BioTechnologia vol. 92(3) • pp. 296-300 • 2011
Journal of Biotechnology, Computational Biology and Bionanotechnology RESEARCH PAPER

Optimization algorithm for *de novo* analysis of tandem mass spectrometry data

MICHAŁ KISTOWSKI [1], ANNA GAMBIN [2,3]*

[1] Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Warszawa, Poland

[2] Institute of Informatics, University of Warsaw, Warszawa, Poland

[3] Mossakowski Medical Research Centre, Polish Academy of Sciences, Warszawa, Poland

* Corresponding author: aniag@mimuw.edu.pl

Abstract

Protein identification is usually achieved by tandem mass spectrometry (MS/MS). Because of the difficulty in measuring complete proteins using MS/MS, typically a protein is enzymatically digested into peptides and the MS/MS spectrum of each peptide is measured. The database searching methods are predominant in the task of peptide identification. Their aim is to find the best match between model spectra generated from the peptides stored in the database and the experimental mass spectrum obtained for an unidentified peptide. In this approach one assumes that the peptide under investigation belongs to the scanned database. Otherwise, so called *de novo* methods have to be applied to determine the peptide sequence. Unfortunately, *de novo* sequencing algorithms are fragile in the presence of missing peaks, background noise or post-translational protein modifications. In this paper, we propose a post-processing method for optimizing the results obtained from *de novo* sequencing algorithms. Our approach in the reconstruction of amino acid sequences employs only spectral features and is robust with respect to missing data. We demonstrate the significant improvement achieved using our method applied to sequences reconstructed using a popular *de novo* sequencing method. The tool is freely available at http://pepygen. sourceforge.net.

Key words: tandem mass spectrometry, *de novo* sequencing, genetic algorithm

Introduction

Rapid development of mass spectrometric (MS) technology offers the possibility of performing exhaustive analysis of complex mixtures containing thousands of peptides in a single experiment. Identification of those peptide signals is a prerequisite for several high-throughput proteomics technologies. In a typical experiment, protein mixtures are digested with trypsin, separated using liquid chromatography and analyzed by tandem mass spectrometry (LC-MS/MS).

In this paper, we focus on the last stage of the processing, i.e. identification of peptides from the knowledge of mass to charge ratios of the fragment ions (MS/MS spectra of peptides). The most recognized peptide identification methods can be classified into two main categories: database searching and *de novo* sequencing (Liu et al., 2007). They can be combined to obtain two other approaches: sequence tagging and consensus of multiple engines (Xu and Ma, 2006b; Kall et al., 2007; Brosch et al., 2009).

Database searching methods compare the input spectrum with theoretical spectra of peptides from virtually digested proteins. Different criteria are used to determine the likelihood that the identified peptide is actually the one seen in the spectrum. The most popular scoring function is based on the number and the intensity of peaks matched with the theoretically computed mass to charge ratios of fragment ions (Perkins et al., 1999; Yates et al., 1995). *De novo* approaches do not refer to the database of previously available peptide sequences. They deduce the sequence of amino acids that gave rise to the investigated spectrum with the use of graph-theoretic frame-work, Bayesian nets or hidden Markov models (Frank and Pevzner, 2005; Fischer et al., 2005; Johnson and Taylor, 2002). *De novo* sequencing often fails to reconstruct the entire amino acid sequence, however even partial information can be exploited in a sequence tagging approach (Tabb et al., 2003).

Recently, it has been demonstrated that taking into account the consensus of multiple peptide sequencing

engines can result in a significant improvement in peptide identification accuracy, coverage and confidence. (Xu and Ma, 2006b).

Methods

De novo reconstruction of peptide sequence is recognized as a highly non-trivial task (Xu and Ma, 2006a). In the proposed approach, the sequence reconstructed by *de novo* method is iteratively adjusted through the application of a genetic algorithm, called PepyGen.

Table 1. The examples of mutation and crossing-over operations

Operation	Input	Output
Mutation	EDELAYKK	EDLLAYKK
Crossing-over	EDLLAYQL KMYALLPE	EDLLAYPE KMYALLQL

The genetic algorithms are used to solve various optimization problems (Mitchell, 1996). Firstly, they work by generating a random multiset (initial population) of potential solutions. Secondly, by evaluating the function being optimized (fitness function) for each of them, culling a percentage of them with low values of such function, cloning and slightly altering (mutating) the rest at random and repeating this process until a satisfactory solution is obtained. In our case, the potential solutions are amino acid sequences, while for a given sequence s of length l and MS/MS spectrum t the fitness function is the following:

$$fitness(s) = \frac{S_{match}}{S_{tot}} - \alpha \frac{|m_{calc} - m_{\exp}|}{m_{\exp}} - \beta \frac{n_{miss}}{l}$$

where S_{match} is the sum of intensities of those peaks in spectrum t that match theoretical spectrum of s, S_{tot} corresponds to the sum of intensities of all peaks in t; m_{calc} the theoretical mass of s; m_{exp} is the ion precursor mass in t and n_{miss} the number of b-ions and y-ions in a theoretical spectrum of s missing corresponding peaks in t. The coefficients α and β were empirically tuned to alter the relative importance of different terms in the fitness function (the presented results were obtained for $\alpha = 1$ and $\beta = 0.02$).

We applied two operations for altering the initial population of sequences: *mutation* corresponding to amino acid substitution and *crossing-over*, when two sequences $a_1 \ldots a_n$ and $b_1 \ldots b_m$ are broken at a random point, say p giving two new sequences $a_1 \ldots a_p b_{p+1} \ldots b_m$ and $b_1 \ldots b_p a_{p+1} \ldots a_n$. After the cloning and altering phase, we eliminated all sequences that were shorter than three amino acids and those with mass exceeding the experimentally measured precursor mass more than some predefined value. Algorithm 1 explains the details. Example of mutation and crossing-over operations are provided in Table 1.

Algorithm 1: PepyGen

Input: Fragmentation spectrum
Output: Sequence of peptide
begin
  $\mathcal{P} \leftarrow$ a multiset of peptide sequences from de-novo method (the initial population);
  while *Not Run Out of Time* **do**
    foreach *sequence* $s \in \mathcal{P}$ **do**
      calculate fragment ion masses $\mathcal{F} \leftarrow fitness(s)$;
      Randomly, based on $\mathcal{F}$, choose one of the following:
      - $\mathcal{P} \leftarrow \mathcal{P} \setminus \{s\}$ – with increasing probability for low values of $\mathcal{F}$
      - $\mathcal{P} \leftarrow \mathcal{P} \setminus \{s\} \cup \{\text{Mutate}(s)\}$;
      - $\mathcal{P} \leftarrow \mathcal{P} \setminus \{s\} \cup \{\text{Mutate}(s)\}$ – with increasing probability for high values of $\mathcal{F}$
    end
  end
  return *the member of* $\mathcal{P}$ *that maximizes* $\mathcal{F}$
end

Results and discussion

We illustrate the performance of PepyGen method on two small examples from the literature (see Table 2). A large set of real MS/MS spectra which is the outcome of post-processing is also presented.

To assess the quality of the results obtained by PepyGen, we used Lutefisk (Johnson and Taylor, 2002) *de novo* sequencing algorithms. It is a popular, easily accessible and user friendly representative example of *de novo* methods.

Table 2. Spectra uses in the examples along with the mass of the parent ion

Example	Source	Peptide	Parent mass	Fragment ion massess
1	Kinter et al. (Kinter and Sherman, 2000)	LFSQVGK	779	114.16, 147.18, 204.23, 261.34, 303.36, 348.42, 431.49, 476.55, 518.57, 575.68, 632.73, 665.75
2	Snyder (Snyder, 2000)	EDLIAYLK	965	147.2, 245.2, 260.3, 358.3, 423.5, 471.5, 494.6, 542.6, 607.1, 705.8, 720.9, 818.9, 965.1

Example 1. In this example, we study the sequence EDLIAYLK (*equine cytochrome c fragment*). The ESI-tandem mass spectrum for this peptide was presented in (Snyder, 2000) (c.f. Table 2). We used the reported m/z values as the input for Lutefisk to identify the correct sequence, except the suffix part of it (see Table 3).

Table 3. The output of Lutefisk method for the spectrum of EDLIAYLK peptide

1	[244.0]LLAYLK
2	[243.1]NLAYLK
3	[259.1]YALNPK
4	[259.1]YALL[226.1]
5	[259.1]YA[211.1]LK
6	[259.1][211.1]AYLK
7	[244.0]LL[250.2]PK
8	[243.1][250.2]LNPK
9	[243.1]NL[250.2]PK
10	[259.1][211.1][250.2]PK

Table 4. The output of Lutefisk method followed by PepyGen post-processing for the spectrum of EDLIAYLK peptide. The improvement is clearly visible: while the Lutefisk reconstructed only a portion of the peptide, our method yielded almost perfect sequence EDLLAYLK (in bold). Amino acids I and L cannot be distinguished within the tolerance vale assumed

1	EDLLAYQL	1	KMYALLPE	1	EDLLAYQL
2	EDLLAYALG	2	LMLLPHLK	2	EDLLAYALG
3	EDLLAYAAV	3	LMLLHPLK	3	KMYALLPE
4	EDLLAYKL	4	DELLAYLQ	4	EDLLAYKL
5	LMLLPHLK	5	QMYALLPE	5	LMLLPHLK
6	EDLLPHLK	6	QMYAPHLM	6	EDDLLAYALG
7	EDLLLAYKL	7	KMYALLAG	7	EDLLAYAAG
8	EDELAYKK	8	LMYALLPE	8	LMLLPHL
9	EDLLAYQL	9	DELLAYLQ	9	EDLLHPLK
10	EDLLAYAAV	10	KMMALLLQ	10	**EDLLAYLK**

The outcomes of Lutefisk algorithm have been processed by PepyGen method that yielded the results presented in Table 4. Among high scored sequences from the three independent runs of PepyGen (500 iterations each), the correct sequence was identified once. We are aware of the fact that the algorithm could achieve better results. However, several other sequences match both the prefixes and suffixes of the target sequence. One can exploit the repeatability of reconstructed prefixes and suffixes for the prediction of a correct sequence (even if it is not known in advance). Also the use of additional non-mass based information should significantly improve the performance of the fitness function.

Example 2. The target of our second example is the spectrum of the sequence LFSQVGK. The objective was to check the robustness of our method to missing peaks. Using Lutefisk algorithm for the dataset missing one or two peaks, we obtained the results that are summarized in Table 5. In a majority of cases, Lutefisk has not been able to reconstruct the correct peptides at all; only in few cases, the outcome of the method matched the portion of target sequence. The output of Lutefisk method was used as the initial population for PepyGen algorithm and after a few seconds (i.e. 500 iterations of evolutionary schema) the correct peptide chain was identified in almost all cases, as presented in Tables 6 and 7.

Example 3. 1038 peptides were derived from *Escherichia coli* and *Schizosaccharomyces pombe* by tryptic digestion, and were analyzed in the Mass Spectrometry Lab at Institute of Biochemistry and Biophysics PAS (Warsaw, Poland) on ESI-FTICR mass spectrometer. The maximal length of each peptide was 15 amino acids and the score from Mascot was at least 50 (i.e. significant with P-value < 0.05). Therefore, we treated the sequences from Mascot as correct ones and calculated the mean length of the longest common subword (lcsw) between them and the peptides sequenced *de novo*. We ran Lutefisk on this dataset. The resulted lcsw equaled 2.962. After post-processing with PepyGen method, lcsw increased to 3.159. The improvement might seem insignificant but what is important is that in many cases (ca. 20%) PepyGen post-processing enabled reconstruction of a correct peptide sequence.

Table 5. The outcome of Lutefisk algorithm when one or two peaks were deleted from the spectrum of the sequence LFSQVGK. This situation, commonly found in practice, clearly constitutes a hurdle for *de novo* approach: the algorithm fails to reconstruct the correct sequence in all cases. However, in some cases, Lutefisk method reported the fragment of the target sequence, mostly the suffix of it

Ion missing	114.16	147.18	204.23	261.34	303.36	348.42	431.49	476.55	518.57	575.68	632.73	665.75
114.16	FGGSKVGK											
147.18	[203.1]VKSGGE	[203.1]VKSGGE										
204.23	FGGSKRK	N[188.0]KSGGE	FGGSKRK									
261.34	FGGSKVGK	[203.1]VKSGGE	FGGSKRK	FGGSKVGK								
303.36	FGVKSGGK	FLSKVGE	FGVKSNK	FGVK[144.0]GK	FGVKSGGK							
348.42	FGGSKVGK	[203.1]VKSGGE	FGGSKRK	FGGSKVGK	FGV[214.1]DK	FGGSKVG						
431.49	FGVKSGGK	FLSKVGE	FGVKSNK	FGVK[144.0]GK	FGVKSGGK	FGV[214.1]DK	FGVKSGGK					
476.55	FGGSKVGK	[203.1]VKSGGE	FGGSKRK	FGGSKVGK	DF[170.1]SGGK	FGGSKVGK	FLSEVGK	FGGSKVGK				
518.57	FGVKSGGK	FLSKVGE	FGVKSNK	FGVK[144.0]GK	FGVKSGGK	FL[314.2]GE	FGVKSGGK	FLSEVGK	FGVKSGGK			
575.68	FNSKVGK	[203.1]VKSNE	FLSKRE	FNSKVGK	FRKSGGK	FNSKVGK	FRKSGGK	FNSKVGK	FRKSGGK	FNSKVGK		
632.73	[204.1]VKSGGK	[204.1]VKSGGK	[204.1]VKSNK	DFSKVGK	[204.1]VKSGGK	DFSKVGK	[204.1]VKSGGK	DFSKVGK	[204.1]VKSGGK	DFSKVGK	[204.1]VKSGGK	
665.75	FGGSKVGK	[203.1]VKSGGE	FGGSKRK	FGGSKVGK	FGVKSGGK	FGGSKVGK	FGVKSGGK	FGGSKVGK	FGVKSGGK	FNSKVGK	[204.1]VKSGGK	FGGSKVGK

Table 6. The results of PepyGen post-processing of data from Table 5.
After 500 iterations (done in a few seconds) the incorrect peptide sequences evolved into the correct one in 94% of cases

Ion missing	114.16	147.18	204.23	261.34	303.36	348.42	431.49	476.55	518.57	575.68	632.73	665.75
1	LFSKVGK											
2	LFSKVGK	LFSKVGK										
3	LFSKVGK	LFSKVGK	LFSKVGK									
4	PYSKVGK	LFSKVGK	LFSKVGK	LFSKVGK								
5	FLSKVGK	LFSKVGK	LFSKVGK	LFSKVGK	LFSKVGK							
6	LFSKVGK	LFSKVGK	LFSKVGK	LFSKVGK	LFSKVGK	LFSKVGK						
7	LFSKVGK	LFSKVGK	LFSKVGK	LFSKVGK	LFSKVGK	LFKSVGK	LFSKVGK					
8	LFSKVGK	LFSKVGK	LFSKVGQ	LFSKVGK	LFSKVGK	LFSKVGK	LFSKVGK	LFSKVGK				
9	FLSKVGK	LFSKVGK	LFSKVGK	LHPKVGK	LFSKVGK	LFSKVGK	LFSKVGK	LFSKVGK	LFSKVGK			
10	LFSKVGK	LFSKVGK	LFSKRK	LFSKVGK	LFSKVGK	LFSKVGK	LFSKVGK	LFSKVGK	LFSKVGK	LFSKVGK		
11	LFSKVGK	LFSKVKG	LFSKVGK	LFSKVGK	LFSNLGK	LFSKVGK	LFKSVGK	LFSKVGK	MGNKSGGK	LFSKVGK	LFSKVGK	
12	FLSKVGK	LFSKVGK	LFSKVGK	LFSKVGK	LFSKVGK	LFSKVGK	MGNKSGGK	LFSKVGK	MGNKSGGK	LFSKVGK	LFSKVGK	LFSKVGK

Table 7. The results of PepyGen post-processing of data from Table 5. Table gives the number of correct sequences obtained in 10 independently started runs of PepyGen. Only in 5 cases out of 78 our tool failed to find the correct sequence, while in most cases all 10 runs yielded the correct answer. The amino acid K and Q cannot be distinguished from each other, given the tolerance value used

	1	2	3	4	5	6	7	8	9	10	11	12
1: 114.16	**9**											
2: 147.18	**10**	**10**										
3: 204.23	7	4	**10**									
4: 261.34	**10**	**10**	**10**	**10**								
5: 261.34	**10**	**10**	**9**	**10**	7							
6: 303.36	**10**	**10**	**10**	**10**	**10**	**10**						
7: 348.42	8	**10**	**10**	**10**	**10**	0	6					
8: 431.49	**10**	**10**	**10**	**10**	3	**10**	**10**	**10**				
9: 476.55	6	**10**	**10**	0	**9**	**10**	6	6	8			
10: 518.57	**10**	**10**	0	**10**	**9**	**10**	**10**	**10**	**9**	**10**		
11: 575.68	**10**	0	6	**10**	4	**9**	1	**10**	1	**10**	**10**	
12: 632.73	0	**10**	**10**	**10**	7	**10**	7	**10**	6	**10**	**10**	**10**

Further research

The most valuable further extension of our method would be incorporation of the non-mass based information acquired routinely in liquid chromatography tandem mass spectrometry analyses, like peptide chromatographic retention time.

Acknowledgments

This work was supported by the Polish Ministry of Education and Science grant PBZ-MNiI-2/1/2005.

References

Brosch M., Yu L., Hubbard T., Choudhary J. (2009) *Accurate and sensitive peptide identifica- tion with mascot percolator*. J. Proteome Res. 8: 3176-3181.

Fischer B., Roth V., Roos F., Grossmann J., Baginsky S., Widmayer P., Gruissem W., Buhmann J.M. (2005) *NovoHMM: a hidden markov model for de novo peptide sequencing*. Anal. Chem. 77: 7265-7273.

Frank A., Pevzner P. (2005) *PepNovo: de novo peptide sequencing via probabilistic network modeling*. Anal. Chem. 77: 964-973.

Johnson R.S., Taylor J.A. (2002) *Searching sequence databases via de novo peptide sequencing by tandem mass spectrometry*. Mol. Biotech. 22: 301-315.

Kall L., Canterbury J.D., Weston J., Noble W.S., MacCoss M.J. (2007) *Semi-supervised learning for peptide identification from shotgun proteomics datasets*. Nature Meth. 4: 923-925.

Kinter M., Sherman N.E. (2000) *Protein Sequencing and Identification Using Tandem Mass Spectrometry*. Wiley-Interscience.

Liu J., Bell A., Bergeron J., Yanofsky C., Carrillo B., Beaudrie C., Kearney R. (2007) *Methods for peptide identification by spectral comparison*. Proteome Sci. 5: 3.

Mitchell M. (1996) *An Introduction to Genetic Algorithms*. MIT Press.

Perkins D.N., Pappin D.J.C., Creasy D.M., Cottrell J.S. (1999) *Probability-based protein iden- tification by searching sequence databases using mass spectrometry data*. Electrophoresis 20: 3551-3567.

Snyder A.P. (2000) *Interpreting protein mass spectra: a comprehensive resource*. An American Chemical Society Publication.

Tabb D.L., Saraf A., Yates J.R. (2003) *GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model*. Anal. Chem. 75: 6415-6421.

Xu C., Ma B. (2006a) *Complexity and scoring function of MS/MS peptide de novo sequenc- ing*. Computational Systems Bioinformatics / Life Sciences Society. Computational Systems Bioinformatics Conference p. 361-369.

Xu C., Ma B. (2006b) *Software for computational peptide identification from MS-MS data*. Drug Discov. Today 11: 595-600.

Yates J.R., Eng J.K., McCormack A.L., Schieltz D. (1995) *Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database*. Anal. Chem. 67: 1426-1436.