
515.79	4	YTRKVPQVSTPTLVEVSR	WHYEHF**T**DKN**LVEVSR**	[200.08][244.07][LP][AH]RP[242.14]LVEVSR	2.5
547.26	3	KVPQVSTPTLVEVSR	**KVA**PG**VSTPTLVEVSR**	no quality sequence found	93.2
571.86	2	KQTALVELLK	KQ**TALVELLK**	KQTALVELLK	1.8
582.29	2	LVNELTEFAK	**LVNELTEFAK**	LVNELTEFAK	11.7
642.36	2	HPEYAVSVLLR	**HPEYAVPSDLR**	no quality sequence found	1.1
653.38	2	HLVDEPQNLIK	**HLVDEP**K**NLLK**	HLVDE[225.15]NLLK	7.8
675.72	3	KVPQVSTPTLVEVSRSLGK	KV**N**PLGMHCAVEVSR**SLGK**	no quality sequence found	0.7
681.84	2	SLHTLFGDELCK	**SLHTLFGDELCK**	[HT]VTL[GV]YE[216.07]K	2.5
693.80	2	YICDNQDTISSK	**YLCDNQDTLSSK**	YL[218.07]NQDTLSSK	22.1
740.39	2	LGEYGFQNALIVR	**LGEYGFQNALLVR**	LWYGFQNALLVR	17.9
756.42	2	VPQVSTPTLVEVSR	**VPQVSTP**NAKE**VSR**	no quality sequence found	1.3
767.70	3	NYQEAKDAFLGSFLYEYSR	QH**SS**FVHT**AQGGSFLYEYSR**	[276.11]GK[SS][MT][199.10]LGSFLYEYSR	2.1
784.34	2	DAFLGSFLYEYSR	**WFLGSFL**ATAAGGN**R**	[186.07]FLGSFLYEYSR	15.9
820.45	2	KVPQVSTPTLVEVSR	K**V**PQ**VST**MAHA**EVSR**	no quality sequence found	2.7
824.74	3	QNCDQFEKLGEYGFQNALIVR	Q**LSE**MF**E**KLW**YGFQNALLVR**	no quality sequence found	0.9

Table 1. The performance of PEAKS and Lutefisk on Albumin (bovine) MS/MS data set. The spectrum quality column s/m shows the average signal intensity of each spectrum.

For the 54 MS/MS spectra, Table 2 gives the numbers of sequences that PEAKS and Lutefisk computed completely correct or partially correct (with at least 6 consecutively correct amino acids). It can be seen that PEAKS performs better than Lutefisk on these 54 spectra. It is important to note that for the 27 spectra of lower quality (s/m between 0.6 and 10), PEAKS computed three times as many completely or partially correct sequences as Lutefisk.

Spectrum quality (s/m)	Total number of spectra	Completely correct sequences		Sequences with 6 consecutively correct amino acids	
		PEAKS	Lutefisk	PEAKS	Lutefisk
> 10	27	13 (48%)	8 (30%)	25 (93%)	18 (67%)
0.6 – 10	27	9 (33%)	3 (11%)	26 (96%)	9 (33%)
Overall	54	22 (41%)	11 (20%)	51 (94%)	27 (50%)

Table 2. The number of completely or partially correct sequences computed by PEAKS and Lutefisk.

Table 3 gives the total number of correct amino acids that PEAKS and Lutefisk computed. From this table it is also evident that PEAKS performs better than Lutefisk. For spectra with lower quality (0.6 – 10), PEAKS computed more than twice as many correct amino acids than Lutefisk.

Spectrum quality (s/m)	Total number of amino acids	PEAKS	Lutefisk
>10	307	262 (85%)	185 (60%)
0.6 – 10	341	261 (77%)	122 (36%)
Overall	648	523 (81%)	307 (47%)

Table 3. The number of correct amino acids computed by PEAKS and Lutefisk.

PEAKS gives a positional confidence score to individual amino acids that it computes. The amino acids that PEAKS gave high confidence are usually the correct amino acids, but PEAKS occasionally makes mistakes. It is also possible that PEAKS computes some correct amino acids but gave low confidence. For the 54 spectra, Figure 4 illustrates the relationship between the amino acids that PEAKS gave a high confidence score (=80%), and the amino acids that PEAKS computed correctly. The figure illustrates that PEAKS' positional confidence scoring is fairly reliable: 92% (= 484/(41+484)) of the amino acids that were given high (=80%) confidence are correct, and 93% (= 484/(39+484)) of the amino acids that were computed correctly have high (=80%) confidence.

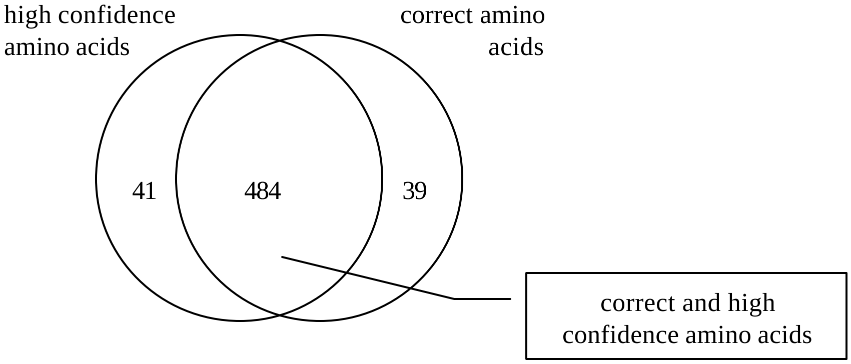
high confidence amino acids
correct amino acids
41
484
39
correct and high confidence amino acids

Figure 4. The relationship between the amino acids that PEAKS gave a high confidence score (=80%), and the amino acids that PEAKS computed correctly.

Both PEAKS and Lutefisk can compute the MS/MS data rapidly. On average, they process each MS/MS spectrum in a few seconds on a Pentium 1GHz PC. PEAKS (including its interface) requires 512M bytes of memory, common to most desktop computers currently available. We do not know Lutefisk's memory requirement but it can be run with no problem on a PC with 512M bytes of memory.

Finally, we want to point out that all of the wrongly assigned amino acids by PEAKS are caused by mass equivalence. Some examples in Table 2 are: mass (SL) = mass(TV) in precursor 454.88, mass(VT)=mass(PC) in precursor 461.72, mass(AL)=mass(LA) in precursor 501.29, and mass(Q)=mass(GA) in precursor 507.79. If the correct sequence is in a database and the computed sequence is partially correct, this type of error can usually be overcome by a careful database search with the sequences. For example, a software system, SPIDER[17], can be fed with sequences containing *de novo* sequencing errors but find the correct sequences in the database.

5. Conclusion

From this initial evaluation, we can see that PEAKS performs very well for *de novo* sequencing of Q-TOF spectra compared with Lutefisk. PEAKS also performed better than other *de novo* sequencing software from manufacturers of mass spectrometers (data not shown). Not only PEAKS computes more correct sequences and amino acids than the other software, but also it outputs positional confidence scores, which reliably determine which sequences or amino acids are correct. Although not discussed in this paper, PEAKS has already been used to successfully compute the peptides with some common post-translational modifications. Future version will include the ability to compute a wider range of more complex modifications. PEAKS should be a very useful tool for the analysis of proteome of both known and unknown genomes.

Acknowledgments. This work was supported by NSERC research grants to BM and KZ, and by the Ontario Research and Development Challenge Fund (ORDCF) to GL and by Bioinformatics Solutions INC (Waterloo , ON). We also thank Dr. R. Johnson for kindly providing the Lutefisk software.

References

1. Perkins DN, Pappin DJC, Creasy DM, and Cottrell JS. *Electrophoresis* 1999; **20**: 3551-3567.

2. Eng JK, McCormack AL, Yates JR III. *J. Amer. Soc. Mass Spectrom.* 1994; **5**: 976-989.

3. Mann M, Wilm M. *Anal. Chem*. 1994; **66**: 4390-4399.

4. Keough T, Lacey MP, Youngquist RS. *Rapid Comm. Mass Spectrom.* 2000; **14**: 2348-2356.

5. Munchbach M, Quadroni M, Miotto G, James P. *Anal Chem.* 2000; **72**: 4047-4057.

6. Uttenweiler-Joseph S, Neubauer G, Christoforidis S, Zerial M, Wilm M. *Proteomics* 2001; **1**: 668-682.

7. Bartels, C. *Biomed. Environ. Mass Spectrom.* 1990; **19**: 363-368.

8. Chen T, Kao M, Tepel M, Rush J, Church G. *J. Computational Biology* 2001; **8**: 325-337.

9. Danc?k V, Addona T, Clauser K, Vath J, Pevzner P. *J. Computational Biology* 1999; **6**: 327-341.

10. Taylor JA, Johnson RS. *Rapid Commun. Mass Spectrom.* 1997; **11**: 1067-1075.

11. Taylor JA, Johnson RS. *Anal. Chem*. 2001; **73**: 2594 - 2604.

12. Fernández de Cossío J, Gonzales J, Besada V. *CABIOS* 1995; **1**: 427-434.

13. Hamm CW, Wilson WE, Harvan DJ. *CABIOS* 1986; **2**: 115-118.

14. Hines WM, Falick AM, Burlingame AL, Gibson BW. *J. Am. Soc. Mass. Spectrom.* 1992; **3**: 326-336.

15. Sakurai T, Matsuo T, Matsuda H, Katakuse I. *Biomed. Mass Spectrum* 1984; **11**: 396-399.

16. Ma B, Zhang K, Liang C. *Symp. Comb. Pattern Matching* 2003; 266-278.

17. Han Y, Ma B, Zhang K. *Unpublished*; available at http://proteome.sharcnet.ca:8080/spider/.