

## PEAKS: A Powerful Software Tool for the *De Novo* Sequencing of Peptides from MS/MS Data

Bin Ma<sup>1</sup>, Kaizhong Zhang<sup>1</sup>, Gilles Lajoie<sup>1</sup>, Amanda Doherty-Kirby<sup>1</sup>, Chengzhi Liang<sup>2</sup>, Ming Li<sup>3</sup>

<sup>1</sup> University of Western Ontario, London, ON, Canada

<sup>2</sup> Bioinformatics Solutions Inc., Waterloo, ON, Canada

<sup>3</sup> University of California at Santa Barbara, Santa Barbara, CA

### 1. Introduction

We developed a software tool, PEAKS, to automatically derive the amino acid sequence from MS/MS experiments of peptides. PEAKS accepts a peak list for the MS/MS spectrum, and tries to “explain” all the peaks in the spectrum by the ions produced with every possible peptide. The peptides that explain the peak list the best will be output as the predicted peptide sequences.

PEAKS' approach is different from the database search methods performed by software systems like Mascot (Perkins *et al.* 1999). Instead of searching in a protein database, PEAKS searches all the possible amino acid strings. Moreover, to speed up the search, PEAKS uses a sophisticated dynamic programming algorithm (Ma *et al.* 2002), which is different than the often used “graph theory” approach (Bartels 1990, Taylor *et al.* 1997, Dancik *et al.* 1999, Chen *et al.* 2001). The dynamic programming algorithm ensures that the best peptides are not ignored. Typically, PEAKS predicts the amino acid sequence of a tryptic peptide in ten seconds with a moderate desktop computer. The accuracy of PEAKS' prediction is superior to most other currently available *de novo* sequencing software tools.

### 2. Methods

PEAKS processes the spectrum with the following three steps:

**(1) Preprocess of the spectrum:** The preprocess includes the noise filtering, deconvolution and translating the doubly/triply charged ions to singly charged ions. For the noise filtering, we estimate the noise level  $h_0$  of the spectrum and reduce the intensities of all the peaks in the spectrum by  $h_0$ , and then remove all the peaks with zero or negative intensities. For the deconvolution, we detect the windows that contain a contiguous sequence of high intensity peaks. Then for each window, we sum up all intensities of the peaks to the “center” of the window, and remove all the peaks in the window except for the “center”. For the doubly/triply charged ions, we use the isotopic ions to determine the charge of the ions and translate them to singly charged ions. Then we add the intensities of all isotopic ion peaks to the corresponding monoisotopic ion peaks and remove the isotopic ion peaks.

**(2) Peptide candidates computation:** We define the matching score between a peptide and a spectrum to be the sum of the “scores” of the peaks that are matched by the a-, b-, c-, x-, y-, z-ions, and the y-, b-ion losing a water or ammonium. The score of a peak matched by an ion is the logarithmic intensity of the peak, multiplied by a factor related to the error between the masses of the ion and the peak. If a peak is matched by two different ions, only the match with higher score is counted. A sophisticated dynamic programming algorithm is used to compute the peptides that match a given spectrum with the highest scores. The best 10000 sequences are computed as candidates.

**(3) Sort of the candidates:** The candidates computed in the second step are further evaluated by a more accurate score computation. In this step, the ammonium ions and the internal cleavage ions are taken into account. The ion mass tolerance is also reduced from the tolerance used in the second step. In addition, the spectrum is recalibrated to incorporate the spectrometer's error caused by temperature change. The factor of the recalibration is computed by fitting the m/z values of the y-ions and the matched peaks, using the Least Square method. This recalibration method is very similar to the one conducted by Taylor *et al.* 2001. After the computation of the new scores of all the candidates, the first  $n$  sequences with the highest scores are output, where  $n$  is specified by the users.

### 3. Experimental results

We applied PEAKS to 20 MS/MS data sets. Some of the data were done in our lab using MicroMass's instruments, and some of the data were collected from other research labs using MicroMass or MDS-Sciex instruments. As a comparison to the existing software, we also applied Lutefisk (Taylor *et al.* 1997,

2001) to the 20 data sets. The results are shown in the following table. Without distinguishing L with I, and K with Q, the underlined (partial) sequences are the amino acids that were correctly predicted by PEAKS or Lutfisk.

<i>real sequences</i>	<i>PEAKS' predictions</i>	<i>Lutfisk's predictions</i>
1. LVNELTEFAK	<u>LVNELTEFAK</u>	<u>LVNELTEFAK</u>
2. HPEYAVSVLLR	<u>HPEYAVSVLLR</u>	<u>HPEYAVSVLLR</u>
3. HLVDEPQNLLK	<u>HLVDEPQNLLK</u>	HLVDE[225.2]FJK
4. MDPHENILLSTLEIK	FVALATCMLLVLPPHK	DMV[226.1][265.2]LLSTLELK
5. ITIPDLTDVNAIDR	TLLPDLTDVNALDR	[214.1]LPDLTDVNALDR
6. STMLAYDASSIQYR	MTSLAYDASSLQYR	[319.1]LAYDASSLKYR
7. YTEYNEPHESR	<u>YTEYNEPHESR</u>	<u>YTEYNEPHESR</u>
8. TNSDLVETLR	<u>TNSDLVETLR</u>	[215.0]SDLVETLR
9. EGVNDNEEGFFSAR	ADVNDNEEGFFSAR	WVNDNEEGFFSAR
10. LVQEVTDFAK	<u>LVKEVTDFAK</u>	[212.1]KEVTDFAK
11. LGEYGFQNAILVR	<u>LGEYGFQNALVR</u>	[170.1]EYGFKNALLVR
12. FPGQLNADLR	FPGQLNADLR	FPGKLNWA[198.1]
13. LGSSEVEQVQLVVDGVK	YYFVEKQVQLVVDGVK	[170.1]SSEVEKVKLVVDGVK
14. LSSPATLNSR	<u>LSSPATLNSR</u>	LSSPATLNDK
15. EALDFFAR	<u>EALDFFAR</u>	[200.0]LDFFAR
16. VLGIDGGEGKEELFR	<u>VLGLDGGEGKEELFR</u>	<u>VLGLDGGEGKEELFR</u>
17. LVPTYESASLR	<u>LVPTYESASLR</u>	<u>LVPTYESASLR</u>
18. DLYANTVLSGGTTMYPGIADR	DFTKWVLSGGTTMYAGDLPR	no output
19. QLFHPEQLITGK	<u>QLFHPEKLLTGK</u>	[304.0]LLK[226.1]H[185.0]GK
20. TTGIVMDSGDGVTHTVPIYEG YALPHAILR	<u>WWVMDSGDRTHTVPLYEGYAL</u> <u>PHVTPR</u>	[371.2]VM[287.1][220.1][293.2][210.2]CETAGYALPHVLG[184.1]

Table 1: The performance of PEAKS and Lutfisk on 20 data sets.

From the comparison with Lutfisk, we find that PEAKS has a better overall performance for these 20 data sets. We observed that in the fourth data set, the y9 to y14 ions are absent. On the other hand, the spectrum contains many high intensity peaks that are coincidentally identified as another series of y1 to y8 ions by PEAKS. For these reasons, PEAKS gave the wrong sequence for the fourth data set. There are many doubly charged ion peaks in data sets 18-20 and this may be the reason that Lutfisk had a poor performance for these three data sets.

#### 4. Discussion

Although our test on PEAKS is not thorough, we can see that PEAKS performs very well for *de novo* sequencing. The algorithm used by PEAKS to compute the candidates is essentially different from the database searching method and the graph theory method used in existing systems. Our recent research has extended PEAKS' algorithm so it can search for the best matching peptides from a protein database as well. The database search feature will be added to PEAKS in the near future.

#### References

- Bartels, C. 1990. *Biomed. Environ. Mass Spectrom* 19, 363-368.
- Chen, T., Kao, M., Tepel, M., Rush J., and Church, G. 2001. *J. Computational Biology* 8(3), 325-337.
- Dančik, V., Addona, T., Clauser, K., Vath, J., and Pevzner, P. 1999. *J. Computational Biology* 6, 327-341.
- Ma, B., Zhang, K., and Liang, C. 2002. *submitted*.
- Perkins, D.N., Pappin, D.J.C., Creasy, D.M., and Cottrell, J.S. 1999. *Electrophoresis* 20, 3551-3567.
- Taylor, J.A., and Johnson, R.S. 1997. *Rapid Commun. Mass Spectrom.* 11, 1067-1075.
- Taylor, J.A., and Johnson, R.S. 2001. *Anal. Chem.* 73, 2594 - 2604.