AIChE WILEY InterScience® DISCOVER SOMETHING GREAT

A Mixed-Integer Optimization Framework for De Novo Peptide Identification

Peter A. DiMaggio, Jr. and Christodoulos A. Floudas
Dept. of Chemical Engineering, Princeton University, Princeton, NJ 08544

DOI 10.1002/aic.11061
Published online December 4, 2006 in Wiley InterScience (www.interscience.wiley.com).

A novel methodology for the de novo identification of peptides by mixed-integer optimization and tandem mass spectrometry is presented in this article. The various features of the mathematical model are presented and examples are used to illustrate the key concepts of the proposed approach. Several problems are examined to illustrate the proposed method's ability to address (1) residue-dependent fragmentation properties and (2) the variability of resolution in different mass analyzers. A preprocessing algorithm is used to identify important m/z values in the tandem mass spectrum. Missing peaks, resulting from 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 identify 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. © 2006 American Institute of Chemical Engineers *AIChE J,* 53: 160–173, 2007

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

Introduction

Of fundamental importance in proteomics is the problem of protein and peptide identification. Over the past few 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. First, the protein of interest is enzymatically digested to produce several peptides that are subsequently separated by HPLC and then volatized and ionized, usually by electrospray ionization (ESI). The mass-to-charge ratios (*m*/*z*) of these gas-phase peptides are measured using a mass spectrometer. Peptides with a specific *m*/*z* value are subsequently fragmented by collision-induced dissociation (CID) and the resulting *m*/*z* values of these ions are recorded by the mass analyzer—thus the term *tandem* mass spectroscopy. The essence of the peptide identification problem is to predict the corresponding peptide based on the peaks in the mass spectrum of its ions. The corresponding protein is then systematically assembled using these peptide predictions. Figure 1 shows a schematic representation of the overall process described for peptide identification.

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 frequent computational approaches reported in literature are (1) de novo and (2) database search methods, both of which can use deterministic, probabilistic, and/or stochastic solution techniques.

The majority of peptide identification methods used in industry are database search methods[1–9] because of their efficiency and ease of implementation. An essential component in most database methods is the use of a probabilistic method for scoring the experimental tandem mass spectra against the theoretical spectra of a peptide obtained from a protein database.[10]

Correspondence concerning this article should be addressed to C. A. Floudas at floudas@titan.princeton.edu.

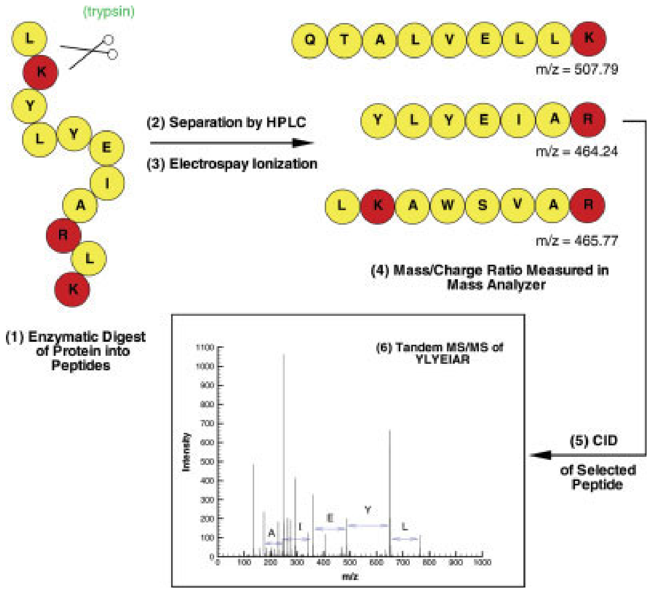

Figure 1. Overview of peptide identification.

[Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

The main difference between these database methods lies in the type of scoring function used to rank-order the most probable peptide matches and the type of sequence database in which the search is conducted. These database types include protein sequence, genomic, or expressed sequence tag.

A variety of techniques for peptide identification using databases currently exist. One approach, as implemented in the SEQUEST algorithm,[1–3] 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. However, this approach is limited to low-resolution data because it can be computationally expensive on large data sets.[11] Modifications to the scoring routine for SEQUEST have been proposed that address the decoupling of the scoring parameter from the size of the database, the search parameters used, and the sequence homologies.[12]

The more frequently used technique, known as *probability-based matching*, computes the probability of theoretically predicted fragments ions and then attempts to match these predictions to the experimental spectrum, starting with the most abundant ion series. This type of approach can be found in search engines such as Mascot.[4] It should be noted that these probability-based models exhibit several variations. The software SCOPE uses a stochastic model to predict the fragment ions for a given peptide in the database and then subsequently scores the match between the theoretical and experimental spectra using a fragmentation probability model, based on a statistical analysis of observed and known cleavages.[6] Another statistical framework uses a likelihood test that compares two explanations for the observed peaks[9]: (1) a statistical model that assumes the peak is a result of peptide fragmentation and (2) another model that assumes the peak is created by some other random process. The method also incorporates experimental and theoretical dependencies in developing the statistical model. A different approach uses a hypergeometric probability model to calculate the probability that a peptide sequence match to the experimental spectrum is random.[7] 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.

It is important to point out that the variety of scoring functions proposed indicates that the peptide with the best score is not necessarily the peptide responsible for generating the tandem mass spectrum. In light of this, significant efforts have been invested in the development of models to validate and interpret the results reported from existing database methods.[10,12–16] These so-called post-database search validation tools also exhibit a variety of forms. Several types of algorithms have been proposed to analyze the results of SEQUEST alone, ranging from statistical models based on linear discriminant analysis[14] to probability-based scoring procedures[13] to even the use of support vector machines for the classification of the scores reported.[16] Other post-validation approaches are aimed toward reducing several independent scoring schemes so that comparisons can be made between different database methods.[10,15]

De novo methods have received considerable interest because 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 identification problem is a spectrum graph approach,[17–27] which in many methods is solved using a dynamic programming algorithm.[18,20,25,26] This approach is based on graph theory, where each ion peak in the experimental spectrum generates multiple nodes (corresponding to different fragment ion types) in a directed acyclic graph, known as a sequence graph or a spectrum graph. Each node is labeled with the corresponding weight of the mass peak in the spectrum from which it was generated. A path through the graph corresponds to a candidate peptide and is constructed by connecting two vertices by a directed edge if the mass difference between these vertices is the weight of some amino acid. However, given that each peak in the experimental spectrum translates into several nodes on the spectrum graph, it is computationally intensive to enumerate all paths. Furthermore, precautions must be taken so that the path corresponding to the candidate sequence does not include multiple nodes associated with the *same* mass peak.[20] The nodes or edges of the spectrum graph are typically assigned scores based on empirically derived probabilities. An interesting post-processing technique is used by the de novo algorithm Lutesfisk,[17,21] where a modified version of FASTA (a homology-based database search program) is used to resolve ambiguous or unknown entries arising from missing ion peaks and isobaric residues. The algorithm EigenMS[27] uses spectral graph partitioning to resolve the problem of peak classification, which in turn results in superior predictions over other de novo techniques that use the best-path algorithm for constructing candidate sequences.

Although the spectrum graph approach is found in the majority of de novo algorithms to date, several alternative techniques have also been developed. For example, the de novo algorithm PEAKS[28] generates 10,000 potential sequences using a dynamic programming algorithm and then in a subsequent step reevaluates the predicted sequences using a stricter confidence scorer. A divide-and-conquer algorithm was recently proposed in which the tandem mass spectra of the *candidate* peptide is subsequently predicted using a quantitative kinetic model.[29,30] Another technique attacks the peptide identification problem by stochastic optimization using genetic algorithms to solve multiobjective models and can empirically test for independence between scoring functions.[31–33] The algorithm NovoHMM[34] uses a hidden Markov model to solve the peptide sequencing problem, where the observable random variables are the observed mass peaks and the hidden variables correspond to the unknown peptide sequence. Despite the vast potential of de novo methods, they are often computationally expensive and exhibit variable prediction accuracies.

Other approaches include sequence-tag–based hybrid methods where a partial sequence internal to the peptide, known as a "sequence tag," is determined using only the spectral information and the remaining portions of the sequence, which span from the ends of sequence tag to the N-terminal and C-terminal of the peptide, are determined by database searching.[35–37] Initially, the subsequence of the peptide (that is, the sequence tag) is derived using abundant intensity peaks in the high mass region of the tandem mass spectra, which correspond primarily to **y**-ions. A sequence similarity search program is then used to look up the remaining unknown portions around the sequence tag in a protein database and then score the corresponding entries. The major differences between the methodologies lie in how these results are ranked, which is primarily based on an empirical fragmentation model developed independently by the authors. This particular approach is advantageous because it combines the strengths from both de novo and database methodologies.

In this article, a novel mixed-integer optimization approach is introduced to efficiently address the de novo peptide identification problem so as to form a basis for a high-throughput computational framework. The section entitled "Mathematical Model for Peptide Sequencing" provides an outline of the formulation used for the sequencing of peptides by mixed-integer optimization, addressing the information concerning sets, parameters, variables, boundary conditions, constraints, and the objective function. The framework for the two-stage algorithmic approach is also presented in this section. In the section "Preprocessing of Spectral Data," an overview of the preprocessing algorithm used to identify certain peaks and to validate boundary conditions before the formulation of the MILP problem is provided. The section "Scoring Candidate Sequences" discusses a method for identifying the most probable sequence by cross-correlating the theoretical spectra of the candidate sequences with the experimental tandem mass spectrum. Computational studies are then presented in the final section.

Mathematical Model for Peptide Sequencing

This section provides a thorough description of the mathematical formulation for the de novo sequencing of peptides by tandem mass spectroscopy. We describe the essential components of the mixed-integer linear programming problem formulation: sets, parameters, binary variables, boundary conditions, constraint equations, and the objective function. The two-stage framework used to address missing peaks in the tandem mass spectrum is then subsequently presented.

Model description

Sets. Let the full variable space for the problem be represented by the matrix M, which constitutes the mass differences between *all* the ion peaks of the tandem mass spectrum:

$$M = \{M_{i,j} = mass(\text{ion peak } j) - mass(\text{ion peak } i): mass(\text{ion peak } j) > mass(\text{ion peak } i)\} \quad (1)$$

Note that the index i represents the rows and the index j represents the columns of the matrix $M_{i,j}$. Consider a set derived from the entries in $M_{i,j}$, denoted by $S_{i,j}$, which is created by the following definition:

$$S = \{S_{i,j} = (i,j): M_{i,j} = \text{ mass of an amino acid}\} \quad (2)$$

Every element in $S_{i,j}$ corresponds to the indices of an entry in $M_{i,j}$ that has a mass value equal to the weight of an amino acid. That is to say, the mass difference between peak i and peak j is equal to the weight of some amino acid for every $(i, j) \in \mathrm{S}_{i,j}$. The problem formulation will be considered only on this set $S_{i,j}$. Another useful set is derived from the following relation:

$$C = \{C_{i,j} = (i,j): mass(\text{ion peak } i) + mass(\text{ion peak } j) = \text{mass of peptide} + 2; i \neq j\} \quad (3)$$

The elements of the subset $\mathrm{C}_{i,j}$ correspond to the indices i and j such that the sum of the mass of peak i and the mass of peak j is equal to the weight of the parent peptide (as determined experimentally). The pairs of peaks for which this criterion is satisfied are known as *complementary ions*. As the charged parent peptide undergoes collision-induced dissociation (CID), it initially fragments into two ion pairs: either **a** and **x**, **b** and **y**, or **c** and **z**, where all three pairs are complementary ions by definition. This property helps to decipher the particular ion type a peak might have. However, one should note that further fragmentation of these ions is possible and frequently observed, which places limitations on how many complementary ions are actually detected in a spectrum.

When discussing the elements of these sets in a conceptual manner, it is important to understand that the index pair $(i,j) \in S$ graphically represents a "path" leading from peak i to a peak j of greater mass by the weight of some amino acid. Likewise, a path leaving peak j to some peak k of greater mass by the weight of an amino acid is represented by the element $(j,k) \in S$, with particular emphasis on the ordering of the indices. The combination of the paired elements (i,j) and (j,k) constitutes a continuous path through the peak j. This process of constructing a continuous, nonoverlapping path between peaks subject to certain constraints is the essence of the peptide sequencing problem. Classifying peak connections using the above sets in conjunction with properly formulated constraints enhances the computational efficiency of the sequencing algorithm by reducing the variable space.

Parameters. The relevant problem parameters correspond to the information contained in the tandem mass spectrum. It is important to note that the mass of the parent peptide and the masses of the ion peaks in the tandem mass spectra are subject to a certain degree of experimental error.[18] The parameters are

m_P = mass of parent peptide
$mass(\text{ion peak } i)$ = mass of ion peak i
λ_{i} = intensity of ion peak i

Binary Variables. Binary 0–1 variables are introduced in the problem formulation to model the selection of specific peaks (p_i) and the paths between peaks ($w_{i,j}$). The use of binary variables allows us to invoke logical inference when formulating the model constraints:

$$p_i = \begin{cases} 1 & \text{if peak}(i) \text{ is selected} \\ 0 & \text{otherwise} \end{cases}$$

$$w_{i,j} = \begin{cases} 1 & \text{if peaks } (i) \text{ and } (j) \text{ are connected by a path} \\ & (\text{that is, } (p_i = p_j = 1) \\ 0 & \text{otherwise} \end{cases}$$

The relationship between these two sets of binary variables is provided in a subsequent section.

Boundary Conditions. At the beginning and end of the candidate peptide sequence, there exists *only one* output and input path, respectively, which must be activated to derive a meaningful result. For instance, the candidate peptide derived using the **y**-ion series must begin at the weight of water (19 Daltons) and terminate at the weight of the parent peptide ($m_P + 1$), whereas deriving the same sequence using the **b**-ion series, the appropriate bounds become one unit mass and the weight of the parent peptide subtracted by the weight of water ($m_P - 17$), in respective order. To model this mathematically, two new sets are created to denote the boundary conditions at the “head” of the peptide and the “tail” of the peptide. Note that the sets presented below consider *only* the possibility for **b**- and **y**-ions in the candidate sequence:

$$BC_i^{\text{head}} = \{i : mass(\text{peak } i) = \{0, 19\} \text{ Daltons}\} \tag{4}$$

$$BC_j^{\text{tail}} = \{j : mass(\text{peak } j) = \{m_P - 17, m_P + 1\} \text{ Daltons}\} \tag{5}$$

Under certain conditions it is necessary to “adjust” the boundary conditions if it is known a priori that specific peaks are missing in the spectrum, as will be described in a later section.

Constraints. Several constraints derived from ion properties and graph theory are formulated in terms of the binary variables by logical inference. One set of constraints is associated with the existence of complementary ions in the mass spectra:

$$p_i + p_j \ \leq \ 1 \quad \forall (i,j) \in C_{i,j} \tag{6}$$

One can infer from Eq. 6 that if peak i is selected (say $p_i = 1$), then peak j, its complementary ion, is not ($p_j = 0$), and vice versa. This is desired because, by definition, peak i and peak j are of different ion types and the candidate peptide is sequenced by connecting ions of the *same* type.[38]

Another constraint to consider is the conservation of mass imposed by the parent peptide on the candidate sequence:

$$\sum_{(i,j) \in S_{i,j}} M_{i,j} \cdot w_{i,j} \leq (m_P - 18) + tolerance \tag{7}$$

$$\sum_{(i,j) \in S_{i,j}} M_{i,j} \cdot w_{i,j} \geq (m_P - 18) - tolerance \tag{8}$$

It is well known that the experimentally measured parent mass/charge ratio is subject to a certain degree of experimental error.[18] Thus, the algorithm must accommodate for this inherent deviation without overconstraining the problem and simultaneously preventing the inclusion of possibility false candidates. This issue is dealt with by allowing for a tolerance of error of ±2 Daltons above and below the parent peptide mass. It is also possible to formulate the *tolerance* term as a variable and then incorporate it into the model such that its value is minimized.

The next constraint, associated with the paths that connect peaks in a tandem mass spectrum, is referred to as the *flow conservation law* from graph theory, which has been used extensively in process synthesis problems.[39–53]

$$\sum_{j \in S_{j,i}} w_{j,i} - \sum_{k \in S_{i,k}} w_{i,k} = 0 \qquad \forall i, \ i \notin BC_i^{\text{head}}, \ i \notin BC_i^{\text{tail}} \tag{9}$$

This constraint ensures that the paths constructed in the sequencing calculations are continuous and nondegenerate. For instance, consider the spectrum graph representation for the tandem mass spectrum of a particular peptide as shown in Figure 2. One can easily observe that there are several paths spanning various portions of the spectrum graph and that at every node the number of possible input and output paths to other nodes varies. Now consider the enlarged nodes shown in Figure 2. Applying Eq. 9 generates the following path constraints:

- Node 66

$$w_{61,66} = 0$$

- Node 67

$$w_{57,67} + w_{62,67} + w_{63,67} - w_{67,76} = 0$$

- Node 68

$$w_{58,68} + w_{59,68} - w_{68,77} = 0$$

It is trivial, but necessary, that $w_{61,66} = 0$ for node 66 because a path representing a candidate sequence can initiate/terminate only with the nodes that are elements of BC^{head}/BC^{tail}, respectively. A path terminating at node 66 would result in a peptide that has a mass much less than that of the parent peptide. Nodes 67 and 68 have the possibility for multiple input paths but only one possible output path. Let us consider only node 67 because node 68 is subject to the same analysis. The equality in the constraint for node 67 implies that the output path for these nodes will be activated (that is, $w_{67,76} = 1$) *if and only if* one corresponding input path is selected (that is, $w_{57,67} = 1$, or $w_{62,67} = 1$, or $w_{63,67} = 1$). This constraint also enforces that at most one input path could be selected, given that only one output path exists. Additionally, if none of the input paths is selected (that is, $w_{57,67} = w_{62,67} = w_{63,67} = 0$) then the output path will not be activated in the construction of the sequence.

Equations 10 and 11 ensure that the candidate sequence has the appropriate C-terminus and N-terminus boundary conditions:

$$\sum_{i \in BC_i^{\text{head}}} \sum_{j \in S_{i,j}} w_{i,j} = 1 \tag{10}$$

$$\sum_{j \in BC_j^{\text{tail}}} \sum_{i \in S_{i,j}} w_{i,j} = 1 \tag{11}$$

The above equations are restricted to the subsets BC^{head} and BC^{tail}, defined in Eqs. 4 and 5, respectively, for which a bound-

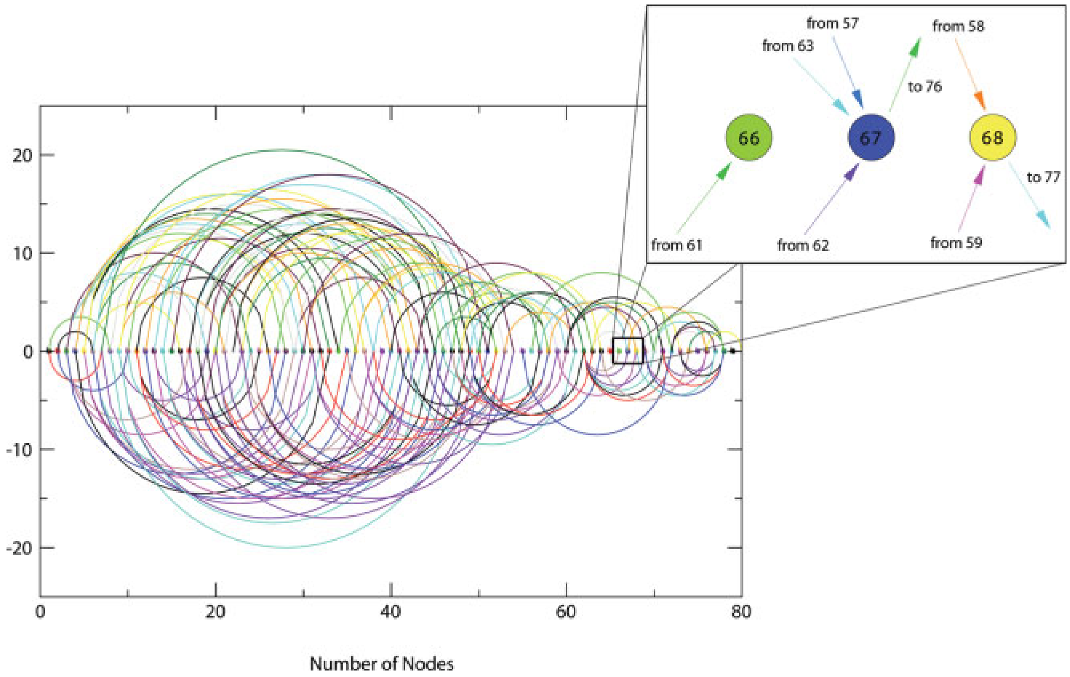

Figure 2. Spectrum graph representation of tandem mass spectrum.

[Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

ary condition is plausible. As previously mentioned, the elements of these boundary conditions can be modified in the event that specific peaks that denote the beginning and end of an ion series are missing in the tandem mass spectrum. Furthermore, these constraints enforce the nondegeneracy of paths, given that *only one* path can initiate and terminate the sequence, respectively.

The final set of constraints establishes the relationship between the binary variables p_i and $w_{i,j}$:

$$\sum_{j \in S_{i,j}} w_{i,j} = p_i \qquad \forall i \in BC_i^{\text{head}} \qquad (12)$$

$$\sum_{j \in S_{i,j}} w_{j,i} = p_i \qquad \forall i \notin BC_i^{\text{head}} \qquad (13)$$

Note that both Eqs. 12 and 13 are necessary. Equation 13 couples the binary variables representative of all the paths (j, i) that *enter* a given node i and is written for all nodes, including those that belong to the "tail" boundary condition set (see Eq. 5). However, the nodes belonging to the "head" boundary condition set (see Eq. 4) do not have entering paths because they begin the candidate sequence and thus should not be included in Eq. 13. These nodes are addressed in Eq. 12, which relates the binary variables for the paths (i, j) *leaving* any "head" boundary condition node i. Equations 12 and 13 together complete the relationship between the binary variables p_i and $w_{i,j}$. These constraints are advantageous because, if it is desired to omit a specific peak from the candidate sequence, simply deactivating the binary variable that represents that peak (p_i), in conjunction with the previous path constraints, deactivates *all* input and output paths to/from that peak. For instance, this is useful for eliminating the precursor ion and multiply charged ions from consideration.

Objective Function. The objective function is postulated as an explicit function of the peak intensities based on the observation that **b**- and **y**-ions are typically the most abundant peak types found in the tandem mass spectrum.[38] By formulating the objective function in this way, the algorithm attempts to maximize the number of **b**- or **y**-ions in the candidate sequence. Note that before the model's formulation, it is decided whether the candidate peptide is to be sequenced using the **b**-ion series or the **y**-ion series. Thus, for each ion type, a different objective function is formulated based on the direction of the sequencing, which for the **b**-ion series is N-terminus to C-terminus and vice versa for the **y**-ion series.

$$\max_{p_k, w_{i,j}} \sum_{(i,j) \in S_{i,j}} \lambda_i \cdot w_{i,j} \qquad \text{for the } \mathbf{b} \text{ ion series} \qquad (14)$$

$$\max_{p_k, w_{i,j}} \sum_{(i,j) \in S_{i,j}} \lambda_j \cdot w_{i,j} \qquad \text{for the } \mathbf{y} \text{ ion series} \qquad (15)$$

It is also important to note the relative trends in intensity of the **b**- and **y**-ion series. The **b**-ions are usually abundant in intensity in the low-mass region of the spectrum and gradually, but almost never monotonically, decrease in intensity with increasing mass-to-charge ratio. Conversely, the **y**-ions are most abundant in the high-mass region and gradually decrease in intensity with decreasing mass-to-charge ratio.[38]

Equations 1–15 constitute the entire mathematical formulation for the de novo peptide identification problem using tandem mass spectra. The entire problem formulation is summarized below for the **y**-ion series:

$$\begin{aligned}
&\max_{p_k, w_{i,j}} \sum_{(i,j)\in S_{i,j}} \lambda_j \cdot w_{i,j} \\
&\text{s.t.} \sum_{(i,j)\in S_{i,j}} M_{i,j} \cdot w_{i,j} \le m_P + tolerance \\
&\sum_{(i,j)\in S_{i,j}} M_{i,j} \cdot w_{i,j} \ge m_P - tolerance \\
&p_i + p_j \le 1 \qquad \forall (i,j) \in C_{i,j} \\
&\sum_{j\in S_{i,j}} w_{i,j} = p_i \qquad \forall i \in BC_i^{\text{head}} \\
&\sum_{j\in S_{i,j}} w_{j,i} = p_i \qquad \forall i \notin BC_i^{\text{head}} \\
&\sum_{i\in BC_i^{\text{head}}} \sum_{j\in S_{i,j}} w_{i,j} = 1 \\
&\sum_{j\in BC_j^{\text{tail}}} \sum_{i\in S_{i,j}} w_{i,j} = 1 \\
&\sum_{j\in S_{j,i}} w_{j,i} - \sum_{k\in S_{i,k}} w_{i,k} = 0 \qquad \forall i, i \notin BC_i^{\text{head}}, i \notin BC_i^{\text{tail}} \\
&w_{i,j}, p_k = 0 - 1 \qquad \forall (i,j), (k)
\end{aligned} \tag{P}$$

The resulting problem (P) is a mixed-integer linear programming (MILP) problem and can be solved to optimality using existing methods (such as CPLEX[54]). Throughout the remainder of this article, the sequencing of the candidate peptides is attempted using first the **y**-ion series and then the **b**-ion series if warranted. The algorithm attempts to derive several candidate sequences by using only *single* amino acid residue weights to connect mass peaks. Because the algorithm seeks the next optimal solution, previous solutions are eliminated by the introduction of integer cuts. Thus, for every solution, an integer cut is incorporated into the model using the following general form[55]:

$$\sum_{(i,j)\in B} w_{i,j} - \sum_{(i,j)\in NB} w_{i,j} \le |B| - 1 \tag{16}$$

where $B = \{(i,j) : w_{i,j} = 1\}$, $NB = \{(i,j) : w_{i,j} = 0\}$, and $|B|$ is the cardinality of B.

Missing ion peaks: two-stage algorithmic approach

In practice, tandem mass spectra are far from ideal because of the absence of certain ion mass peaks. These missing peaks can be attributed to fragmentation issues primarily associated with proline (P) and glycine (G) and how they affect the transfer of the proton along the peptide backbone (according to mobile proton theory),[38] whereas repeated amino acids or subsequences within the peptide do not inhibit this mechanism of reaction. For spectra that are inherently missing mass peaks, it is still possible to derive nonoptimal candidate sequences using solely single amino acid weights. For this reason, two consecutive MILP problems are solved, which are referred to as stages 1 and 2. The first stage consists of using only single amino acid weights to derive the candidate peptide sequences. A subsequent MILP problem, stage 2 of the algorithm, is then solved for which the possibility of using *two* amino acid weights, if needed, is allowed to connect mass peaks in the tandem mass spectra. That is to say, in this stage the additional option is available to generate a path between peaks by the weight of any two combined amino acids, a full list of which is available elsewhere.[56] However, in the second stage the emphasis is again placed on primarily using the weights of single amino acids to construct the candidate sequences. Use of the two combined amino acid weights is "penalized" in the expression for the objective function by multiplying those peak intensities by a weighting fraction <1 and decreases with increasing mass error. As a result, the driving force for the algorithm is the single residue weights, whereas the double residue weights are used only to bridge the gap between disjoint single residue segments of the candidate sequence. Note that the proposed approach can predicted repeated amino acids and repeated subsequences.

Solution information from the first-stage computations are used in the second-stage MILP. For instance, if a single amino acid residue in the optimal candidate sequence connects two peaks (say p_i and p_j) whose intensities are greater than that of the $\mathbf{y}_1$-ion and are *both* active in the complementary ion constraints, then their binary variables are activated for the second-stage MILP (that is, $p_i = 1$ and $p_j = 1$). The reason for activating these mass peaks is based on the assumption that their complementary ions are in fact **b**-ions. Note that a peptide derived using the **b**-ion series is the same as that derived by the **y**-ion series but its residues are in reverse order, so these complementary **b**-ions serve as a "validation" of the proposed candidate peptide that was sequenced using the **y**-ion series. Furthermore, integer cuts from the first stage are also passed onto the second stage to eliminate previous solutions from being revisited. All candidate sequences from both the first- and second-stage computations are then examined individually for validity.

Preprocessing of Spectral Data

Before formulating the MILP problem, the MS/MS data are analyzed using a preprocessing algorithm with the intent of elucidating key spectral features. In particular, certain ion types are sought to confirm the proposed boundary conditions previously mentioned. First, the raw spectrum is scanned and compared with a data file for the existence of the typically abundant in intensity $\mathbf{b}_2$-ion,[38] whose validity can be confirmed by its complementary $\mathbf{y}_{n-2}$-ion. If the corresponding $\mathbf{y}_{n-2}$-ion is found, then the two possible $\mathbf{y}_{n-1}$-ions are computed using the mass of the parent peptide and the weights of the amino acids, which constitute the $\mathbf{b}_2$-ion (see Table 1), and the spectrum is once again searched to confirm these proposed peaks as well as existence of the $\mathbf{a}_2$-ion. This step is important because

Table 1. Ions Identified by the Preprocessing Algorithm

Ion Type	Relation to the $\mathbf{b}_2$-ion
$\mathbf{y}_{n-2}$	$(m_P + 2) - \mathbf{b}_2$
$\mathbf{a}_2$	$\mathbf{b}_2 - 28$
$\mathbf{b}_1$	AA_1 or AA_2 where $AA_1 + AA_2 + 1 = \mathbf{b}_2$
$\mathbf{y}_{n-1}$	$(m_P + 2) - AA_1$ or $(m_P + 2) - AA_2$ where $AA_1 + AA_2 + 1 = \mathbf{b}_2$

the absence of a $\mathbf{y}_{n-1}$-ion indicates that the peptide cannot be completely sequenced from the raw spectral data using the **y**-ion series alone. However, if it is suspected that only the $\mathbf{y}_{n-1}$-ion is missing, then the original N-terminus or "tail" boundary condition for the **y**-ion series (Eq. 5) can be adjusted by terminating the proposed sequence at the mass of the $\mathbf{y}_{n-2}$-ion. The appropriate $\mathbf{y}_{n-1}$-ion can then be subsequently determined independently of the MILP problem from the $\mathbf{b}_2$-ion and other spectral information, such as existing immonium ions. The mathematical relationship between these ions is provided in Table 1.

Each probable $\mathbf{b}_2$-ion pair is assigned a score based on the intensities of the supporting ions found (that is, $\mathbf{y}_{n-2}$, $\mathbf{y}_{n-1}$, $\mathbf{a}_2$) and their isotopic offsets and neutral losses of water and ammonia. These scores are normalized according to the location of the $\mathbf{b}_2$-ions' mass in the tandem mass spectrum. This is necessary because ion intensities near the ends of the mass spectrum are statistically the lowest[57] and the normalization removes the scoring bias for heavier $\mathbf{b}_2$-ions. In the case where the algorithm is unable to find any probable $\mathbf{b}_2$-ions—which implies the absence of a feasible boundary condition—the high-mass end of the spectrum is reexamined and the peaks with largest relative intensity are selected as the most probable upper bounds for the **y**-ion series. A presequencing MILP is formulated that computes only the optimal candidate peptide using *each* peak as the upper bound of the **y**-ion series. The peak corresponding to the maximum objective function value is then selected to be used as the appropriate boundary condition throughout the subsequent sequencing calculations. This instance arises mostly as the result of fragmentation issues associated with the peptide and can often be attributed to the presence of a basic internal residue such as arginine, histidine, or lysine.

The preprocessing algorithm can also be used to elucidate which peak is the C-terminal peak for the **y**-ion series. For instance, if the peptide of interest is the product of proteolytic digestion using trypsin, then the $\mathbf{y}_1$-ion must be either a C-terminal lysine, identified by an *m*/*z* peak at 147, or a C-terminal arginine, identified by an *m*/*z* peak at 175.[38] In the examples considered, it is known a priori that the peptide is a tryptic peptide and this information is used when constructing boundary conditions.

The existence of immonium ions in the spectrum is significantly important in the interpretation process. These low-mass ions are indicative of specific residues and, although they do not provide any information regarding the position of the amino acids in the peptide, they are useful for validating residues in the *predicted* sequence. A thorough list of immonium ions is reported elsewhere.[38,56]

For high-resolution mass spectra, the preprocessing algorithm helps to decipher ions with a charge >1, based on the offset of its corresponding isotopic carbon peaks. For instance, an isotopic offset of 1.0 Dalton indicates that the ion peak is singly charged; an isotopic offset of 0.5 Daltons indicates that the ion peak is doubly charged; an isotopic offset of 0.33 Daltons indicates that the ion peak is triply charged; and so on.[38] To avoid misinterpreting multiply charged ions as singly charged ions, the preprocessing algorithm examines the isotopic carbon offsets of ion peaks to determine its charge and then postulates new peaks by multiplication of each charged *m*/*z* value by its corresponding charge value. Using this approach allows for the freedom of either keeping the originally multiply charged ions in the spectrum or eliminating them by the deactivation of their corresponding binary variables. Another feature of the algorithm is the option to deactivate ion peaks that offset other peaks by either 17, 18, or 28 Daltons, signifying losses of ammonia, water, and carbon monoxide, respectively (a characteristic of low-energy CID[38,56]). A filtering technique is applied to every tandem mass spectrum and only the top 125 peaks of highest intensity are used in the problem formulation.

Scoring Candidate Sequences

To determine the most probable sequence from a rank-ordered list of candidate sequences and their permutations (in the case of weights in the sequence), the theoretical MS/MS spectrum for each sequence is predicted and then compared to the given experimental tandem mass spectrum. Several techniques have been developed to assess the degree of similarity between the experimental and theoretical spectra of the predicted sequences. In particular, probabilistic matching[4,6,9] and cross-correlation[1–3] have proved to be effective tools for this purpose. In this section, we describe a method for generating theoretical tandem mass spectra from predicted sequences to be correlated with the experimental spectrum based on a modified version of the SEQUEST algorithm.[1]

Recall that the candidate peptides were sequenced using the **y**- or **b**-ion series only. Thus, it would be beneficial to use various other types of ions when scoring these candidate peptides to exploit as much information as possible from the tandem mass spectrum. To minimize the assignment of random matches between the theoretical tandem mass spectra of predicted sequences and experimental spectra, the types of ions featured in the theoretical spectra were selected based on observations reported from various sources in the literature. The isotopic carbon offsets of **b**-ions (that is, $\mathbf{b} + 1$ and $\mathbf{b} + 2$ and similarly for the **y**-ions) were included based on observations that isotopic shifts in the **b**- and **y**-ion series are nearly as common as the **b**- and **y**-ion series themselves and *more common* than various other ions, such as **a**-ions and ions resulting from neutral losses of water and ammonia.[24] The neutral losses of water, ammonia, and combinations thereof from **b**-ions (that is, $\mathbf{b}$-H_2O, $\mathbf{b}$-NH_3, $\mathbf{b}$-H_2O-NH_3, $\mathbf{b}$-H_2O-H_2O, and similarly for the **y**-ions) are also included because their existence serves as a measure of support for their corresponding **b**- or **y**-ions. Doubly charged **b**-ions were not included in the theoretical spectra because the major pathways of backbone fragmentation (based on the mobile proton theory) do not facilitate their formation, whereas doubly charged **y**-ions are generated by these mechanisms.[38] A common dissociation reaction pathway for **b**-ions is the elimination of carbon monoxide to form the **a**-ion series.[58] It should be noted that, based on empirical observations, the doubly charged **y**-ions and **a**-ions are predicted for only the first-half of the theoretical tandem mass spectrum.[26] Although it has been reported that the energy of fragmentation in low-energy CID might be insufficient to break the bond between the α-carbon and the carbonyl,[57] **x**-ions are included in the C-terminal ion series of the theoretical MS/MS. Also included in the peaks of this spectra are what are known as internal fragment ions[56,59–61] (ions that have lost both their C-terminal and N-terminal ends).

There has been a recent surge of interest in developing adaptive models that incorporate residue chemistry and position dependencies into intensity predictions. For instance, a 236-parameter kinetic model based on the mobile proton theory has been developed for the simulation of ion-trap tandem MS.[29] Another study statistically quantified the influence of neighboring residues on fragmentation and relative intensity trends for ion-trap spectra.[57] Although these studies provide valuable insight toward the development of sequence-dependent predictive models, they are restricted to use with ion-trap spectra only. Because spectra analyzed by our proposed algorithm are not restricted to any one type of instrument, a modified SEQUEST representation for ion intensities was adopted. That is, using a normalized scale, **y**- and **b**-ions were assigned an intensity of 1, their isotopic offsets were assigned an intensity of 1/2, and neutral losses from these ions were assigned an intensity of 1/5, as described in the original SEQUEST model.[1] In addition, certain intensity dependencies on residue types were introduced into the model; that is, a neutral loss of water from a fragment ion that contains either a D, E, S, or T residue and a neutral loss of ammonia from a fragment ion that contains either an N or Q residue is assigned an intensity of 1/3 rather than 1/5. The reason for favorably weighting these offsets is based on the observation that these types of fragments are statistically likely to contain those residues mentioned.[29,57]

To increase the predictive capabilities of the model, a reward/penalty system was created for certain ion types. For instance, a match between a predicted **y**-ion and a peak in the experimental spectra is more probable if the corresponding **y**-ion offsets also have matches in the experimental spectra.[26] Therefore, the score for a **y**-ion is rewarded by 5% for each existence of a supporting ion (that is, a spectrum match for an isotopic offset or neutral loss ion); otherwise, its score remains unaltered. Conversely, the isotopic offsets and neutral loss ions are penalized in the absence of a match in the experimental spectrum for its corresponding **y**- or **b**-ion. By adopting this convention we aim to reduce the contributions of random matches to the overall score of a given peptide. The overall score is weighted by the ratio of the number of hits between the theoretical and experimental spectra to the total number of ions in the theoretical spectra.

Computational Studies

In this section we discuss computational studies for three classes of problems: (1) a benchmark spectrum, (2) quadrupole time-of-flight spectra, and (3) ion-trap spectra. All subsequent MILPs reported in this section and throughout the document were solved to optimality using CPLEX.[54] In the tables of candidate sequences presented throughout this section, "Obj" denotes the value in the objective function, "Avg" corresponds to the average of the peak intensities, "STD" is the standard deviation of the peak intensities, and "Comp Ions" is an abbreviation for the number of complementary ions that are activated in the constraint set.

One should also note the information conveyed by bold and italicized residues in these candidate sequences. The italicized residues indicate that the mass peak on the amide end of the amino acid has an intensity *less* than that of the $\mathbf{y}_1$-ion. The bold residues, on the other hand, denote that the peaks on both the amide and carboxyl ends of the amino acid are complementary ions and have intensities *greater* than that of the $\mathbf{y}_1$-ion. These conventions are adopted throughout the remainder of the article. From computational experience, it is generally observed that the accuracy of a peptide sequence follows the trends of maximum objective function, maximum average and minimum standard deviation of peak intensities, maximum number of bold residues, and minimum number of italicized residues. These terms provide site-specific and average estimates of the likelihood that **b**- or **y**-ions were used to construct the candidate sequences because these ion intensities are consistently abundant and follow approximate trends,[38] which these measures attempt to quantify.

Benchmark problem

The first step to validating an algorithm is to test the approach on small-scale examples whose solutions have been reported and are established. For this step, several example problems were chosen from the literature[38] as a cornerstone for evaluating the performance of the algorithm. The solutions to these examples as provided by Kinter and Sherman[38] were determined by a nine-step methodology specifically used to interpret the product ion spectra of tryptic peptides. The specific example presented here was selected because it was categorized as a difficult spectrum to interpret. No filtering of the data was required for these spectra because only peaks with high intensities were provided.

TXAMoDGTEGXVR. This example is taken from Kinter and Sherman[38] and the tandem mass spectrum is provided in Figure 4.30 of this reference. This peptide was selected because it illustrates fragmentation issues characteristic of an internal glycine residue. The preprocessing algorithm identifies the C-terminal amino acid for the peptide to be arginine, assigned to a strong intensity peak at an *m*/*z* of 175.10 Th. Furthermore, immonium ion peaks of 86.10 and 102.10 Th in the low-mass region suggest that the residues (L/I) and E are most likely constituents of the peptide. Two strong intensity *m*/*z* peaks at 215.20 and 286.20 Th result in several possible $\mathbf{b}_2$-ion combinations. However, the higher-intensity peak at an *m*/*z* of 215.20 also exhibits a complementary $\mathbf{y}_{n-2}$-ion of higher intensity, suggesting (L/I)T, T(L/I), VD, or DV are the more probable residue pairs. A $\mathbf{b}_1$-ion of moderate intensity, consistent with the amino acid threonine (T), confirms the appropriate choice for the $\mathbf{b}_2$-ion to be T(L/I). Because there exists no probable $\mathbf{y}_{n-1}$-ions, the "tail" boundary condition of the N-terminus, shown in Eq. 5, is replaced by the $\mathbf{y}_{n-2}$-ion. The first-stage problem formulation contains 94 variables and 172 constraint equations.

In 0.17 CPU seconds, CPLEX determines that no solutions are feasible for this first-stage MILP, forcing the predictions to be based solely on the candidate sequences from the second stage, as reported in Table 2. The second-stage model, which consists of 94 variables and 174 constraint equations, is solved to optimality in a CPU time of 0.46 s. Nine out of the ten sequences reported in Table 2 are consistent with the optimal peptide, A(Mo/F)DGTE[170.20]VR, and so we shall consider only this peptide for subsequent analysis. The only internal doublet weight to resolve is 170.20 Th, which corresponds to either AV, VA, (L/I)G, or G(I/L). The poor fragmentation characteristics of glycine in low-energy CID are well known,

Table 2. Distinct Solutions from the Second-Stage MILP for TXAMoDGTEGXVR

Candidate Sequence	Obj	Avg	STD	Comp Ions
A*(Mo/F)*DG*TE*[170.20]*V*R	3.3660	0.4690	0.3503	3
A*(Mo/F)*DG*TE*[269.30]R	3.2160	0.5044	0.3520	3
A*(Mo/F)*DG*T[299.20]V*R	3.1320	0.4811	0.3693	3
A*(Mo/F)*DG*[230.10]*[170.20]*V*R	3.1100	0.4967	0.3597	3
A*(Mo/F)*DG*[141.70][258.60]V*R	3.0920	0.4767	0.3730	3
A*(Mo/F)*DG*[230.10]*[269.30]R	2.9600	0.5400	0.3586	3
A*(Mo/F)*D[158.10]*E*[170.20]*V*R	2.7680	0.5167	0.3353	3
A*(Mo/F)*D[158.10]*E*[269.30]R	2.6180	0.5625	0.3270	3
A*(Mo/F)*D[287.10][170.20]*V*R	2.5480	0.5537	0.3382	3
A*(Mo/F)*D[158.10]*[299.20]V*R	2.5340	0.5363	0.3529	3

but unlike the intensity variations observed with proline, glycine yields no high abundance ions.[38] Despite the lack of validating information, this inherent gap in the sequence is most likely explained by the residue pairs G(I/L) or (I/L)G. Finally, offsets of 64 Th from the **y**-ion series in the high-mass region of the spectrum suggests the residue Mo over F in the candidate peptide. Thus, the most complete sequence to be reported based purely on experimental observations, without resorting to guessing the most probable order of the doublet residue pair, is T(L/I)AMoDGTE[170.20]VR. The cross-correlation method was not used to resolve the weight of 170.20 Th because the mass spectrum provided was labeled with a small number of mass peaks. A database technique would perform reasonably well for this benchmark problem because it can provide a peptide identification without complete fragmentation information. However, one should be cautious of the validity of such a database identification because the "most probable" peptide is reported *relative* to all the other peptides in the protein database searched, which can obscure the actual quality of the overall identification.

Quadrupole time-of-flight MS/MS data

To test the effectiveness of the proposed methodology on full-scale problems, the tandem mass spectra for doubly charged peptides generated from the quadrupole time-of-flight and ion-trap mass spectrometers were examined. Two different instruments are studied because the quality of the data that they generate is vastly different. Thus, it is important to develop a robust methodology whose prediction results are *instrument independent*. The MS/MS data for the peptides in this section were obtained from a quadrupole time-of-flight mass analyzer,[28] which is known for its accurate *m*/*z* resolution and extensive coverage of the *m*/*z* range.[38] The examples presented were selected to illustrate the diversity of the various spectral artifacts and how the framework handles issues encountered in tandem mass spectra.

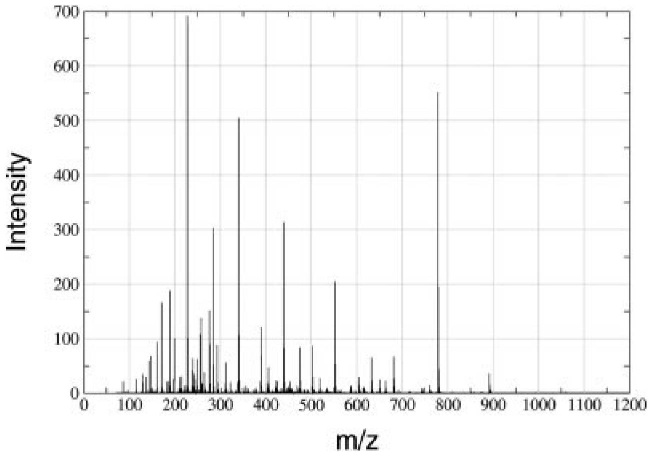

Figure 3. Tandem mass spectrum for STLPEIYEK.

Table 3. Distinct Solutions from the First-Stage MILP for STLPEIYEK

Candidate Sequence	Obj	Avg	STD	Comp Ions
(**L**/**I**)P*E*(L/I)YEK	2.9595	0.4949	0.3846	2
(**L**/**I**)P*Q(L/I)*VGHK	2.7803	0.4200	0.4036	2
(**L**/**I**)P*E*M*H*HK	2.7428	0.4678	0.4113	2
(**L**/**I**)P*QN*YEK	2.7399	0.4675	0.4080	2
(**L**/**I**)Q*P(L/I)*VGHK	2.7197	0.4133	0.4101	2
(**L**/**I**)P*QV(L/I)*GHK	2.6734	0.4082	0.4150	2
(**L**/**I**)P*Q(L/I)*RHK	2.6257	0.4532	0.4181	2
(**L**/**I**)Q*PV(L/I)*GHK	2.6127	0.4014	0.4211	2
(**L**/**I**)Q*PN*YEK	2.6792	0.4599	0.4163	2

STLPEIYEK. This peptide illustrates how the algorithm handles missing $\mathbf{y}_{n-1}$-ions. Figure 3 contains the raw tandem mass spectrum for the peptide, originally consisting of 11,437 mass peaks.

The preprocessing algorithm identifies the top-scoring N-terminal pairs to be ST and GM, which both share the same $\mathbf{b}_2$- and $\mathbf{y}_{n-2}$-ion masses. However, no supporting $\mathbf{y}_{n-1}$-ions exist in the tandem mass spectra for ST and GM and therefore the ordering of their residues is not known. Because the $\mathbf{y}_{n-1}$-ion is missing for these two pairs, the de novo sequencing algorithm will terminate the **y**-ion series at the mass of the $\mathbf{y}_{n-2}$-ion (891.44 Th). A significant peak at an *m*/*z* of 147.11 indicates that the C-terminal amino acid of the peptide is lysine (K), and the binary variables that represent the peaks for this amino acid are activated. The stage 1 formulation contains 571 binary variables and 276 constraint equations and is solved to optimality by CPLEX in 3.96 CPU seconds. The statistically distinct solutions from stage 1 are shown in Table 3.

Note that the optimal candidate sequence, (L/I)PE(L/I)YEK, also exhibits the maximum average and minimum standard deviation of peak intensities. Several of the suboptimal candidate sequences reported in Table 3 contain an internal histidine (H), which is not a very common feature of doubly charged tryptic peptides, given that a basic residue internal to the sequence typically results in a charge state >2. The problem formulation for the second stage contains 2152 variables and 288 constraint equations and is solved to optimality in 5.87 CPU seconds. The statistically distinct solutions from stage 2 are shown in Table 4.

The missing N-terminal mass for the above sequences is 188.04 Th, which, as identified by the preprocessing algorithm, corresponds to ST, TS, MG, or GM. To address this and the other weights in the candidate peptides in Table 4, permutations of amino acids consistent with the weights are substituted into the sequences. The theoretical spectrum for each distinct sequence is then generated and cross-correlated with the

Table 4. Distinct Solutions from the Second-Stage MILP for STLPEIYEK

Candidate Sequence	Obj	Avg	STD	Comp Ions
(L/I)P*[242.11]*YEK	2.5762	0.5233	0.4063	2
(L/I)D*[224.12]*YEK	2.5698	0.5142	0.4180	2
(L/I)Q*[211.09]*YEK	2.5701	0.5147	0.4174	2
(L/I)N*[225.13]*YEK	2.5678	0.5114	0.4219	2
(L/I)P*[241.10]*VGHK	2.6166	0.4630	0.4089	2
(L/I)P*QV[170.08]*HK	2.5084	0.4398	0.4318	2
(L/I)P*E*[211.11]GHK	2.6149	0.4852	0.3920	2
(L/I)P*Q[212.12]*GHK	2.5929	0.4577	0.4142	2

Table 5. Distinct Solutions from the First-Stage MILP for DAFLGSFLYEYSR

Candidate Sequence	Obj	Avg	STD	Comp Ions
F(L/I)GSF(L/I)YH*G*ANR	2.9850	0.3065	0.3420	7
F(L/I)GSF(L/I)YH*AG*NR	2.9674	0.3052	0.3431	7
F(L/I)GSF(L/I)YQHNR	2.9499	0.3292	0.3469	7
F(L/I)GSF(L/I)YH*Q*NR	2.9374	0.3281	0.3478	7
F(L/I)GSF(L/I)YEYSR	2.8547	0.3212	0.3435	7
F(L/I)GSF(L/I)YEH*(L/I)*R	2.7544	0.3129	0.3498	7
F(L/I)GSF(L/I)YE(L/I)*H*R	2.7444	0.3120	0.3505	7
F(L/I)GSF(L/I)YYESR	2.6968	0.3081	0.3511	7
F(L/I)GSF(L/I)YYTDR	2.6391	0.3033	0.3542	7

experimental tandem mass spectrum. The sequence with the maximum cross-correlation score with the experimental spectrum is ST(L/I)PELYEK, which is the correct peptide. Thus, the correct N-terminal assignment (ST) is made.

DAFLGSFLYEYSR. The peptide DAFLGSFLYEYSR was selected because it illustrates the utility of the cross-correlation method for identifying a suboptimal candidate sequence as the correct peptide. The tandem mass spectrum for this peptide consists of 16,962 points and is shown in Figure 4. The preprocessing algorithm identifies the C-terminal amino acid to be arginine, assigned to the strong intensity peak at an *m*/*z* of 175.12, and also reports an abundant phenylalanine (F) immonium ion at 120.08 Th. The N-terminal pairs SV, DA, and GE exhibit comparable scores for this spectrum. However, no supporting $\mathbf{y}_{n-1}$-ions were found to distinguish the correct pair. Thus, the sequencing calculations are terminated at the $\mathbf{y}_{n-2}$-ion by replacing the "tail" boundary condition (Eq. 5) with the mass of this ion (1381.69 Da). The first-stage MILP consists of 592 variables and 291 constraint equations and is solved to optimality in a CPU time of 3.88 s. The rank-ordered list of candidate sequences is reported in Table 5.

As shown in Table 5, the correct peptide is ranked 5th with reference to the objective function value. However, all the candidate sequences of higher objective function value contain internal histidine (H) residues, which is not common for a doubly charged tryptic peptide. The C-terminal region of the candidate peptides reported in the stage 1 calculations varies from sequence to sequence, so it will be necessary to examine several distinct peptides for cross-correlation with the experimental spectrum. Notice that the first six N-terminal residues for all sequences are bold, indicating that this subsequence of the peptide is verified by the **b**-ion series and thus will be fixed in the second-stage calculations. The second-stage MILP contains 1937 variables and 307 constraint equations and ten candidate peptides are sequenced in a CPU of 5.60 s. The distinct peptides are shown in Table 6.

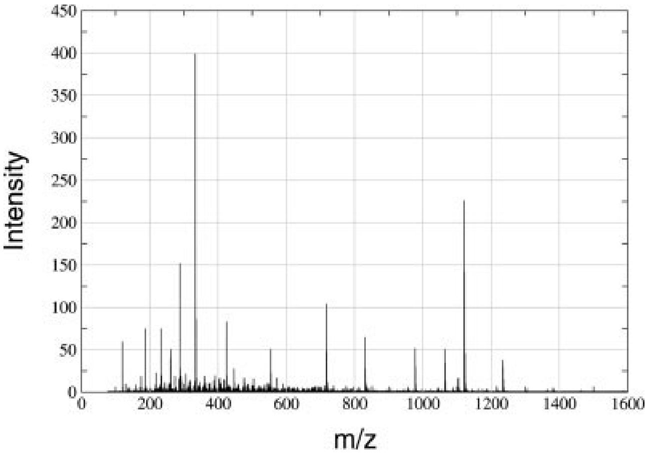

Figure 4. Tandem mass spectrum for DAFLGSFLYEYSR.

The candidate sequences from the second-stage solutions also exhibit variability in the C-terminal region of the peptide. The missing N-terminal weight of 185.99 Th, eliminated from the sequencing stages, corresponds to the amino acid combinations of DA, SV, or GE. Amino acid permutations consistent with the weights for all the distinct candidate peptides are substituted into the sequences and the theoretical spectrum for each is predicted and cross-correlated with the experimental tandem mass spectrum. The sequence with the maximum correlation score from this set is DAFLGSFLYEYSR, which is the correct peptide. Recall that in Table 5, this peptide is ranked 5th with reference to the objective function value. This example demonstrates the ability of the cross-correlation method to exploit spectral information that was not used in the stage 1 and stage 2 calculations to identify the correct peptide sequence.

Ion-trap MS/MS data

The peptides sequenced in this section were examined from a tryptic digest of *Homo sapien* proteins and the tandem mass spectrometry data are accessible on the Open Proteomics Database.[62] The data for the peptides were gathered using a ThermoFinnigan ESI–ion-trap mass analyzer, which characteristically has a low *m*/*z* cutoff value and has a mass resolution of

Table 6. Distinct Solutions from the Second-Stage MILP for DAFLGSFLYEYSR

Candidate Sequence	Obj	Avg	STD	Comp Ions
F(L/I)GSF(L/I)Y[194.06]ANR	2.8323	0.3289	0.3471	7
F(L/I)GSF(L/I)Y[208.10]*G*NR	2.8148	0.3275	0.3484	7
F(L/I)GSF(L/I)Y[265.12]NR	2.7847	0.3545	0.3520	7
F(L/I)GSF(L/I)[172.04]*QG*ANR	2.6501	0.2988	0.3475	7
F(L/I)GSF(L/I)[171.04]*EG*ANR	2.6501	0.2988	0.3475	7
F(L/I)GSF(L/I)[171.04]*SV*ANR	2.6351	0.2977	0.3484	7
F(L/I)GSF(L/I)[171.04]*GE*ANR	2.6351	0.2977	0.3484	7
F(L/I)GSF(L/I)[171.04]*DA*ANR	2.6351	0.2977	0.3484	7
F(L/I)GSF(L/I)[172.04]*NA*ANR	2.6351	0.2977	0.3484	7

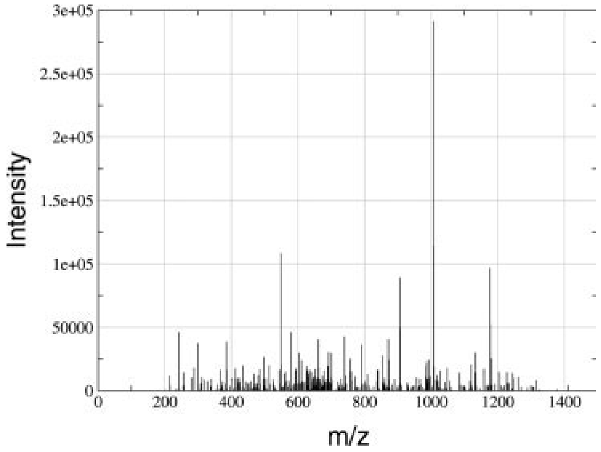

Figure 5. Tandem mass spectrum for LEGLTDEINFLR.

approximately unit resolution (that is, it has the ability to resolve different ions with *m*/*z* values that differ by 1 Dalton) throughout the *m*/*z* range.[38] The low *m*/*z* cutoff unfortunately prevents the identification of immonium ions, the C-terminal $\mathbf{y}_1$-ion, and various other ions necessary for identifying the complete set of boundary conditions for the problem formulation.

LEGLTDEINFLR. The tandem mass spectrum for this peptide consists of 345 mass peaks. As is typical of an ion-trap mass spectrometer, the low-mass region of the spectrum was cut off at an *m*/*z* of 200 Th, which in turn prevents the ability to identify the correct C-terminal amino acid or any immonium ions (see Figure 5). However, it is known that the peptide was digested with trypsin, which implies the C-terminal amino acid must be lysine or arginine. Through use of this information, the algorithm *assumes* the existence of both lysine and arginine as the $\mathbf{y}_1$-ion and then checks for their complementary $\mathbf{b}_{n-1}$-ions in the high-mass region of the spectrum. Those that are validated by their complementary $\mathbf{b}_{n-1}$-ion are added to the data set for further consideration. However, if *neither* is validated then both *m*/*z* values of 147.17 and 175.19 are added to the data file. The latter case is encountered in this example so both C-terminal amino acids are considered equally probable.

The preprocessing algorithm identifies the top-scoring N-terminal pairs as (Q/K)N, N(Q/K), E(L/I), and (L/I)E, which share the same $\mathbf{b}_2$-ion peak at an *m*/*z* of 243.10 Th. Neither of these pairs has a supporting $\mathbf{y}_{n-1}$-ion so the corresponding "tail" boundary condition for the **y**-ion series (Eq. 5) is replaced with the weight of the $\mathbf{y}_{n-2}$-ion (1177.56 Th). The resulting MILP consists of 435 variables and 282 constraint equations. Ten candidate peptides are generated in a CPU of 4.68 s and the statistically distinct sequences are shown in Table 7.

The peptide sequence of maximum objective function and maximum average of peak intensities, G(L/I)TDE(L/I)NF(L/I)R, contains a subsequence of five consecutive bold residues, indicating that these residues were confirmed by the **b**-ion series. The second-stage formulation contains 1961 variables and 298 constraint equations and ten candidate peptides were generated in a CPU of 7.97 s. The distinct solutions are reported in Table 8.

The N-terminal regions of the candidate peptides in Table 8 are considerably homologous to one another. However, the C-terminal region is quite variable from sequence to sequence and must be resolved with more detailed spectral information. Each of these candidate sequences is also missing the N-terminus amino acids, which for the weight of 242.44 Th could be combinations of N(Q/K), (I/L)E, GAN, or GG(Q/K). All permutations of amino acids consistent with the weights in the sequences in Tables 7 and 8 are substituted into the candidate peptides. The theoretical tandem mass spectrum for each sequence is predicted and cross-correlated with the experimental mass spectrum. The sequence with the maximum cross-correlation score is NQG(L/I)TDE(L/I)NF(L/I)R, which is the correct peptide except for the residue assignment of NQ at the N-terminus.

SQIHDIVLVGGSTR. This peptide was selected because it illustrates an instance where the preprocessing algorithm is unable to find *any* information regarding possible boundary conditions. The tandem mass spectrum for this peptide contains 351 data points and exhibits an *m*/*z* cutoff of 236 Th, which prevents the identification of the proper C-terminal amino acid (even when searching for a possible complementary $\mathbf{b}_{n-1}$-ion as a confirmation) (see Figure 6). Thus, the *m*/*z* values of 147.17 and 175.19 Th are added to the spectrum data to allow for the possibility of a C-terminal lysine or arginine, respectively. No potential $\mathbf{b}_2$-ions were identified in the data set, which is essential for verifying or adjusting the "tail" boundary conditions for the **y**-ion series.

When faced with the absence of an N-terminal amino acid pair, the algorithm automatically examines the high-mass region of the MS/MS data for prominent peaks that could possibly be acceptable "tail" boundary conditions for the **y**-ion series. Several peaks of high intensity are selected and a MILP is formulated and solved using each as the "tail" boundary

Table 7. Distinct Solutions from the First-Stage MILP for LEGLTDEINFLR

Candidate Sequence	Obj	Avg	STD	Comp Ions
G*(L/I)***TDE(L/I)N***F(L/I)*R	3.2322	0.3860	0.4065	10
G*(L/I)***TDQ***N***N***F(L/I)*R	3.1404	0.3790	0.4119	10
G*(L/I)***TD(L/I)***E***N***F(L/I)*R	3.1069	0.3762	0.4143	9
G*(L/I)***TY***PP***N***F(L/I)*R	3.0552	0.3690	0.4198	8
G*(L/I)***TD**VG*H(L/I)***P***(L/I)*R	3.0515	0.3376	0.4046	7
G*(L/I)***TD**VP*P(L/I)***P***(L/I)*R	3.0234	0.3380	0.4044	7
G*(L/I)***TD**VG*HP(L/I)(L/I)*R	2.9518	0.3293	0.4102	7
G*(L/I)***TD**R*H(L/I)***P***(L/I)*R	2.9469	0.3588	0.4174	7
G*(L/I)***TD**VP*PP(L/I)(L/I)*R	2.9237	0.3297	0.4100	7

Table 8. Distinct Solutions from the Second-Stage MILP for LEGLTDEINFLR

Candidate Sequence	Obj	Avg	STD	Comp Ions
G*(L/I)***TDE(L/I)N***[260.20]*R	3.1083	0.4184	0.4133	9
G*(L/I)***TDE(L/I)N***[288.22]*K	3.1083	0.4184	0.4133	8
G*(L/I)***TDE(L/I)**Y*[210.92]*R	3.1060	0.4173	0.4144	8
G*(L/I)***TDE(L/I)H***[265.11]*K	3.1046	0.4148	0.4168	8
G(L/I)TDE(L/I)NF[269.15]	3.0763	0.4146	0.4167	9
G(L/I)TDE(L/I)H*T*[292.42]	3.0343	0.4085	0.4225	8
G*(L/I)***TDE(L/I)[261.29]***(L/I)*R	2.8282	0.4177	0.4139	9
G*(L/I)***TDE(L/I)**[231.18]*[171.09]*K	2.7691	0.4146	0.4170	7
G(L/I)TDE(L/I)[261.29][269.15]	2.6723	0.4531	0.4227	8

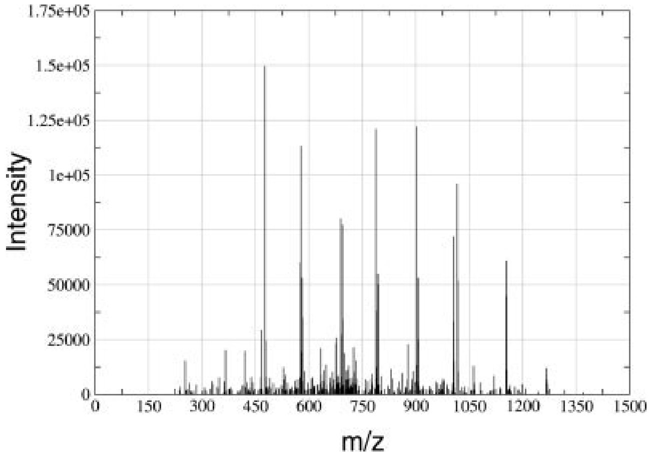

Figure 6. Tandem mass spectrum for SQIHDIVLVGGSTR.

condition. From these results, the peak that produces the optimal candidate sequence is selected and then used as the N-terminal boundary condition (Eq. 5) in the subsequent formulations. Eleven high-mass peaks were examined and their corresponding optimal solutions are shown in Table 9.

From Table 9, the optimal boundary condition corresponds to the mass peak of 1266.53 Th. The subsequent MILP consists of 2265 variables and 277 constraint equations using this mass peak as the "tail" boundary condition for the **y**-ion series. The ten candidate solutions are generated in 7.78 CPU seconds and the distinct sequences are shown in Table 10.

There is a high degree of homology among the sequences reported in Table 10. The two sequences of maximum objective function value and maximum standard deviation of peak intensities, (L/I)HD(L/I)V(L/I)VGG[216.34]K and (L/I)HD(L/I)V(L/I)VGG[188.32]R, also exhibit a subsequence of seven consecutive bold residues that were confirmed by the **b**-ion series. The only major difference between all the candidate peptides reported in Table 10 lies in the C-terminal region of the sequences, where the glycines (G) prevented complete fragmentation. The N-terminal weight of 215.11 Th corresponds to amino acid combinations of S(Q/K), TN, and GAS. Amino acids consistent with these weights are substituted into the distinct candidate sequences, whose theoretical spectrum is then predicted and cross-correlated with the experimental tandem mass spectrum. The sequence with the maximum cross-correlation score with the experimental spectrum is NT(L/I)HD(L/I)V(L/I)VGGSTR, which is correct except for the N-terminal assignment of NT. However, it should be noted that the correct sequence ranks a close second to the reported peptide. For this peptide, there was sufficient information in the tandem mass spectrum to resolve the C-terminal ambiguities, but the low mass cutoff of the spectrum prevented the correct N-terminal assignment.

Table 9. Results for Different N-Terminal Boundary Conditions (Pre Solve for SQIHDIVLVGGSTR)

m/z of Upper Bound	Maximum Obj
1266.53	5.7117
1153.63	5.4047
1119.59	5.3150
1081.16	3.0288
1062.25	3.7464
1016.55	4.7640
1005.61	2.8743
1003.51	3.9914
988.57	4.7640
976.75	3.7414
975.55	3.1712

Table 10. Distinct Solutions from the Second-Stage MILP for SQIHDIVLVGGSTR

Candidate Sequence	Obj	Avg	STD	Comp Ions
(L/I)H**D(L/I)V(L/I)VGG***[216.34]*K	5.7117	0.5733	0.3603	8
(L/I)H**D(L/I)V(L/I)VGG***[188.32]*R	5.7117	0.5733	0.3603	8
(L/I)H**D(L/I)V(L/I)VGGV**[245.73]	5.6398	0.5679	0.3685	8
(L/I)H**D(L/I)V(L/I)VG**[244.98]R	5.6220	0.6219	0.3341	7
(L/I)H**D(L/I)V(L/I)VG**[273.00]K	5.6207	0.6219	0.3341	7
(L/I)H**D(L/I)V(L/I)VE***[201.04]*K	5.6140	0.6146	0.3464	6
(L/I)H**D(L/I)V(L/I)V**N*[216.34]*K	5.5127	0.6134	0.3486	7
(L/I)H**D(L/I)VD*P*E***[201.04]*K	5.2650	0.5829	0.3821	5
(L/I)H**D(L/I)V(L/I)V**N*[188.32]*R	5.5127	0.6134	0.3486	7

Conclusions

A novel mixed-integer linear optimization framework was proposed for the de novo identification of peptides using tandem mass spectroscopy. For a specific MS/MS spectrum, the algorithm generates a rank-ordered list of potential candidate sequences and a cross-correlation technique is used to assess the degree of similarity between the theoretical tandem mass spectra of predicted sequences and experimental tandem mass spectra. The examples presented in this article demonstrate the reliability of the predictions of the proposed de novo algorithm for quadrupole time-of-flight and ion-trap mass analyzers. For situations where a spectrum does not contain sufficient information to unambiguously assign certain residues (for instance, in ion-trap mass analyzers where the data in the low-mass region are not recorded), the internal portions of the sequence are predicted correctly. This high degree of confidence for the predictions made by the de novo algorithm is important for the effective use of a homology search program to determine positive hits in a protein database because the slightest variability in a peptide sequence can lead to false protein matches. It is our experience from using database methods that if the peptide of interest is not in the database or the tryptic digestion was incomplete, then the highest scoring peptide reported by these methods can be completely incorrect in terms of residue accuracy. Our de novo method would be extremely useful for independently validating these database identifications because it provides consistent sequence accuracy and can elucidate regions of high confidence for the predicted candidate peptide. Furthermore, the use of several peptide identification methods, both de novo and database, could serve as a consensus-based approach to determine the identification of a peptide. The proposed de novo algorithm is not currently available for public use but we do plan to make it accessible to the scientific community as a Web-based tool.

Acknowledgments

The authors gratefully acknowledge financial support from the U.S. Environmental Protection Agency, EPA Grant R 832721-010; Siemens

Corporation; and the National Institutes of Health. Although the research described in the article has been funded wholly or in part by the U.S. EPA'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.

Literature Cited

1. Eng JK, McCormack AL, Yates JR. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom.* 1994;5:976–989.
2. Yates JR, Eng JK, McCormack AL, Schieltz D. Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal Chem.* 1995;67:1426–1436.
3. Yates JR, Eng JK, McCormack AL. Mining genomes: Correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. *Anal Chem.* 1995;67:3202–3210.
4. Perkins DN, Pappin DJC, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis.* 1999;20:3551–3567.
5. Pevzner PA, Mulyukov Z, Dancik V, Tang CL. Efficiency of database search for identification of mutated and modified proteins via mass spectrometry. *Genome Res.* 2001;11:290–299.
6. Bafna V, Edwards N. SCOPE: A probabilistic model for scoring tandem mass spectra against a peptide database. *Bioinformatics.* 2001; 17:S13–S21.
7. Sadygov RG, Yates JR. A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases. *Anal Chem.* 2003;75:3792–3798.
8. Hernandez P, Gras R, Frey J, Appel RD. Popitam: Towards new heuristic strategies to improve protein identification from tandem mass spectrometry data. *Proteomics.* 2003;3:870–878.
9. Havilio M, Haddad Y, Smilansky Z. Intensity-based statistical scorer for tandem mass spectrometry. *Anal Chem.* 2003;75:435–444.
10. Nesvizhskii AI, Aebersold R. Analysis, statistical validation and dissemination of large-scale proteomics datasets generated by tandem ms. *Drug Discovery Today.* 2004;9:173–181.
11. Steen H, Mann M. The abc's (and xyz's) of peptide sequencing. *Nat Rev Mol Cell Biol.* 2004;5:699–711.
12. MacCoss MJ, Wu CC, Yates JR. Probability-based validation of protein identifications using a modified SEQUEST algorithm. *Anal Chem.* 2003;74:5593–5599.
13. Moore RE, Young MK, Lee TD. Qscore: An algorithm for evaluating SEQUEST database search results. *J Am Soc Mass Spectrom.* 2002;13:378–386.
14. Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by ms/ms and database search. *Anal Chem.* 2002;74:5383–5392.
15. Fenyo D, Beavis RC. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal Chem.* 2003;75:768–774.
16. Anderson DC, Li W, Payan DG, Noble WS. A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: Support vector machine classification of peptide ms/ms spectra and sequest scores. *J Proteome Res.* 2003;2:137–146.
17. Taylor JA, Johnson RS. Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom.* 1997;11:1067–1075.
18. Dancik V, Addona TA, Clauser KR, Vath JE, Pevzner PA. De novo peptide sequencing via tandem mass spectrometry. *J Comp Biol.* 1999;6:327–342.
19. Fernandez de Cossio J, Gonzalez J, Satomi Y, Shima T, Okumura N, Besada V, Betancourt L, Padron G, Shimonishi Y, Takao T. Automated interpretation of low-energy collision-induced dissociation spectra by seqms, a software aid for de novo sequencing by tandem mass spectrometry. *Electrophoresis.* 2000;21:1694–1699.
20. Chen T, Kao MY, Tepel M, Rush J, Church GM. A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry. *J Comp Biol.* 2001;10:325–337.
21. Taylor JA, Johnson RS. Implementation and uses of automated de novo peptide sequencing by tandem mass spectrometry. *Anal Chem.* 2001;73:2594–2604.
22. Lubeck O, Sewell C, Gu S, Chen X, Cai DM. New computational approaches for de novo peptide sequencing from ms/ms experiments. *Proc IEEE.* 2002;90:1868–1874.
23. Jarman KD, Cannon WR, Jarman KH, Heredia-Langner A. A model of random sequences for de novo peptide sequencing. Proceedings of the 3rd IEEE International Symposium on BioInformatics and BioEngineering (BIBE 2003), Bethesda, MD, Mar. 10–12; 2003:206–213.
24. Cannon WR, Jarman KD. Improved peptide sequencing using isotope information inherent in tandem mass spectra. *Rapid Commun Mass Spectrom.* 2003;17:1793–1801.
25. Chen T, Bingwen L. A suboptimal algorithm for de novo peptide sequencing via tandem mass spectrometry. *J Comp Biol.* 2003;10: 1–12.
26. Frank A, Pevzner P. PepNovo: De novo peptide sequencing via probabilistic network modeling. *Anal Chem.* 2005;77:964–973.
27. Bern M, Goldberg D. De novo analysis of peptide tandem mass spectra by spectral graph partitioning. *J Comp Biol.* 2006;13:364–378.
28. Ma B, Zhang KZ, Hendrie C, Liang C, Li M, Doherty-Kirby A, Lajoie G. PEAKS: Powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom.* 2003; 17:2337–2342.
29. Zhang Z. Prediction of low-energy collision-induced dissociation spectra of peptides. *Anal Chem.* 2004;76:3908–3922.
30. Zhang Z. De novo peptide sequencing based on a divide-and-conquer algorithm and peptide tandem spectrum simulation. *Anal Chem.* 2004;76:6374–6383.
31. Heredia-Langner A, Cannon WR, Jarman KD, Jarman KH. Sequence optimization as an alternative to de novo analysis of tandem mass spectrometry data. *Bioinformatics.* 2004;20:2296–2304.
32. Malard JM, Heredia-Langner A, Baxter DJ, Jarman KH, Cannon WR. Constrained de novo peptide identification via multi-objective optimization. Proceedings of the 3rd International Workshop on High Performance Computational Biology (HiCOMB 2004). Santa Fe, NM, Apr. 26; 2004.
33. Malard JM, Heredia-Langner A, Cannon WR, Mooney R, Baxter DJ. Peptide identification via constrained multi-objective optimization: Pareto-based genetic algorithms. *Concurr Comput Pract Exp.* 2005; 17:1–18.
34. Fischer B, Roth V, Roos F, Grossmann J, Baginsky S, Widmayer P, Gruissem W, Buhmann JM. NovoHMM: A hidden Markov model for de novo peptide sequencing. *Anal Chem.* 2005;77:7265–7273.
35. Mann M, Wilm M. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal Chem.* 1994;66:4390–4399.
36. Tabb DL, Saraf A, Yates JR. GutenTag: High-throughput sequence tagging via an empirically derived fragmentation model. *Anal Chem.* 2003;75:6415–6421.
37. Sunyaev S, Liska AJ, Golod A, Shevchenko A, Shevchenko A. MultiTag: Multiple error-tolerant sequence tag search for the sequence-similarity identification of proteins by mass spectrometry. *Anal Chem.* 2003;75:1307–1315.
38. Kinter M, Sherman NE. Protein Sequencing and Identification Using Tandem Mass Spectrometry. New York:Wiley; 2000.
39. Schrijver A. Theory of Linear and Integer Programming. New York: Wiley; 1986.
40. Floudas CA, Grossmann IE. Synthesis of flexible heat exchanger networks for multiperiod operation. *Comput Chem Eng.* 1986;10:153–168.
41. Floudas CA, Grossmann IE. Synthesis of flexible heat exchanger networks with uncertain flow rates and temperatures. *Comput Chem Eng.* 1987;11:319–336.
42. Floudas CA, Paules GE IV. A mixed-integer nonlinear programming formulation for the synthesis of heat-integrated distillation sequences. *Comput Chem Eng.* 1988;12:531–546.
43. Floudas CA, Anastasiadis SH. Synthesis of general distillation sequences with several multicomponent feeds and products. *Chem Eng Sci.* 1988;43:2407–2419.
44. Paules GE IV, Floudas CA. APROS: Algorithmic development methodology for discrete-continuous optimization problems. *Oper Res J.* 1989;37:902–915.
45. Floudas CA, Ciric AR. Strategies for overcoming uncertainties in heat exchanger network synthesis. *Comput Chem Eng.* 1989;13:1133–1152.
46. Ciric AR, Floudas CA. A retrofit approach of heat exchanger networks. *Comput Chem Eng.* 1989;13:703–715.

47. Ciric AR, Floudas CA. A mixed-integer nonlinear programming model for retrofitting heat exchanger networks. *Ind Eng Chem Res*. 1990;29:239–251.
48. Ciric AR, Floudas CA. Application of the simultaneous match-network optimization approach to the pseudo-pinch problem. *Comput Chem Eng*. 1990;14:241–250.
49. Kokossis AC, Floudas CA. Optimization of complex reactor networks—I: Isothermal operation. *Chem Eng Sci*. 1990;43:595–614.
50. Aggarwal A, Floudas CA. Synthesis of general separation sequences—Nonsharp separations. *Comput Chem Eng*. 1990;14:631–653.
51. Ciric AR, Floudas CA. Heat exchanger network synthesis without decomposition. *Comput Chem Eng*. 1991;15:385–396.
52. Kokossis AC, Floudas CA. Synthesis of isothermal reactor-separator-recycle systems. *Chem Eng Sci*. 1991;46:1361–1383.
53. Kokossis AC, Floudas CA. Optimization of complex reactor networks—II: Nonisothermal operation. *Chem Eng Sci*. 1994;49:1037–1051.
54. CPLEX, ILOG Inc. *ILOG CPLEX 9.0 User's Manual*. Mountain View, CA:ILOG; 2005.
55. Floudas CA. *Nonlinear and Mixed-Integer Optimization*. Oxford, UK:Oxford Univ. Press; 1995.
56. Papayannopoulos IA. The interpretation of collision-induced dissociation tandem mass spectra of peptides. *Mass Spectrom Rev*. 1995;14:49–73.
57. Tabb DL, Smith LL, Breci LA, Wysocki VH, Lin D, Yates JR. Statistical characterization of ion trap tandem mass spectra from doubly charged tryptic peptides. *Anal Chem*. 2003;75:1155–1163.
58. Yalcin T, Csizmadia IG, Peterson MR, Harrison AG. The structure and fragmentation of B_n ($n \geq 3$) ions in peptide spectra. *J Am Soc Mass Spectrom*. 1996;7:233–242.
59. Burlet O, Yang CY, Gaskell SJ. Influence of cysteine to cysteic acid oxidation on the collision-activated decomposition of protonated peptides: Evidence for intraionic interactions. *J Am Soc Mass Spectrom*. 1992;3:337–344.
60. Hunt DF, Yates JR, Shabanowitz J, Winston S, Hauer CR. Protein sequencing by tandem mass spectrometry. *Proc Natl Acad Sci USA*. 1986;83:6233–6237.
61. Bean MF, Carr SA. Tandem mass spectrometry of peptides using hybrid and four-sector instruments: A comparative study. *Anal Chem*. 1991;63:1473–1481.
62. The Open Proteomics Database (OPD). Austin, TX:University of Texas Bioinformatics, Proteomics, and Functional Genomics. Available at http://bioinformatics.icmb.utexas.edu/OPD/. Accessed December 4, 2005.

Manuscript received Aug. 15, 2006, and revision received Oct. 12, 2006.