

New Computational Approaches for *de Novo* Peptide Sequencing From MS/MS Experiments

OLAF LUBECK, CHRISTOPHER SEWELL, SHENG GU, XIAN CHEN, AND D. MICHAEL CAI

Invited Paper

We describe computational methods to solve the problem of identifying novel proteins from tandem mass spectrometry (tandem MS or MS/MS) data and introduce new approaches that will give more accurate solutions. These new approaches integrate chemical information and knowledge into a graph-theoretic framework. Two sources of chemical information that we investigate are mass tagging and dissociation chemistry in the tandem MS process itself. We describe machine learning techniques that are used to classify peaks according to ion types based on known dissociation chemistry. We describe the algorithms that are implemented in a software code called PepSUMS. Using PepSUMS, we give results on the effectiveness of the new methods on the ultimate goal of improved protein identification.

Keywords—Computational biology, *de novo* sequencing, dissociation chemistry, machine learning, mass spectrometry (MS), PepSUMS, peptide sequencing, protein identification, proteomics, tandem mass spectrometry (tandem MS or MS/MS).

I. INTRODUCTION

Although large quantities of deoxyribonucleic acid (DNA) have been sequenced, cataloged, and annotated, we have come to the realization that this information is not sufficient to infer biological function. Informatic advances in decoding genomic sequences have not progressed enough to accurately predict protein products from genomic data [1], [2]. Moreover, modifications of proteins after transcription and translation cannot be deduced from gene sequence. These modifications add chemical state to the basic protein sequence and are crucial to activation of function and to cell signaling. In addition, proteins form complexes and interact in complicated networks [3]. As a result, proteomics, the large-scale analysis of proteins, is a growing research area and will be crucial to our understanding in the postgenomic era. Unambiguous protein identification forms the basis of

proteomic studies. Pinpointing biological targets of human diseases for subsequent intervention by pharmaceuticals invariably requires the identification of unknown proteins, as does, for example, an investigation into the mechanisms that microbial pathogens use to infect human cells.

An important breakthrough tool for proteomic studies has been protein identification by mass spectrometry (MS) [4] because the method is suitable to perform sequence deduction based on very small molecular quantities. This has been achieved, in part, by improvements in ionization, separation, and other experimental techniques. However, the attainment of a high-throughput genomewide identification is lacking computational methods to systematically interpret the data. The primary information coming from these mass analysis techniques is the protein sequence and its post-translational modifications. There are two methods of determining the sequence of proteins. The first is an approach that correlates known proteins (from a sequence database) with the measured MS spectrum [5]. The second is a *de novo* interpretation of the data that is capable of sequencing unknown (novel) proteins and their modifications. The statistical methods for the database lookup approach are well in hand, and this method is widely used in practice. However, this approach is obviously limited, because it finds only matches with already sequenced proteins. It can be used, for example, to determine the composition of *common* proteins. However, new highly sensitive separation techniques will increasingly allow for improved capture and measurement of low-abundance proteins, and these are likely to be novel. These recent technology advances motivate new complementary computational methods to analyze the MS data. A biologist who attempts to discover the basic sequence of a *new* protein and/or its current modification state based on MS data needs *de novo* rather than just database matching algorithms [6].

The *de novo* peptide sequencing problem, then, is to derive the sequence and post-translational modifications of the peptides *directly* from the masses of their fragments. A

Manuscript received April 30, 2002; revised September 8, 2002.

O. Lubeck, C. Sewell, S. Gu, and X. Chen are with the Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM 87545 USA.

D. M. Cai is with the NIS Division, Los Alamos National Laboratory, Los Alamos, NM 87545 USA.

Digital Object Identifier 10.1109/JPROC.2002.805301

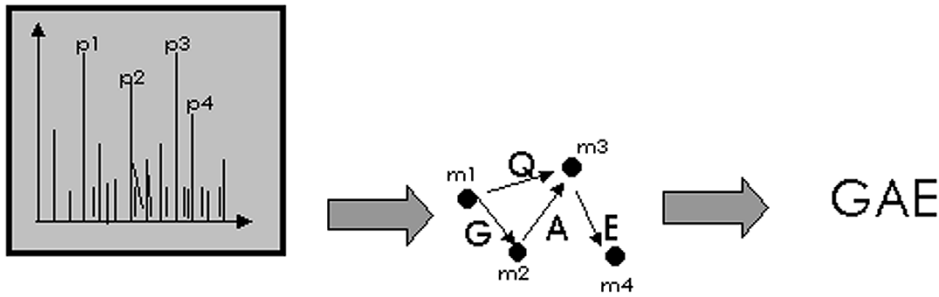


Fig. 1. Construction of spectrum graph.

tandem mass spectrometer is capable of ionizing a mixture of peptides and measuring their respective parent mass/charge ratios, recursively dissociating each peptide into more fragments and subsequently measuring the mass/charge ratios of these ionized fragments [tandem mass spectrometry (tandem MS; also known as MS/MS)]. In an ideal fragmentation process with an ideal instrument, the sequence of a peptide could be determined by matching the mass difference of two consecutive ions with the mass of an amino acid. This ideal experiment would require that an instance of each peptide be cleaved into two pieces exactly at the bond of two amino acids, that a single charge is deposited consistently on the same terminus of the peptide, and that instances of all these partial peptides exist and are measured.

In reality, fragmentation processes are far from ideal, and *de novo* protein sequencing remains an open yet crucial question for high-throughput protein identification. In this paper, we describe the major issues that must be addressed and a new approach that combines novel applications of machine learning techniques with chemical information that can make a difference in solving these issues.

II. BACKGROUND AND NEW APPROACH

Dancik *et al.* [6] have characterized previous attempts to develop *de novo* algorithms for solving the sequencing problem as employing either global or local search paradigms [6]. A global approach involves the generation of all amino acid sequences and the calculation of all theoretically possible fragment masses for each sequence. The goal is to find a sequence with the best match between the theoretical and experimental spectra. Different pruning methods were used to limit the exponential growth of sequences with peptide length. These global methods have largely been abandoned because pruning methods frequently discarded the solution and MS data were used only *after* the theoretical sequences were generated.

Local approaches tend to be more efficient because they use spectral data *before* any candidate sequence is evaluated. Different local strategies have employed branch and bound, dynamic programming, and graph-theoretical algorithms. Most recently, a group at Harvard Medical Center has reported on a dynamic programming approach [7]; a pharmaceutical and university consortium have

shown some progress with graph-theoretical approaches [6]; and a Web-available implementation (SEQMS) has been described [8]. However, recent work by us comparing methods indicates that the solution quality does not differ significantly among these techniques. The problem lies not in the choice of basic algorithmic approach but in the identification and classification of mass peaks in the spectrum. This aspect of the problem has not been the focus of previous work but is the key to an automated, high-throughput *de novo* computational interpretation. In the next section, we lay the framework for understanding why and the research approach that we are taking. We do employ a graph-theoretical organization, but our main research results affecting ion identification can be employed in any of the basic algorithmic approaches.

A. Graph-Theoretical Organization

In a graph-theoretical approach, spectral peaks are transformed to a directed acyclic graph (spectrum graph).

The vertices of the graph correspond to mass peaks in the spectral data where the mass value (x axis) is the sum of the component amino acids (residues) for a partial peptide, and the relative intensity is the y axis value. Two partial peptide masses are connected with an edge if their mass difference corresponds to an amino acid. The edge is labeled with the residue and assigned a weight that is a function of, for example, the intensity of the peaks. The edge is *directed* because the peptide sequence must be ordered from the protein's N terminus to its C terminus, which is customary convention. Post-translation modifications can be accounted for by expanding the dictionary of amino acids to include the modification states. A spectrum S of a peptide P is called *complete* if S contains at least one ion for every fragmentation of P [6]. Given a complete spectrum, *de novo* sequencing is thus cast as a problem of finding the longest path in a directed acyclic graph. It is important to realize that the mass measurement of a spectral peak is not directly the partial peptide mass, but rather an ion of the partial peptide. It is necessary to convert the ion mass into the partial peptide mass during the construction of the spectrum graph (see Fig. 1).

In the simplest (but most probable) model, where a peptide can be cleaved only at a single amide bond (a partial peptide), two peaks resulting from the two partial peptides (one an N

terminal, the other a C terminal ion) typically appear in the spectrum S . However, a mass peak, p_i , in S could correspond to either an N terminus (b ion) or C terminus ion (y ion). Constructing the spectrum graph requires that the vertices represent masses corresponding to the same terminus. The mass difference between a C terminus ion and an N terminus ion could match an amino acid, resulting in the *aliasing* of an improper edge into the spectrum graph. Construction of a correct spectrum graph requires a classification of the mass peaks into their respective ion types. Assuming that ion identification is possible, y ions can be converted into their b ion equivalents, because the total mass of the peptide is known.

We have considered only the simplest model for peptide dissociation—breaks at a single amide bond. In practice, many different types of ions are present owing to several complications to the simple model that we have been considering.

- 1) Breaks can occur at bonds internal to an amino acid.
- 2) Multiple amide bond breaks can occur for a single peptide.
- 3) After dissociation, ions can interact chemically to form secondary ions—for example, a b ion can lose a carbon monoxide molecule to form an a ion.
- 4) Mixture of ion types is instrument and protocol dependent.

Each of these ions can be characterized by an integer value d_j , such that $p_i + d_j$ gives the partial peptide mass corresponding to spectral peak p_i . This mass becomes a vertex in the spectrum graph. Because of the considerations previously listed, there is a set $D = \{d_1, d_2, \dots, d_n\}$ of ion types. The most common ion types for an ion-trap mass spectrometer have been observed and identified elsewhere [6]. For example, high-occurrence N terminal ions include b , a , $b\text{-H}_2\text{O}$, and $b\text{-NH}_3$ corresponding to the set $D = \{-1, 27, 17, 16\}$. Current graph-theoretical methods construct the spectrum graph by creating n vertices corresponding to each of n possible ion types for each p_i in S , $p_i + d_j$, $j = 1, n$. The number of ion types that must be considered is typically $10 < n < 20$. Since the same partial peptide can create multiple ions, some vertices can be expected to align, and are consequently merged into one. Edge weights are calculated and longest path algorithms are employed to find the highest ranking path in the spectrum graph.

The major problem with these current approaches is that, in addition to the noise in the experimental data, computational noise is generated during the construction process. Of the n vertices created in the spectrum graph for each peak for each feasible ion type, only one of these vertices corresponds to the correct partial peptide mass while $n - 1$ vertices are spurious. Edge weight scoring schemes are employed only *after* the spurious vertices are introduced and after an exponential number of candidate paths is created from them. The major contribution of this paper is to introduce a different approach that reduces the number of candidate vertices and paths by discriminating among and predicting ion types *prior* to scoring. For example, a precise prediction of the ion type of a spectral peak would introduce only one vertex. A less

precise determination that gives the ion's terminus would eliminate about half of the vertices. In this paper, we investigate the effect of two methods that integrate chemical information about the ions: 1) mass tagging of amino acids with stable isotopes; and 2) machine learning techniques that capture ion correlation information occurring in the dissociation chemistry.

The first method requires additional experimental protocols; the second uses information that already exists in the tandem MS data. We have developed a software implementation called PepSUMS that contains known *de novo* methods, and have extended it to include the new approaches that are described in this paper. The Appendix contains a description of the software. We apply the methods in PepSUMS to data from a Finnegan LCQ Deca ion-trap mass spectrometer.

III. THE IMPACT OF INTEGRATING CHEMICAL INFORMATION

A. Mass Tagging—Discriminating Between N Terminus and C Terminus Ions

Stable isotope labeling adds an isotopic “mass tag” to each labeled peptide at specified amino acids. In this investigation, deuterium is substituted for hydrogen at the fourth and fifth carbon atom position in lysine, causing a four-Dalton (Da) mass change of the peptide. The cells are 50% labeled—natural lysine and deuterium-enriched lysine were at the same concentration in the minimal media. Lysine residues are located at the C terminal sites of proteolytic peptides, for many but not all proteolytic cleavages such as trypsin, and Lys-C digestions. This will introduce a 4-Da split pattern into the lysine-containing peptides (see Fig. 2).

We have 18 pairs of labeled and unlabeled MS spectra. The experimental conditions under which these were obtained have been reported previously. A 4-Da mass shift with respect to the corresponding unlabeled counterpart was observed for each fragment peak originating from the lys- d_4 -labeled parent peptide, indicating the presence of a lysine residue in these fragments. PepSUMS was extended to include a differential analysis of the pairs of labeled/unlabeled spectra. Any peak above background in the unlabeled data is attempted to be matched with peaks of similar intensity at +4 Da in the labeled data. Those that match are designated C terminus; those for which there is definitely no signal at +4 Da are designated N terminus. Others which are ambiguous are left undesignated.

Tables 1 and 2 show the effect of incorporating mass tagging data into PepSUMS. Table 1 shows the percentage of the 18 peptides whose correct protein sequences are ranked by PepSUMS in the top one, five, ten, etc., of the candidate sequences with and without the introduction of mass tagging. PepSUMS is twice as effective in ranking the correct sequence at the top with the incorporation of mass tagging information.

Another metric to consider is the ladder distance [6] between the correct peptide sequence and the top ranked sequence from PepSUMS. Table 2 shows these results. The ladder distance between two sequences is the number of differences (larger than a tolerance of 1.5 Da) in their partial

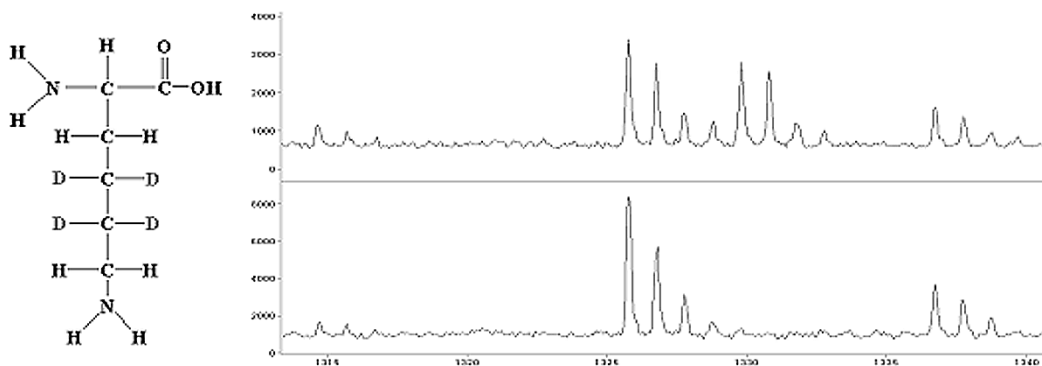


Fig. 2. Mass split pattern of 50% lysine-labeled peptides (upper 50% labeled, bottom unlabeled).

Table 1
Sequence Rank

	Tagged	Untagged
1	33	16
5	38	38
10	44	38
50	61	44
100	61	50
1000	61	55
100,000	66	66

Table 2
Ladder Distance

	Tagged	Untagged
0	33	16
1	38	22
2	44	27
3	44	27
4	44	27
5	50	27
6	50	27
7	50	33
8	55	38

sums. A ladder distance of zero is an exact match, and, for example, the sequences NDFEK and GGDFEK have a ladder distance of one because the residues GG have the same mass as *N*. Table 2 shows that the top ranked sequence is significantly closer to the correct one with mass tagging than without.

B. Ion Clustering and Classification—Integration Into the Spectrum Graph

We seek to develop an algorithm that will classify a mass peak in the spectrum S according to its ion type. We consider three classes: *b* ion, *y* ion, and other. We do this by estimating the probability of ion types (*b* and *y*) for each peak p_i in S . The process consists of two steps. First, we establish profiles of *b* and *y* ions based on tandem MS of *known* sequences. Then we use the established profile to classify the peaks of a tandem MS from *unknown* sequences into their respective ion classes. Fig. 2 shows how this process is accomplished.

We take hundreds of tandem MS measured from known sequences, and label each significant peak p_i as a *b* or *y* ion (unlabeled if neither). We construct a feature vector (profile) for each instance that contains correlation information of data in the vicinity of p_i . This feature vector contains chemical information that can discriminate between ion types. Based on known dissociation chemistry, we expect, for example, that the presence of two mass peaks differing by the mass of a CO molecule will be indicative of a *b* ion/*a* ion pair. However, the same dissociation will not happen for *y* ions. This difference rationalizes that *b* ions and *y* ions could have distinguishable feature vectors, which point to different directions in n -dimensional space. Although we have discussed only the correlation of *b* ions with *a* ions for the sake of brevity, there are many other ions that are present because of the dissociation chemistry. Each of these ions would be represented in the feature vector.

After establishing feature vectors for both ions, two primarily different approaches are applied to separate these ions in n -dimensional space, namely, unsupervised learning (clustering) and supervised learning (classification). The goal of clustering is to group together objects with similar properties. This can also be viewed as the reduction of the dimensionality or complexity of the system. The goal of supervised learning is to construct classifiers, using techniques such as linear discriminants, decision trees, or support vector machines, which assign predefined classes to a given ion profile—in our case, *b* or *y* ions. (See Fig. 3.)

Clustering and classification methods can have inaccuracies, and a subset of the ion types will be correctly predicted. The details of the machine learning algorithms are a sub-problem, and will be reported elsewhere. In this paper, we report on the larger picture—the effect that ion prediction accuracy ultimately has on protein identification. To assess this effect, we use our graph-theoretical *de novo* software, PepSUMS, to analyze data from known proteins where we incorporate varying levels of prediction accuracy into the spectrum graph construction. For each of 46 MS/MS peptide datasets, peaks are labeled as *b* or *y* ions. We then randomly pick a subset of the labeled peaks as being “predicted” accurately. All other peaks are treated as unknown.

The percentage of predicted peaks varies from 0 (no ion type prediction) to 100 (every *b* and/or *y* ion is classified cor-

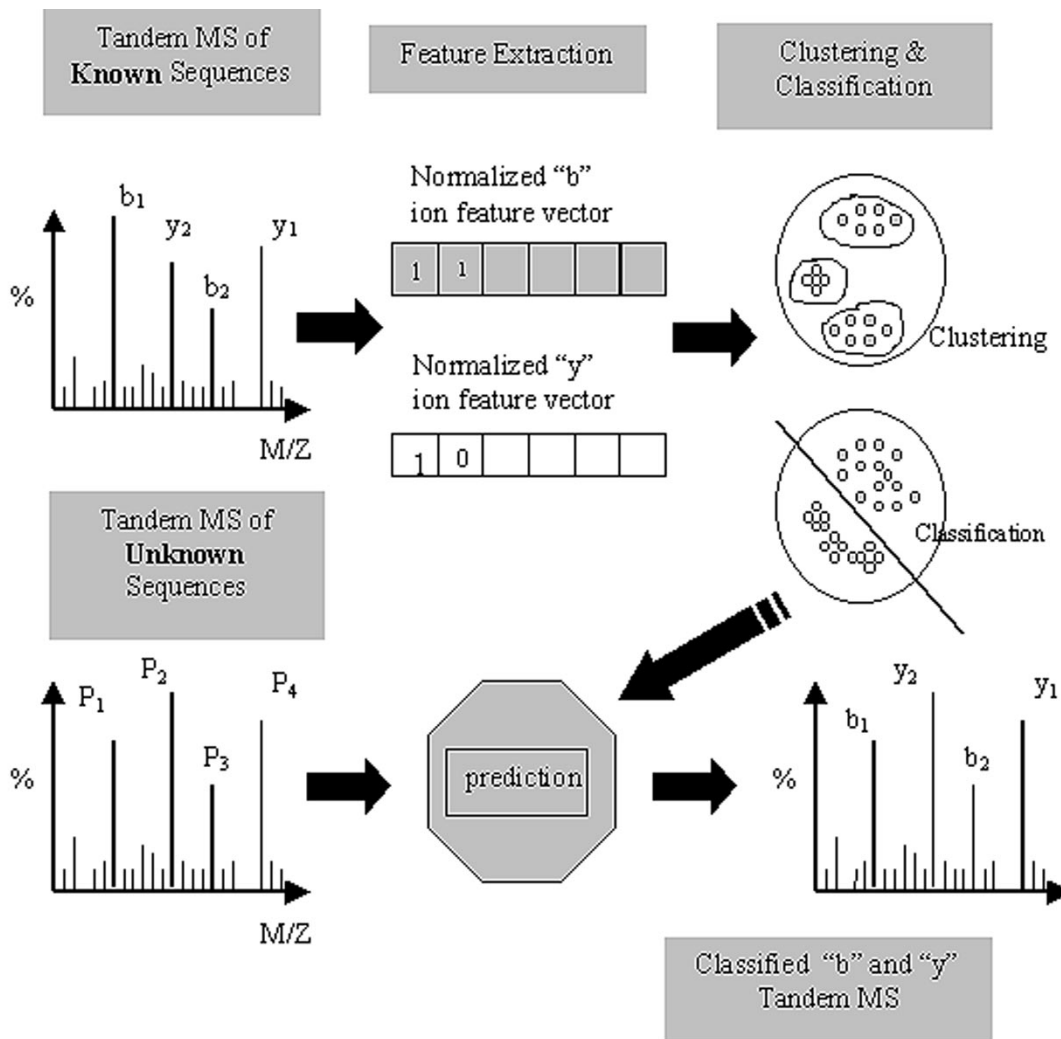


Fig. 3. Clustering and classifying “b” ion and “y” ions using machine learning approaches.

Table 3
Impact of B and Y Ion Prediction

	100	75	50	25	0
1	54	50	45	38	10
5	67	64	63	52	17
10	71	68	68	57	23
50	78	76	75	66	32
100	78	78	76	68	34
1000	86	83	80	73	43
100,000	86	86	86	80	67

Table 4
Impact of Y Ion Prediction

	100	75	50	25	0
1	39	37	34	26	10
5	54	53	49	41	17
10	56	56	56	46	23
50	69	64	63	55	32
100	69	67	65	59	34
1000	76	73	69	66	43
100,000	89	87	84	79	67

rectly). Results are summarized below in Tables 3 and 4. Separate cases of only b ion prediction and both b and y ion prediction are depicted. Tables 3 and 4 show the percentage of the 46 peptides whose correct protein sequences are ranked in the top one, five, ten, etc., by PepSUMS. For example, in Tables 3 and 4, one can see that if no prediction scheme is used, the correct peptide sequence is identified as the top candidate in only 10% of the cases and ranks in the top five in only 17% of the cases. Table 3 shows that if a conservative 25% of the b and y ions were predicted, this would have an almost fourfold increase in the number of correctly identified peptide sequences. This is a significant result and points

out that a relatively low number of predicted ion types can have a large impact on sequence accuracy. Note that a limit is reached where about 85% of the peptides can be sequenced at all in the case of complete knowledge of ion types. The remaining 15% do not have enough data to construct a complete spectrum graph.

As in the mass tagging results, we show ladder distances in Tables 5 and 6. The results show the large gain in sequence accuracy for a relatively small number of ion predictions. About 37% of all the peptide cases can be sequenced correctly if we can classify 25% of the b and y ions. This is a factor of four over no classification.

Table 5
Ladder Distance: *B* and *Y* Ion Prediction

	100	75	50	25	0
0	54	51	45	37	10
1	58	55	51	43	15
2	73	74	69	65	23
3	73	75	71	69	26
4	76	78	76	73	32
5	76	78	79	76	32
6	82	84	83	81	43
7	84	84	83	83	50
8	84	84	84	84	54

Table 6
Ladder Distance: *Y* Ion Prediction

	100	75	50	25	0
0	39	37	34	26	10
1	43	42	39	32	15
2	65	60	60	50	23
3	65	63	63	56	26
4	69	67	68	63	32
5	78	75	76	69	32
6	84	84	82	75	43
7	86	86	83	78	50
8	86	86	84	80	54

IV. SUMMARY, DISCUSSION, FUTURE WORK

It is clear that current *de novo* sequencing methods can benefit from the integration of additional chemical knowledge into the algorithms. We have investigated two ways that this can be accomplished and the effect that this additional information will have on the ultimate goal of producing an automated high-throughput sequencing capability using tandem MS. In the case of mass tagging, the information is gotten with additional “hard” experimental evidence. In the case of ion type classification, the information is already present in the MS spectrum in the form of peak correlation, but further computational analysis is needed. We approach the ion classification problem from the standpoint of machine learning techniques and discover that relatively conservative estimates of the accuracy of these techniques can significantly increase the precision of the putative peptide sequence. We have implemented a software code, PepSUMS, which serves as a vehicle for further research.

Further work will be directed toward increasingly better machine learning methods for ion classification. We intend to combine the mass tagging data with machine learning methods. PepSUMS can also be used to direct mass tagging experiments by pointing out areas where the putative sequence has low confidence. We also intend to apply PepSUMS to data from instruments that have more accurate mass measurements.

APPENDIX SOFTWARE IMPLEMENTATION

PepSUMS is a C++ implementation of our *de novo* sequencing algorithm. The program reads masses and abundances of spectral peaks from files in dta format. It considers

only a specified number of largest abundance peaks. The possibly inaccurate parent mass read from the file is corrected using a combinatorial algorithm that maximizes the number of peaks in common between the original spectrum and its complement [6]. It then builds a spectrum graph in which the vertices are the masses of the peptide fragments, obtained by adding the masses of ion types produced by the mass spectrometer to each large peak in the spectrum. If mass tagging is used, vertices are created only for ion types of the correct terminus. If ion prediction is used, a vertex is created only for the correct ion type. Boost Graph Library data structures and graph algorithms are used for the spectrum graph [10]. Vertices with similar masses are merged until no two vertices are closer in mass than a specified minimum distance. Edges are created between two vertices whose mass difference, or the mass difference between any two peaks in the merged vertices (“bridge” edges) is equal to the mass of an amino acid, within tolerance. The graph (if not too large) can be visualized using AT&T Graphviz. Each edge is assigned a weight by subtracting the sum of the probabilities of all absent ions for the target vertex of the edge from the sum of the probabilities of all present ions. If this result is positive, it is multiplied by the total abundance; if negative, it is divided by the total abundance. The ion probabilities are determined from mass spectrometry spectra of known sequences. The weight is greatly increased for an edge that connects to the parent mass with an arginine or lysine, since correct sequences of trypsin-digested peptides should end with one of these residues. If ion prediction is being used, edges leading to a vertex that includes a predicted peak are also greatly increased in weight to ensure that the correct path accounts for this fragment mass. The program outputs a requested number of longest paths from the vertex with zero mass to the vertex with the total mass of the peptide, using Eppstein’s *k* shortest paths algorithm and the Bellman–Ford algorithm [11]. The candidate solution sequences are the set of amino acids corresponding to the edges along these paths.

REFERENCES

- [1] A. Krogh, *Guide to Human Genome Computing*. San Diego, CA: Academic, 1998, pp. 261–274.
- [2] I. Dunham *et al.*, “The DNA sequence of human chromosome 22,” *Nature*, vol. 402, pp. 489–495, 1999.
- [3] D. Eisenberg *et al.*, “Protein function in the post-genomic era,” *Nature*, vol. 405, pp. 823–826, 2000.
- [4] A. Pandey *et al.*, “Proteomics to study genes and genomes,” *Nature*, vol. 405, pp. 837–406, 2000.
- [5] J. Eng *et al.*, “An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database,” *J. Amer. Soc. Mass Spectrom.*, vol. 5, pp. 976–989, 1994.
- [6] V. Dancik *et al.*, “De novo sequencing via tandem mass spectrometry,” *J. Comput. Biol.*, vol. 6, pp. 327–342, 1999.
- [7] T. Chen *et al.*, “A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry,” in *Proc. 10th SIAM Conf. Discrete Mathematics*, 2000.
- [8] J. Fernandez-de-Cossio *et al.*, “Automated interpretation of low-energy collision-induced dissociation spectra by SEQMS, a software aid for de novo sequencing by tandem mass spectrometry,” *Electrophoresis*, vol. 21, pp. 1694–1699, 2000.
- [9] T. Chen *et al.*, “A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry,” in *Proc. 10th SIAM Conf. Discrete Mathematics*, 2000.
- [10] Boost C++ libraries [Online]. Available: www.boost.org.

