17th European Symposium on Computer Aided Process Engineering – ESCAPE17
V. Plesu and P.S. Agachi (Editors)



# De Novo Peptide Identification Via Mixed-Integer Linear Optimization And Tandem Mass Spectrometry

Peter A. DiMaggio Jr. and Christodoulos A. Floudas[a]

[a]*Department of Chemical Engineering, Princeton University, Princeton, NJ 08544, USA, floudas@titan.princeton.edu*

## Abstract

A novel methodology for the de novo identification of peptides via mixed-integer linear optimization (MILP) and tandem mass spectrometry is presented. The overall mathematical model is presented and the key concepts of the proposed approach are described. A pre-processing algorithm is utilized to identify important m/z values in the tandem mass spectrum. Missing peaks, due to residue-dependent fragmentation characteristics, are dealt with using a two-stage algorithmic framework. A cross-correlation approach is used to resolve missing amino acid assignments and to select the most probable peptide by comparing the theoretical spectra of the candidate sequences that were generated from the MILP sequencing stages with the experimental tandem mass spectrum. The novel proposed de novo method, denoted as PILOT, is compared to existing popular methods such as Lutefisk, PEAKS, PepNovo, EigenMS and NovoHMM for a set of spectra resulting from QTOF instruments.

**Keywords:** mixed-integer linear optimization (MILP), de novo peptide identification, tandem mass spectrometry (MS/MS)

## 1. Introduction

Of fundamental importance in proteomics is the problem of peptide and protein identification. Over the past couple decades, tandem mass spectrometry

(MS/MS) coupled with high performance liquid chromatography (HPLC) has emerged as a powerful experimental technique for the effective identification of peptides and proteins. In recognition of the extensive amount of sequence information embedded in a single mass spectrum, tandem MS has served as an impetus for the recent development of numerous computational approaches formulated to sequence peptides robustly and efficiently with particular emphasis on the integration of these algorithms into a high throughput computational framework for proteomics. The two most frequently reported computational approaches in the literature are (a) de novo and (b) database search methods, both of which can utilize deterministic, probabilistic and/or stochastic solution techniques.

The majority of peptide identification methods used in industry are database search methods [1-5] due to their accuracy and their ability to exploit organism information during the identification. A variety of techniques for peptide identification using databases currently exist. One approach, as implemented in the SEQUEST algorithm [1], uses a signal-processing technique known as cross-correlation to mathematically determine the overlap between a theoretical spectrum as derived from a sequence in the database and the experimental spectrum under investigation. The more frequently used technique, known as probability-based matching, utilizes a probabilistic model to determine whether an ion peak match between the experimental and theoretical tandem mass spectrum is actual or random [2,4,5]. Despite the sophistication of these database methods, they are ineffective if the database in which the search is conducted does not contain the corresponding peptide responsible for generating the tandem mass spectrum.

De novo methods have received considerable interest since they are the only efficient means for applications such as finding novel proteins, amino acid mutations and studying the proteome before the genome. A prominent methodology for the de novo peptide sequencing problem is a spectrum graph approach [6-10]. Various types of graph representations have been proposed, but the majority of methods map the peaks in the tandem mass spectrum to nodes on a directed graph, where the nodes are connected by edges if the mass difference between them is equal to the weight of an amino acid. Despite the vast potential of de novo methods, they can be computationally demanding and may exhibit inconsistent prediction accuracies.

## 2. Novel Method for De Novo Peptide Identification

### *2.1. Mathematical Model and Algorithmic Framework*

A tandem mass spectrum is comprised of the mass of the parent peptide ($m_P$) and a set of data point pairs corresponding to the mass-to-charge ratio of the ion

peaks (*mass*(*ion peak i*)) and their intensities ($\lambda_i$). The following sets are defined using these parameters:

$S = \{S_{i,j} = (i,j)$: $M_{i,j} \equiv$ *mass*(*ion peak j*) – *mass*(*ion peak i*) = mass of an amino acid, *mass*(*ion peak j*) > *mass*(*ion peak i*)$\}$ (1)

$C = \{C_{i,j} = (i,j)$: *mass*(*ion peak i*) + *mass*(*ion peak j*) = $m_P + 2\}$ (2)

The set S contains the pairs of peaks whose mass difference ($M_{i,j}$) is equal to the weight of an amino acid and the set C contains the pairs of peaks which are known as *complementary ions*. It is important to note that the pair (*i*,*j*) in C indicates that ion peak *i* and ion peak *j* are of different ion type.

A peptide sequence is derived from tandem mass spectrum data by connecting ion peaks of similar ion type by the weights of amino acids. This is nontrivial since the type of an ion peak (i.e., **a**, **b**, **c**, **x**, **y**, **z**) is not known a priori. The key idea of the proposed approach is two model the selection of peaks and connections between peaks using binary variables. We define the binary variable $p_k$ to equal one if ion peak *k* is used in the construction of the candidate sequence (i.e., $p_k = 1$), else $p_k$ is equal to zero. We also define the binary variable $w_{ij}$ to equal one if peaks *i* and *j* are connected in the construction of the candidate sequence (i.e., $p_i = p_j = 1$) and to be zero otherwise.

Based on the observation that **y**- and **b**-ions are typically the most abundant in intensity in a tandem mass spectrum, we postulate the objective function of maximimizing the intensities of the peaks used in the construction of the candidate sequence so as to maximize the number of **b**- or **y**-ions used.

$$\text{MAX} \sum_{(i,j)\in S_{i,j}} \lambda_j \cdot w_{i,j} \tag{3}$$

Several contraints can be added to the problem defined in Eq. (3) in order to incorporate various ion peak properties and fragmentation characteristics, as shown in Eqs. (4) – (11).

$$\text{MAX} \sum_{(i,j)\in S_{i,j}} \lambda_j \cdot w_{i,j}$$

s.t.

$$\sum_{(i,j)\in S_{i,j}} M_{i,j} \cdot w_{i,j} \le \left(m_P - 18\right) + tolerance \tag{4}$$

$$\sum_{(i,j)\in S_{i,j}} M_{i,j} \cdot w_{i,j} \ge \left(m_P - 18\right) - tolerance \tag{5}$$

$$p_i + p_j \le 1 \qquad \forall \left(i,j\right) \in C_{i,j} \tag{6}$$

$$\sum_{j \in S_{i,j}} w_{i,j} = p_i \qquad \forall i \in BC_i^{head} \tag{7}$$

$$\sum_{j \in S_{i,j}} w_{j,i} = p_i \qquad \forall i \notin BC_i^{head} \tag{8}$$

$$\sum_{i \in BC_i^{head}} \sum_{j \in S_{i,j}} w_{i,j} = 1 \tag{9}$$

$$\sum_{j \in BC_j^{tail}} \sum_{i \in S_{i,j}} w_{i,j} = 1 \tag{10}$$

$$\sum_{j \in S_{j,i}} w_{j,i} - \sum_{k \in S_{i,k}} w_{i,k} = 0 \qquad \forall i, i \notin BC_i^{head}, i \notin BC_i^{tail} \tag{11}$$

$$w_{i,j}, p_k = \{0,1\} \qquad \forall (i,j), k$$

The mass balance for the peptide is defined by Eqs. (4) and (5), which ensures that the sum of the masses used to derive the candidate sequence is within some error tolerance of the experimental mass of the parent peptide minus water. To eliminate ion peaks of a different ion type from being used in the peptide sequence, Eq. (6) enforces that if ion peak *i* is selected (i.e, $p_i = 1$) then its complementary ion, ion peak *j*, will not be selected (i.e., $p_j = 0$) since ion peak *i* and ion peak *j* are complementary ions (see Eq. (2)). The relationship between the binary variables *p* and *w* given in Eqs. (7) and (8) ensures that if ion peak *i* and ion peak *j* are choosen (i.e., $p_i = p_j = 1$) then a path exists between these two peaks ($w_{i,j} = 1$). Eqs. (9) and (10) require that the candidate sequence has the correct N- and C-terminus boundary conditions, which are predefined in the sets $BC_i^{head}$ and $BC_i^{tail}$, respectively. Eq. (11) enforces that the number of input paths entering and the number of output paths leaving an ion peak *i* are equal. The peptide identification problem is defined by Eqs. (3)-(11).

A preprocessing algorithm is used to filter the peaks in the raw tandem mass spectrum and to validate the existence of ion peaks pertaining to the N- and C-terminus boundary conditions of the ion series. To accommodate missing peaks in the tandem mass spectrum, a two-stage framework is employed in which the first stage sequences the candidate peptides using single amino acid weights and the second stage allows for combinations of two to three amino acids weights to be used in the construction of the candidate sequences.

Residue assignment ambiguities are subsequently resolved using a modified SEQUEST algorithm [1] so as to exploit the information in the tandem mass spectrum which was not utilized in the sequencing calculations. This postprocessing component of the method replaces weights in the candidate peptide sequences derived from the second stage calculations with permutations of amino acids consistent with these weights. The theoretical tandem mass spectrum for each candidate sequence is predicted and cross-correlated with the experimental tandem mass spectrum and the highest scoring sequence is

reported as the most probable peptide. This overall framework is denoted as PILOT, which stands for **P**eptide identification via mixed **I**nteger **L**inear **O**ptimization and **T**andem mass spectrometry.

### *2.2. Case study*

In this section we present a comparative study with several existing de novo peptide identification methods to demonstrate the predictive capabilities of the proposed framework PILOT. The algorithms examined in the comparison, that is, Lutefisk, LutefiskXP, PepNovo, PEAKS, EigenMS, NovoHMM, were selected on the basis of availability, reported popularity and performance. In the studies presented, assignments to isobaric residues (i.e., Q and K, I and L) are considered to be equivalent. To test the method's performance on quadrupole time-of-flight (QTOF) tandem mass spectra, we selected an existing data set that is publicly available [11]. These spectra were collected with Q-TOF2 and Q-TOF-Global mass spectrometers for a control mixture of four known proteins: alcohol dehydrogenase (yeast), myoglobin (horse), albumin (bovine, BSA), and cytochrome C (horse). The top-ranked sequence reported from each of these methods were compared using a number of metrics.

### *2.3. Results*

A summary of the identification results for the de novo methods on the 38 quadrupole time-of-flight spectra are presented in Table 1.

Table 1: Identification Rates for the 38 QTOF Spectra

| | Lutefisk | LutefiskXP | PepNovo | PEAKS | EigenMS | PILOT |
|---|---|---|---|---|---|---|
| Correct Peptides | 10 | 9 | 16 | 21 | 20 | 25 |
| with in 1 residue | 11 | 10 | 17 | 22 | 21 | 25 |
| with in 2 residues | 23 | 22 | 25 | 29 | 29 | 33 |
| with in 3 residues | 23 | 25 | 27 | 32 | 30 | 35 |
| Correct Residues | 245 | 294 | 337 | 366 | 353 | 381 |

In terms of correct peptide identifications, PILOT is superior to the other de novo methods with an identification rate of about 66 percent, followed by PEAKS and EigenMS, both at about 53 percent. A common limitation of de novo methods is the inability to assign the correct N-terminal amino acid pair or resolve isobaric residues (i.e., Q or GA, W or SV, etc.). Thus, to accommodate this limitation in the comparison, we also reported the percentage of predictions for which there are only one, two, or three incorrect amino acid assignments in the entire sequence. In Table 1, it is seen that allowing for up to three *incorrect* amino acids increases the identification rate for all methods on the order of 30 percent, indicating that these limitations affect the results reported by all the de novo methods. The last entry in Table 1 reports the number of correctly assigned residues normalized by the total number of actual residues (which is

418 for the 38 doubly-charged peptides considered). PILOT outperforms the other de novo methods with a residue accuracy of 91 percent.

## 3. Conclusions

A novel mixed-integer linear optimization framework, PILOT, was proposed for the automated de novo identification of peptides using tandem mass spectroscopy. For a given experimental MS/MS spectrum, PILOT generates a rank-ordered list of potential candidate sequences and a cross-correlation technique is employed to assess the degree of similarity between the theoretical tandem mass spectra of predicted sequences and the experimental tandem mass spectrum. A comparative study for 38 quadrupole time-of-flight spectra was presented to benchmark the performance of the proposed framework with several existing methods. For the case study presented, PILOT consistently outperformed the other de novo methods in several measures of prediction accuracy.

## Acknowledgements

The authors gratefully acknowledge financial support from the US Environmental Protection Agency, EPA (R 832721-010), Siemens Corporation, and the National Institutes of Health. Although the research described in the article has been funded in part by the U.S. Environmental Protection Agency's STAR program through grant (R 832721-010), it has not been subjected to any EPA review and therefore does not necessarily reflect the views of the Agency, and no official endorsement should be inferred.

## References

1. J.K. Eng, A.L. McCormack, and J.R. Yates, J. Am. Soc. Mass Spectrom., 5 (1994) 976.
2. D.N. Perkins, D.J.C. Pappin, D.M. Creasy, and J.S. Cottrell, Electrophoresis, 20 (1999) 3551.
3. P.A. Pevzner, Z. Mulyukov, V. Dancik, and C. L. Tang, Genome Research, 11 (2001) 290.
4. V. Bafna and N. Edwards, Bioinformatics, 17 (2001) S13.
5. M. Havilio, Y. Haddad, and Z. Smilansky, Anal. Chem., 75 (2003) 435.
6. J.A. Taylor and R.S. Johnson, Rapid Commun. Mass Spectrom., 11 (1997) 1067.
7. V. Dancik, T.A. Addona, K.R. Clauser, J.E. Vath, and P.A. Pevzner, J. Comp. Biol., 6 (1999) 327.
8. T. Chen, M.Y. Kao, M. Tepel, J. Rush, and G.M. Church, J. Comp. Biol., 10 (2001) 325.
9. A. Frank and P. Pevzner, Anal. Chem., 77 (2005) 964.
10. M. Bern and D. Goldberg, J. Comp. Biol., 13 (2006) 364.
11. B. Ma, , K.Z. Zhang, C. Hendrie, C. Liang, M. Li, A. Doherty-Kirby, and G. Lajoie, Rapid Commun. Mass Spectrom., 17 (2003) 2337.