Anal. Chem. **2005,** *77,* 7265−7273

NovoHMM: A Hidden Markov Model for de Novo Peptide Sequencing

Bernd Fischer,[†] Volker Roth,[†] Franz Roos,[§] Jonas Grossmann,[‡] Sacha Baginsky,[‡] Peter Widmayer,[§] Wilhelm Gruissem,[‡] and Joachim M. Buhmann[†]

Institute of Computational Science, Institute of Plant Sciences, and Institute of Theoretical Computer Science, ETH Zurich, Switzerland

De novo sequencing of peptides poses one of the most challenging tasks in data analysis for proteome research. In this paper, a generative hidden Markov model (HMM) of mass spectra for de novo peptide sequencing which constitutes a novel view on how to solve this problem in a Bayesian framework is proposed. Further extensions of the model structure to a graphical model and a factorial HMM to substantially improve the peptide identification results are demonstrated. Inference with the graphical model for de novo peptide sequencing estimates posterior probabilities for amino acids rather than scores for single symbols in the sequence. Our model outperforms state-of-the-art methods for de novo peptide sequencing on a large test set of spectra.

In the process of high-throughput protein identification, mass spectrometry has attained considerable importance during the most recent years.[1] Analysis based on mass spectrometry typically starts with a complex protein mixture which is fractionated by either gel electrophoresis or other fractionation methods to reduce the complexity of the sample. The proteins are then digested by a specific enzyme, such as trypsin. The resulting set of peptides is measured by a tandem mass spectrometer coupled with a high performance liquid chromatography device. In the first measurement stage of a tandem mass spectrometer, the total mass of the different peptides eluting at a certain time point from the column is determined. In the second stage, a subset of peptides of a certain small mass range is selected. These peptides are fragmented by low-energy collision with a noble gas. This fragmentation process finally yields the MS/MS spectra that ideally contain the masses of all N-terminal and C-terminal fragments of the designated peptides.

The inherently high noise level of mass spectra in high-throughput experiments strongly favors probabilistic models of the data generation process. In this paper, we propose NovoHMM, a novel and completely generative statistical model for mass spectra which arise from peptide sequences.[1] (A preliminary version of this model was presented at the NIPS conference.[2])

The data generation process is assumed to be adequately described by a hidden Markov model with hidden amino acid states along the mass axis of a spectrum and observable emissions of mass peaks for each state. The model is generative in the sense that typical spectra can be sampled from the estimated model, given a peptide sequence. The complete HMM produces computational costs which scale quadratically in the number of states. To reduce the computational and model complexity, the complete model is approximated by a factorial hidden Markov model. The factorial model implicitly estimates whether a certain peak has been generated by an N-terminal or a C-terminal fragment. An expectation-maximization algorithm is used to estimate these hidden variables. The complexity regularization mechanism by factorization leads to an improved predictive power of the model. As a particular advantage for inference, NovoHMM explicitly represents probabilistic estimates of the posterior, both for the whole sequence and for single amino acids.

In the Experimental Section, we compare NovoHMM with other de novo peptide sequencing methods, especially with PepNovo[3] which presumably is the best existing de novo sequencing method for doubly charged ion trap spectra. The experiments effectively show that our approach outperforms PepNovo and other competitors on the same benchmark dataset on which PepNovo was tested. In addition, we show that the model yields reliable estimates of the peptide's parent masses. The Bayesian modeling approach allows us to provide a fully probabilistic posterior estimate for each amino acid.

Related Work. The interpretation of the tandem MS spectra usually starts with a database search in cases for which a database of known proteins is available for the organism under investigation. The most popular database search tools are SEQUEST[4] and MASCOT.[5] In recent publications, probabilistic models for database searches have been reported, such as OLAV[6] and SCOPE.[7] Data analysts typically observe, however, that only a relatively

[†] Institute of Computational Science.

[‡] Institute of Plant Sciences.

[§] Institute of Theoretical Computer Science.

[1] Aebersold, R.; Mann, M. *Nature* **2003**, *422*, 198−207.

[2] Fischer, B.; Roth, V.; Buhmann, J. M.; Grossmann, J.; Baginsky, S.; Gruissem, W.; Roos, F.; Widmayer, P. A Hidden Markov Model for De Novo Peptide Sequencing. In *Neural Information Processing Systems*; MIT Press: Cambridge, MA, 2005, Vol 17.

[3] Frank, A.; Pevzner, P. *Anal. Chem.* **2005**, *77*, 964−973.

[4] Eng, J. K.; McCormack, A. L.; Yates, J., III. *Am. Soc. Mass Spectrom.* **1994**, *5*, 976−989.

[5] Hirosawa, M.; Hoshida, M.; Ishikawa, M.; Toya, T. *Comput. Appl. Biosci.* **1993**, *9*, 161−167.

[6] Colinge, J.; Masselot, A.; Giron, M.; Dessingy, T.; Magnin, J. *Proteomics* **2003**, *3*, 1454−1463.

[7] Bafna, V.; Edwards, N. *Bioinformatics* **2001**, *17*, S13−S21.

10.1021/ac0508853 CCC: $30.25 © 2005 American Chemical Society
Published on Web 10/11/2005

small fraction of all MS/MS spectra can be assigned to sequences from the database. Many of the spectra, for which the assignment procedure fails, do not contain enough fragment ions to infer the underlying peptide sequence. Large-scale quality assessment studies of mass spectra, on the other hand, show that a considerable fraction of failed assignments cannot be explained by low spectrum quality.[8] The main reason for the failure to identify these spectra is the lack of information in protein databases for splice variants, mutations, or posttranslational modifications. As a heuristic work-around, it has been proposed to start the analysis with database searches and subsequently increase the complexity of the search space.[9] The largest complexity of the search space is reached in the stage of de novo sequencing, in which a minimum of restrictions on the set of potential sequences is provided (e.g., by constraining the parent mass or by exploiting knowledge of tryptic digestion).

Lutefisk[10,11] and PEAKS[12] are two widely used de novo peptide sequencing methods. Lutefisk creates a N−C spectrum graph and searches for the best matching peptides using a simple scoring scheme. PEAKS creates a similar spectrum graph and generates a candidate list of peptides by searching the spectrum graph with a simple scoring scheme. PEAKS further refines this search by a modified score that takes into account the y-, x-, y − H_2O- and y − NH_3-ions. To obtain the sequences, a dynamic programming approach[13] has been developed that searches for the highest scoring antisymmetric path in the N−C spectrum graph. A first probabilistic scoring scheme for de novo sequencing has been presented by Dancik.[14] It simply estimates the fragmentation pattern at a small number of positions around the b- and y-ions. An improved scoring scheme[3] refines the noise model and uses a Bayesian network to model the fragmentation patterns.

Preliminaries on Tandem Mass Spectrometry. We consider double positively charged peptides which typically occur in the widely used electro-spray ionization. The total elementary mass of a peptide is

$$M = 1 + \sum_{i=1}^{n} \text{mass}\ (\alpha_i) + 17 + 2 \tag{1}$$

where the first term on the right side of the equation is the N terminus, the second is the amino acid residues, the third is the C terminus, and the last is the positive charge ($2H^+$).

The term *elementary mass* of a molecule refers to the number of protons and neutrons in the molecule. The exact mass of a peptide differs slightly from the elementary mass due to mass deficits from relativity theory. In a mass range up to 2500 Da, however, we can neglect this mass deficit. In the case of high-resolution instruments, such as FT-ICR (1 ppm precision), the mass deficit is clearly detectable. Data analysis of mass spectra with such a high resolution is beyond the scope of this paper.

The first MS measurement yields an estimate of the total peptide mass (parent mass). The difficulty in measuring this parent mass mainly arises from heavy ^{13}C isotopes, which yield different visible isotope peaks. Due to fluctuations and limitations in the resolution of the data, only rough estimates of the monoisotopic mass can be achieved with a resulting uncertainty of about ±1 Da. All masses used within this paper are monoisotopic masses.

In the first measurement scan, the peptides from a small mass window are selected and fragmented by collision with a noble gas. Most often, the peptide breaks at the peptide bond, yielding one N-terminal fragment and one C-terminal fragment. In most cases, both fragment ions are singly charged. An ideal MS/MS spectrum contains the masses of all fragments, yielding a list of N-terminal and C-terminal fragment masses. Deviations from this ideal case are caused by, for example, isotope shifts which result in problems in determination of the exact monoisotopic mass of fragments. Further complications originate from a neutral loss of water (H_2O), ammonia (NH_3) or other uncharged molecules in the collision process. The spectrum probably also contains a small amount of doubly charged ions. Moreover, spectrometers do not uniformly detect ions with the same sensitivity along the mass scale of the spectrum. Furthermore, mass spectrometry measurements are noisy.

THE HIDDEN MARKOV MODEL

A hidden Markov model is a statistical model describing sequential data with hidden information. For a general introduction to HMMs, the reader is referred to, for example, the textbook by Durbin et al.[15] HMMs are widely used in genome analysis, for example, for the detection of CpG islands. Markov models assume as their main modeling restriction that the probability of one state depends only on its predecessor. In a DNA model based on HMMs, for instance, the probability of observing a nucleotide A at position i depends only on the nucleotide at position $i - 1$ while being independent of the nucleotides at positions 1, ..., $i - 2$. In a HMM, we have to distinguish between hidden and observable random variables. In a Markov model for DNA sequences, the individual nucleotides are instances of an observable random variable. The hidden variable in the above CpG island example decides whether a nucleotide position belongs to a CpG island. In our NovoHMM model, the observable random variables correspond to the observed mass peaks, whereas the hidden variables represent the unknown underlying sequence.

We will derive a hidden Markov model that generates mass spectra as a finite automaton over states that correspond to masses. The elementary mass unit defines a natural granularity for the states. The elementary masses of fragment ions are clustered around centers with ~1.000 45-Da spacing. Thus, the mass axis of a spectrum is discretized in 1.000 45-Da steps. We will first assume that the exact parent mass is known. The problem how to estimate the parent mass will be discussed in the section titled Inferring the Sequence and Posterior Probabilities. The description of the model is divided in two parts. We first derive the transition probabilities between the individual model states

[8] Bern, M.; Goldberg, D.; McDonald, W. H.; Yates, J., III. *Bioinformatics* **2004**, *20*, i49−i54.

[9] Sadygov, R. G.; Cociorva, D.; Yates, J., III. *Nat. Methods* **2004**, *1*, 195−202.

[10] Taylor, J. A.; Johnson, R. S. *Rapid Commun. Mass Spectrom.* **1997**, *11*, 1067−1075.

[11] Taylor, J. A.; Johnson, R. S. *Anal. Chem.* **2001**, *73*, 2594−2604.

[12] Ma, B.; Zhang, K.; Hendrie, C.; Liang, C.; Li, M.; Doherty-Kirby, A.; Lajoie, G. *Rapid Commun. Mass Spectrom.* **2003**, *17*, 2337−2342.

[13] Chen, T.; Kao, M.-Y.; Tepel, M.; Rush, J.; Church, G. M. *J. Comput. Biol.* **2001**, *8*, 325−337.

[14] Dancik, V.; Addona, T. A.; Clauser, K. R.; Vath, J. E. *J. Comput. Biol.* **1999**, *6*, 327−342.

[15] Durbin, R.; Eddy, S. R.; Krogh, A.; Mitchison, G. *Biological Sequence Analysis*; Cambridge University Press: Cambridge, 1999.

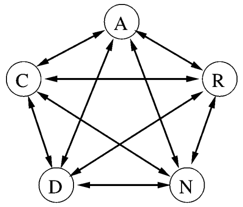

Figure 1. State transitions in the finite state automaton for amino acid sequences. The graph is drawn for five amino acids.

in the section titled Transition Probabilities. The second part concerns the emission probabilities which specify the process of generating peaks of certain heights, given the individual states (see Emission Probabilities (for N-Terminal Fragments)). Finally, we combine both sets of probabilities to a factorial hidden Markov model.

Transition Probabilities. The simplest model of an amino acid sequence is a list of random variables with 20 different states. Each random variable represents one amino acid in the sequence. The transitions are the conditional probabilities of observing a certain amino acid at position t, given the observation at position $t - 1$. Figure 1 shows the complete state transition graph for this model for only five amino acids. The graph is fully connected, since each amino acid can be followed by any other amino acid. Every state transition in this model corresponds to a one amino acid step.

A random variable in a one-dimensional Markov model depends only on the preceding variable in the sequence. Since the parent mass constraint has to be fulfilled, we have to know the total mass of all preceding amino acids. This information is only available in the model depicted in Figure 2, where we introduce a counter for the mass of each amino acid. The new model has, thus, a step size of one elementary mass unit.

The amino acid-based model is augmented by replicating each state a n_a times, where n_a is the elementary mass of the amino acid a. The transition graph of the augmented model is depicted by the solid lines in Figure 2. A missing edge in this graph denotes a vanishing transition probability. Once the first state of an amino acid has been selected, the transitions are constrained to all subsequent states of this amino acid. At the end state of an amino acid, a transition occurs according to the corresponding transition probabilities in the amino acid-based model from Figure 1. Each amino acid sequence corresponds to exactly one state sequence $s_1, ..., s_M$. The transition probabilities of such a state sequence are denoted by

$$a(s_m, s_{m-1}) = P(S_m = s_m | S_{m-1} = s_{m-1}) \tag{2}$$

The resulting model can represent amino acid sequences of arbitrary length with a state sequence of 1-Da step size. In tandem mass spectrometry, however, we actually have additional information in the form of the measured parent mass, M. After M steps in the Markov model we can easily check, if the parent mass constraint is fulfilled: the constraint is satisfied if we have reached a final state of an individual amino acid. According to the outcome of this check, a positive and a negative end state are added to the model. If the parent mass constraint is not fulfilled after M steps, the only possible transition is one to the negative end state. Thus, the transition probabilities change in a deterministic way after exactly M steps. The new transition probabilities are denoted a'. In Figure 2, the dotted arrows represent the transition probabilities at step M.

Thus far, we have modeled the prior distribution

$$P^{\text{prior}}(\text{peptide}) \tag{3}$$

$$= P^{\text{prior}}(S_1 = s_1, ..., S_M = s_M | S_{M+1} = s_+) \tag{4}$$

$$\propto \prod_{m=1}^{M} a(s_m, s_{m-1}) a'(s_M, s_{M+1}) \tag{5}$$

on all sequences of a certain parent mass. For any possible peptide, this distribution specifies the probability that the peptide occurs in the dataset, given a certain parent mass. It is 0 whenever the parent mass constraint is not fulfilled, and it is strictly positive otherwise. Since a' can be clearly identified by the occurrence of the index M in its argument, we will only refer to one transition probability a in the sequel (with a slight abuse of notation).

To estimate the transition probabilities, we implemented the usual maximum likelihood method which reduces to observing the frequencies of amino acids in the training dataset in this case.

Emission Probabilities (for N-Terminal Fragments). After specifying the transition probabilities in the automaton model, the emission process that finally generates the observable spectra with all its complexity has to be modeled. Again, we make a simplification: we assume that mass peaks are solely generated by N-terminal ions (a, b, b − H_2O, etc.). Within the state sequences corresponding to a single amino acid (due to the introduced mass counter, we will refer to these states as *counter states* in the sequel), each individual state has a certain function in the generation process. Figure 3 illustrates the different effects of the counter states. Assume the states of amino acid A are numbered from s_1 to s_{n_A}. The main peak (b-ion) is generated by the state s_{n_A}. The water loss (b − H_2O) is generated by the state s_{n_A-18}, the carbon monoxide loss (a-ion) is generated by the state s_{n_A-28}, and so on. A similar analysis is possible for the isotopes: the first isotope of the b-ion is generated by the first counter state of the consecutive amino acid. Thus, the state s_1 generates the first isotope shift of the b-ion of the previous amino acid. At most states, we do not expect any fragment ion. For example, it is not observed that the fragment ion loses an ion of mass 22 Da. These states, however, may generate noise peaks instead.

For any of these effects, we model an emission probability distribution of the corresponding peak height. The spectrum under investigation is denoted by $x_1, ..., x_M$. Thus, x_m is the peak height at mass m. These emission probabilities, e, are now defined as the probability to observe a peak with a certain peak intensity x_m given a state s_m, that is,

$$e_{sm}(x_m) = P(X_m = x_m | S_m = s_m) \tag{6}$$

For any effect (e.g., b-ion, y-ion, b − H_2O-ion, noise, etc.) we have a different emission distribution. In our experiments, we have first discretized the peak heights by introducing a set of equally populated bins for intensities. The number of bins was selected

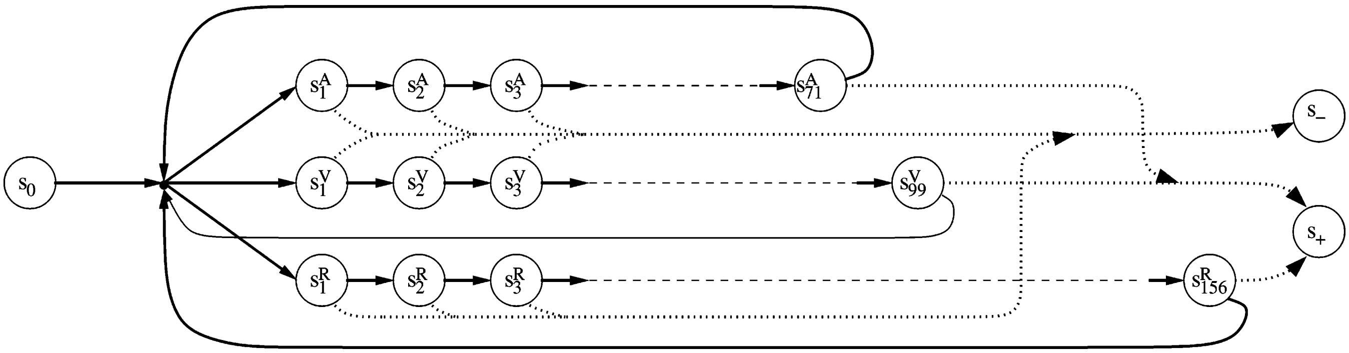

Figure 2. State transitions in the finite state automaton for tandem mass spectra. Solid arrows denote the possible transitions while generating a peptide. After *M* steps, the automaton is forced to take the positive or negative end state (dotted arrows).

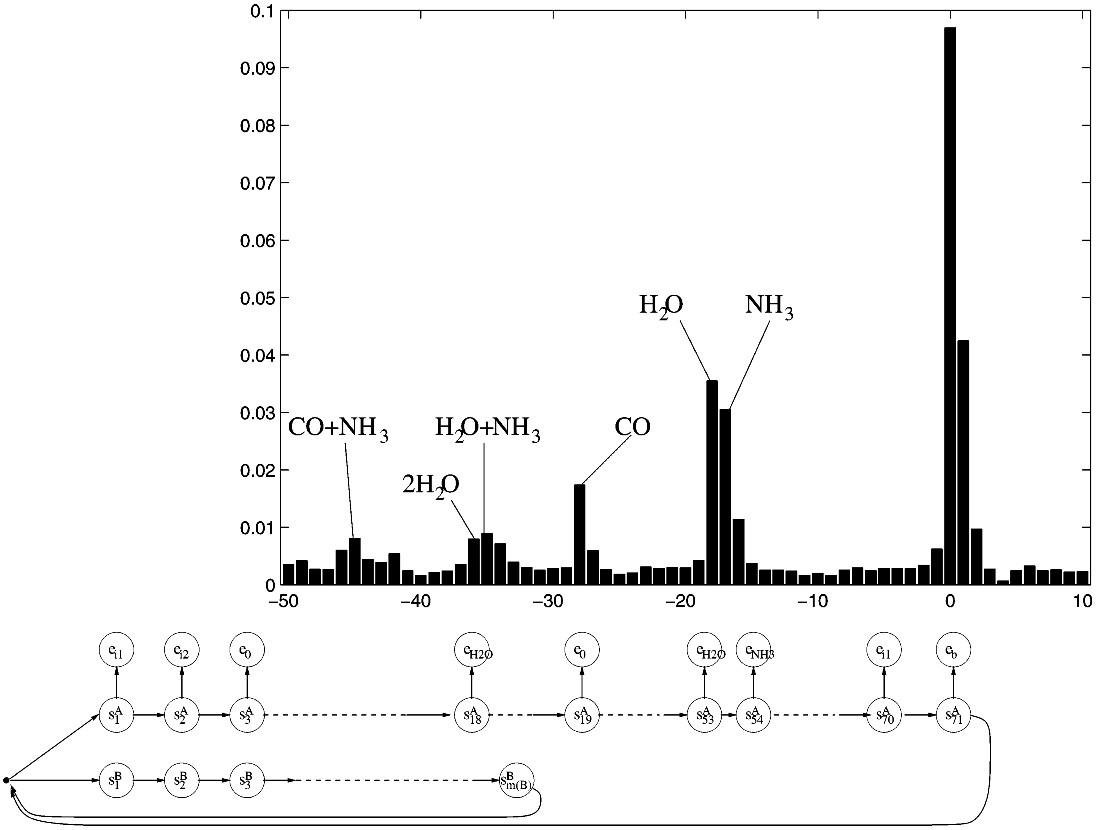

Figure 3. Emission probabilities for N-terminal fragment peaks are coupled with the counter states of the Markov model. Each counter state has a certain function, for example, the emission of an a-ion, a b-ion, a H_2O loss or the emission of a noise peak.

by cross-validation, with optimal prediction performance being obtained for five bins.

The joint probability distribution of sequence and spectrum is the basis of the subsequent inference method. For spectra that contain only N-terminal fragment ions, it is given by

$$P^N(\text{peptide, spectrum}) \tag{7}$$

$$= P^N(s, x) = \prod_{m=1}^{M} a(s_m, s_{m-1}) e_{s_m}(x_m) \tag{8}$$

An analogous model can be derived for the C-terminal fragments. The overall goal is to infer the peptide that most probably has generated the spectrum. In Inferring the Sequence and Posterior Probabilities, it is shown how the peptide is inferred given the spectrum by only using the above joint probability distribution.

The Factorial Hidden Markov Model. We now focus on the combination of the two models for N- and C-terminal fragments. The main aspect of a hidden Markov model is that the generation of an event (here a peak) depends only on the preceding event. In particular, an event cannot depend on the future. In a model that describes tandem mass spectrometry as a state sequence with increasing mass, however, such a situation occurs: the peak at a certain mass position cannot only be produced by the N-terminal fragment, but also by the C-terminal fragment. For example, if there is an N-terminal fragment ion in the high-mass region, then there should be evidence for an C-terminal fragment ion in the low-mass region. This means that the actual peak depends on a future event in the mass-ordered state model. This dependency structure is often the reason a complicated dynamic programming approach[13] is applied to search for antisymmetric paths. We propose to pursue a different strategy to overcome this problem.

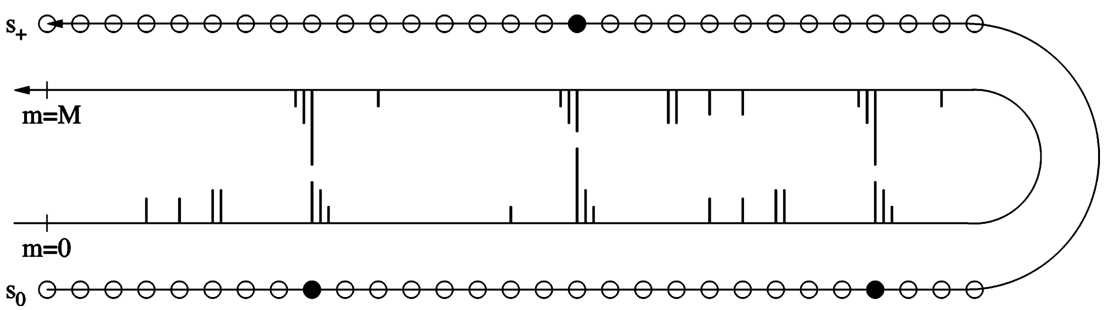

Figure 4. Folding the spectrum in the middle illustrates the internal mirror symmetry of the problem. The Markov chain models a sequence with four amino acids. The filled circles correspond to the amino acid boundaries. Around each amino acid boundary a peak pattern is generated, once for the N-terminal fragments and once for the C-terminal fragments.

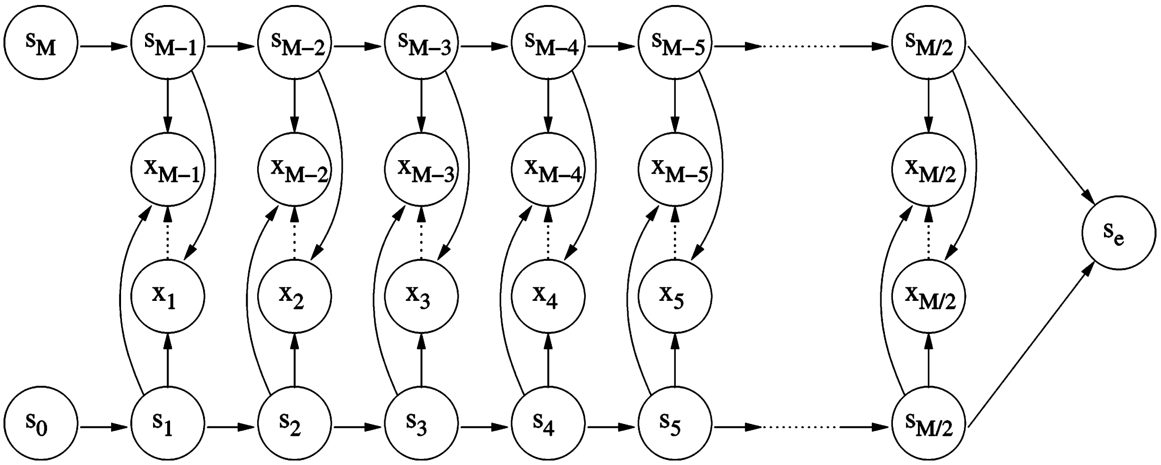

Figure 5. The dependency structure of the factorial hidden Markov model consists of two Markov chains, one for the first half of the peptide and one for the second half of the peptide. The emission variables depend on both Markov chains, thereby coupling them.

We divide the Markov chain into two Markov subchains: one to generate the sequences from low mass up to half the parent mass and another to generate the sequences from parent mass down to one-half the parent mass (see Figure 4). As a consequence, the generated spectrum is folded in the middle. The situation depicted in the figure represents a situation in which a peptide contains four amino acids. The black dots mark the amino acid boundaries. The first amino acid boundary (left lower black dot) generates an N-terminal fragment pattern in the low-mass region and an C-terminal fragment pattern in the high-mass region. Analogously, the third amino acid boundary generates an N-terminal fragment pattern in the high-mass region and an C-terminal fragment pattern in the low-mass region. The reader should notice that the main peaks which correspond to the y- and b-ions exhibit a clear mirror symmetry, whereas the other peaks break the symmetry.

The dependency structure of the graphical model corresponding to the above Markov subchain structure is drawn in Figure 5 by the solid arrows. One subchain models the low-mass amino acid sequence (states s_0, ..., $s_{M/2}$), whereas the other generates the high-mass amino acid sequence (states s_M, ..., $s_{M/2}$). The two Markov chains are coupled exclusively by the emission probabilities. Each emission variable (a peak x_m in the spectrum) depends on two states (s_m and s_{M-m}), because the peak can be generated either by an N-terminal ion or by an C-terminal ion. The emission probability in the factorial HMM is

$$e_{s_m,s_{M-m}}(x_m) = P(x_m|s_m, s_{M-m}) \tag{9}$$

The joint probability distribution of sequence and spectrum is given by

$$P(\text{peptide, spectrum}) = P(s, x) \tag{10}$$

$$= \prod_{m=1}^{M/2} a(s_m, s_{m-1}) a(s_{M-m}, s_{M-m-1}) e_{s_m,s_{M-m}}(x_m) e_{s_{M-m},s_m}(x_{M-m}) \tag{11}$$

This model combines a HMM for N-terminal fragment ions and a HMM for C-terminal fragment ions, where the emission symbols (the peaks) are shared by both models. Such a hidden Markov model in which the Markov chain factorizes in two (or more) Markov chains that are only coupled by the emission probabilities is called a factorial hidden Markov model.[16]

Frank and Pevzner[3] introduced a Bayesian network to represent the fragmentation pattern. Dependencies between emission probabilities can be incorporated into the factorial HMM, as well. The design of the structure of the network, however, poses a challenging task: one has to find a suitable tradeoff between a refined model of the dependency structure and overfitting problems induced by an increased model complexity. In practical applications with a limited number of training samples, the overfitting problem is of particular importance. It is, thus, not too surprising that in our experiments, a relatively simple network structure showed the best results: we model only the dependency

(16) Ghahramani, Z.; Jordan, M. I. *Mach. Learn.* **1997**, *29*, 245–273.

of the y-ion on the b-ion. All other dependencies, such as the dependency of the water loss on the b-ion, are discarded. Figure 5 depicts the complete model used in the experiments. The dependencies between b-ion and y-ion emissions are plotted as dotted arrows.

Approximation of the Model. Since the state space of the factorial hidden Markov model is squared in the number of states of the simple model, decoding becomes very time-consuming. Furthermore, the model complexity (measured in terms of the number of free parameters in the model) is extremely high, which imposes severe problems for a statistical inference process that is based on relatively small training samples. Sequencing would become much easier if we knew a priori whether a peak is generated from an N-terminal ion or an C-terminal ion. A second set of hidden variables is introduced to distinguish between N- and C-terminal fragment peaks. Let B_m denote a binary random variable that takes the value 1 if the peak is an N-terminal fragment peak and 0 otherwise. In the sequel, these variables will be called N/C-bits. The emission probabilities from the factorial model are replaced by a mixture model:

$$e_{s_m,s_{M-m}}(x_m) = b_m e_{s_m}(x_m) + (1 - b_m) e_{s_{M-m}}(x_m) \tag{12}$$

$$= e_{s_m}(x_m)^{b_m} \cdot e_{s_{M-m}}(x_m)^{1-b_m} \tag{13}$$

$$e_{s_{M-m},s_m}(x_{M-m}) = e_{s_{M-m}}(x_{M-m})^{b_{M-m}} \cdot e_{s_m}(x_{M-m})^{1-b_{M-m}} \tag{14}$$

The N/C-bit b_m indicates whether the peak is an N-terminal ion or an C-terminal ion and, thus, on which subchain the peak x_m depends. Note that the second equation holds, because the N/C-bits are either 0 or 1. A similar idea of separating the spectrum into N-terminal and C-terminal peaks has been described recently.[17] The joint probability distribution is now a probability of peptide, spectrum, and N/C-bits and can be written as

$$P(\text{peptide, spectrum, } N/C\text{-bits}) = P(s, x, b) \tag{15}$$

$$= \prod_{m=1}^{M/2} a(s_m, s_{m-1}) e_{s_m}(x_m)^{b_m} e_{s_m}(x_{M-m})^{1-b_{M-m}} \tag{16}$$

$$\prod_{m=1}^{M/2} a(s_{M-m}, s_{M-m-1}) e_{s_{M-m}}(x_m)^{1-b_m} e_{s_{M-m}}(x_{M-m})^{b_{M-m}} \tag{17}$$

The two sets of hidden variables (the peptides and the N/C-bits) are coupled. Since a priori knowledge about the origin of the peaks is hardly available in a de novo scenario, we propose to use the classical expectation-maximization (EM) algorithm[18] (also known as Baum–Welch training[19]) to estimate the hidden N/C-bit variables.

The expectation-maximization algorithm iterates two steps. In the E step, the expectation values of the hidden variables (peptide sequence and N/C-bits) are computed given the observed data (the training spectra) while keeping the model parameter fixed. In the M step, the maximum likelihood estimate of the model parameter is computed, given the training spectra and the expectation values of the sequences and N/C-bits. We decided to infer the sequence and the assignment variables by transductive inference: the sequence and the assignment variables are estimated within one EM loop involving both the training set and the new test spectrum. In the E step, the expectation over the joint probability distribution of all hidden variables (the sequence and the assignments) has to be computed for the new spectrum. Since this is computationally intractable, we decided to decouple both types of variables, leading to an iterative update scheme: if $P(s_m)$ is given, the expectations of the N/C-bits are estimated as

$$P(b_m) = \frac{\sum_{s_m} P(x_m|s_m)P(s_m)}{\sum_{s_m} P(x_m|s_m)P(s_m) + \sum_{s_{M-m}} P(x_m|s_{M-m})P(s_{M-m})} \tag{18}$$

Keeping the N/C-bits fixed, the probabilities $P(s_m)$ are recomputed by the forward–backward algorithm. The expectation-maximization algorithm replaces the binary variables, b_m, in the joint probability distribution by their expectation values that are equivalent to $P(b_m)$.

Figure 6 shows the estimation of the N/C-bits for one example spectrum. The original spectrum (top panel) is divided into a spectrum of N-terminal fragment peaks (middle panel) and C-terminal fragment peaks (bottom panel). Both the N-terminal and the C-terminal spectra are derived by multiplying the original spectrum by the estimation of the N/C-bits. It can be seen that the N-terminal fragment peaks are clearly separated from the C-terminal fragment peaks. Note that some peaks can be explained as neutral loss of b-ions and as y-ions.

Inferring the Sequence and Posterior Probabilities. Since the measurement of the total peptide mass typically differs from the true parent mass by up to ±1 Da, the hidden Markov model first is used to obtain a precise estimate of the parent mass. The maximum likelihood estimate of the parent mass is given by

$$\hat{M} = \operatorname*{argmax}_M P\{x|s_+, M\} \tag{19}$$

$$= \operatorname*{argmax}_M \sum_s P\{x, s|s_+, M\} \tag{20}$$

where the parent mass M is now considered as a random variable. The sum over all sequences can be computed efficiently by dynamic programming using the forward or backward algorithm. In our experiments, we tested the given elementary masses and one mass unit more or less. The parent mass is estimated using the model without the grouping variables, b, because the computation of the corresponding estimates in the model with grouping would require an integration over all b variables, which is computationally intractable.

The next step is the decoding, that is, the computation of the sequence that best matches the spectrum given the parent mass. The maximum posterior estimate of the sequence is

$$s^* = \operatorname*{argmax}_s P\{s|x, s_+, \hat{M}\} = \operatorname*{argmax}_s P\{s, x|s_+, \hat{M}\} \tag{21}$$

[17] Bern, M.; Goldberg, D. EigenMS: De Novo Analysis of Peptide Tandem Mass Spectra by Spectral Graph Partitioning. In *LNCS*; Springer: Berlin, 2005; Vol. 3500.

[18] Dempster, A. P.; Laird, N. M.; Rubin, D. B. *J. R. Stat. Soc., Series B* **1977**, *39*, 1–38.

[19] Baum, L. E. *Inequalities* **1972**, *3*, 1–8.

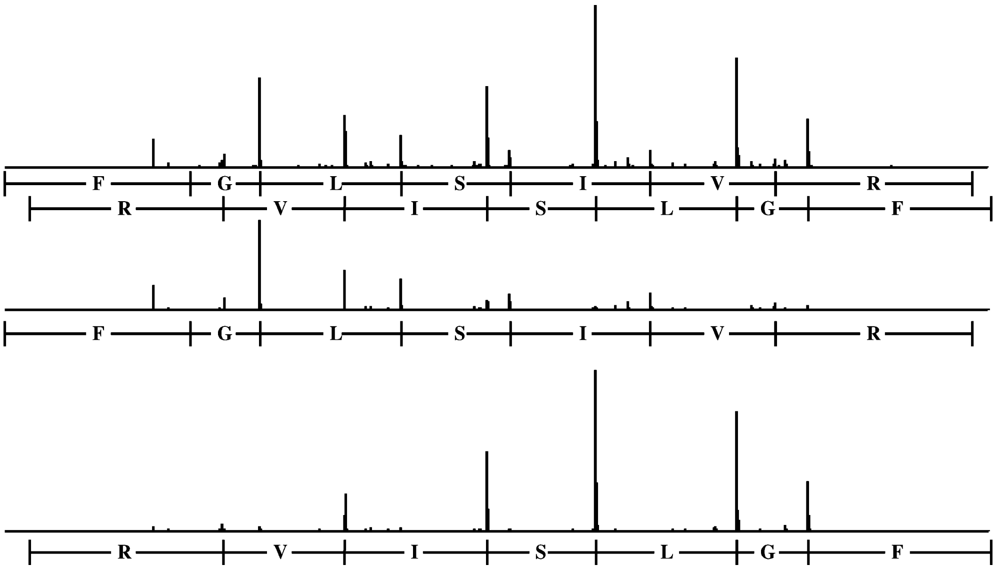

Figure 6. Example of spectrum partitioning. The first image shows the original spectrum. It is partitioned into two spectra, the spectrum of N-terminal fragments (middle) and that of C-terminal fragments (bottom).

The best sequence can efficiently be found by the Viterbi algorithm.[20] The Viterbi algorithm can be applied to both models (with and without grouping).

The probabilistic model, on the other hand, provides more information than just the best sequence. In particular, it enables the user to compute posterior probabilities of observing the predicted sequences. The posterior of the whole sequence is

$$P(s|x, s_+, \hat{M}) = \frac{P\{s, x|s_+, \hat{M}\}}{P\{x|s_+, \hat{M}\}} \qquad (22)$$

The denominator, also known as *evidence*, can be computed by applying the forward−backward algorithm.

Despite the fact that the overall prediction accuracy of de novo sequencing has constantly increased during the most recent years, it is nowadays still not possible to predict whole sequences de novo from a complex sample. Thus, a very useful quality scoring criterion in practical applications is the posterior value of single amino acids. Assume an amino acid starts at mass *m′* and ends at mass *m″*. This amino acid-specific posterior probability is given by

$$P(s_{m'}, ..., s_{m''}) = \frac{\sum_{s_1,...s_{m'-1}} \sum_{s_{m''+1},...s_M} P\{s, x|s_+, \hat{M}\}}{P\{x|s_+, \hat{M}\}} \qquad (23)$$

Both the numerator and the denominator can again be computed by way of the forward−backward algorithm. Here, as for the parent mass calculation, the exact computation of the posterior values for the model with grouping would require an intractable integration over all *b* variables; therefore, the posterior values for the grouping model are approximated by the posterior estimates of the model without grouping.

EXPERIMENTS

To compare the hidden Markov model with other de novo peptide sequencing methods, we have chosen the benchmark test composed by Frank and Pevzner.[3] The dataset contains 1252 tandem mass spectra of doubly charged tryptic digest peptides. The spectra originate from the 18-protein mixture dataset from Keller et al.[21] and the open proteomics database from Prince et al.[22] The spectra are divided into two sets: a training set containing 972 spectra and a test set containing 280 spectra. The corresponding sequences are validated by a SEQUEST search against a 20-Mb nonredundant protein database with nonspecific digestion.

In a preprocessing step, the mass scale of the spectra is discretized in ∼1-Da mass units. The peak heights are discretized into five equally populated bins. The emission probability distribution is then modeled as a multinominal distribution over these bins. To cope with the case that the peak height varies strongly with the relative mass position, we have divided the spectrum into seven equally sized mass regions. Different emission probability tables are learned for these seven regions. In a first step, we use the HMM for estimating the parent mass. For 95.7% of the spectra, this estimate is correct. Having derived the parent mass, the HMM is applied to infer the peptide sequence.

The prediction accuracy and recall is measured as

$$\text{precision} = \frac{\text{number of correct amino acids}}{\text{number of predicted amino acids}} \qquad (24)$$

$$\text{recall} = \frac{\text{number of correct amino acids}}{\text{number of all amino acids in test set}} \qquad (25)$$

The recall value is the fraction of correctly inferred amino acids as compared to the total number of amino acids in the test data. The precision value is the fraction of correctly inferred amino acids as compared to the total number of inferred amino acids. By restricting the length of subsequences predicted by an algorithm, one can vary the precision and recall values. Shorter subsequences

[20] G. David Forney, J. *Proc. IEEE* **1973**, *61*, 268−278.

[21] Keller, A.; Purvine, S.; Nesvizhskii, A. I.; Stolyar, S.; Goodlett, D. R.; Kolker, E. *OMICS* **2002**, *6*, 207−212.

[22] Prince, J. T.; Carlson, M. W.; Wang, R.; Lu, P.; Marcotte, E. M. *Nat. Biotechnol.* **2004**, *22*, 471−472.

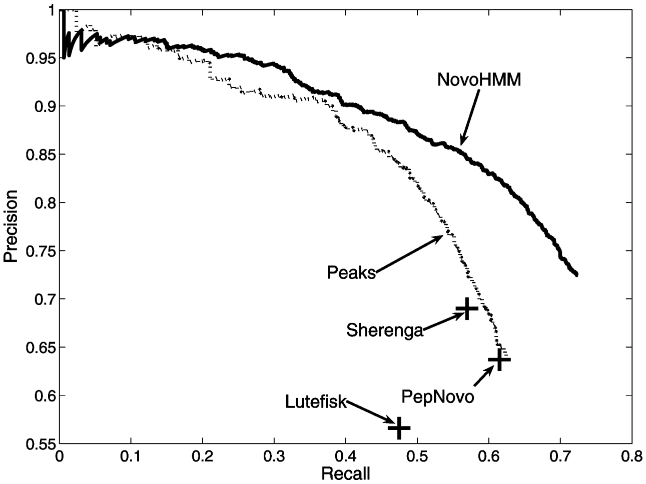

Figure 7. The precision−recall curves for the HMM as compared with other de novo sequencing methods. Tolerance criterion: exact elementary mass. The closer the curve to (1, 1), the better.

Table 1. Prediction Accuracy with Mass Tolerance 2.5 Da

algorithm	precision	recall	av length
NovoHMM[a]	0.769	0.703	9.59
NovoHMM (whole seq.)[b]	0.737	0.736	10.47
NovoHMM simple[c]	0.702	0.671	10.02
PepNovo	0.727	0.703	10.39
Sherenga	0.69	0.570	8.65
PEAKS	0.673	0.662	10.32
Lutefisk	0.566	0.475	8.79

[a] HMM in which the recall is chosen identical to PepNovo. [b] Precision−recall values for the whole sequence. [c] HMM without N/C-bit estimation.

with high confidence give a large precision value, but the recall is low, because only a small fraction of amino acids is predicted. Conversely, if longer subsequences with lower confidence are provided, the whole sequence will have a lower precision, but a higher recall, since more amino acids are predicted. In the HMM model, the sequence can be restricted to amino acids that exceed a certain posterior value. If the posterior cutoff increases, the precision will increase, but the recall will decrease. Varying cutoff values gives the precision−recall curves in Figure 7. The closer the precision−recall curve approaches the point (1, 1), the better the prediction method is.

We did not distinguish between leucine (L) and isoleucine (I) and between lysine (K) and glutamine (Q), which have almost the same mass and cannot be distinguished by low-resolution tandem mass spectrometry. In the first comparison, we considered two amino acids to be correct if the difference in mass position of an amino acid in the original spectrum and in the predicted spectrum is ≤2.5 Da, as proposed by Frank and Pevzner.[3]

In Table 1, NovoHMM is compared to other de novo sequencing methods. The posterior cutoff is chosen such that the recall is identical to that of PepNovo. Furthermore, we show the precision and recall for NovoHMM, inferring the whole sequence. To emphasize the improvement by introducing the N/C-bits, we have added the precision−recall values of the HMM with N/C-bits fixed to 0.5 (HMM simple). Note that NovoHMM outperforms all other competitors in terms of prediction accuracy. Table 2 presents the relative frequency of correctly labeled subsequences of length at least *x*. Whereas PepNovo is superior for short subsequences, NovoHMM outperforms PepNovo for long ones.

In a second comparison, we considered an amino acid to be correct if the label is correct and if the elementary mass of the amino acid in the original spectrum and the predicted spectrum is identical (see Table 3). NovoHMM outperforms PepNovo for all parameter settings. Moreover, the HMM is advantageous over all other investigated methods, since the latter are systematically worse in terms of both precision and recall. The results for the exact matching criterion in Table 3 show that when compared to PepNovo, the HMM supports a much better localization of the amino acids. Since the values for different recall and precision values are difficult to compare, we plotted the whole precision−recall curves in Figure 7. For NovoHMM and for PEAKS, the curves are computed by varying the posterior cutoff value. For the other methods, it was not possible to vary any parameter; therefore, they appear as only a single point.

The HMM model (and also PepNovo) was trained and tested on data of only a small number of proteins. Training and test data are generated on the same machines. To show that the results are not an overfitting toward a certain machine or to certain proteins, we trained the HMM on 522 spectra derived from a complex sample of vacuola proteins (*Arabidopsis thaliana*). The precision is 0.739 at a recall of 0.703 in the accuracy measure with 2.5-Da tolerance and a precision of 0.769 at a recall of 0.615 for the exact elementary mass measurement. The values are certainly smaller, since the training and test instances are measured under different conditions, but they are only slightly worse. The best results on de novo sequencing will be derived

Table 2. Percentage of Correct Subsequences of Length at Least *x*[a]

	predictions with correct subsequences of length at least							
algorithm	$x = 3$	$x = 4$	$x = 5$	$x = 6$	$x = 7$	$x = 8$	$x = 9$	$x = 10$
NovoHMM[b]	0.893	0.796	0.711	0.589	0.486	0.404	0.293	0.193
NovoHMM (whole seq.)[c]	0.911	0.829	0.743	0.632	0.546	0.464	0.336	0.229
NovoHMM simple[d]	0.864	0.761	0.636	0.542	0.446	0.379	0.279	0.186
PepNovo	0.946	0.871	0.800	0.654	0.525	0.411	0.271	0.193
Sherenga	0.821	0.711	0.564	0.364	0.279	0.207	0.121	0.071
PEAKS	0.889	0.814	0.689	0.575	0.482	0.371	0.275	0.179
Lutefisk	0.661	0.521	0.425	0.339	0.268	0.200	0.104	0.057

[a] Prediction accuracy with mass tolerance 2.5 Da. [b] HMM in which the recall is chosen identical to PepNovo. [c] Precision−recall values for the whole sequence. [d] HMM without N/C-bit estimation.

Table 3. Accuracy with Exact Elementary Mass Matching

algorithm	precision	recall	average length
NovoHMM[a]	0.823	0.615	10.28
NovoHMM (whole seq.)[b]	0.725	0.724	10.42
NovoHMM simple[c]	0.693	0.662	10.04
PepNovo	0.637	0.615	10.12

[a] HMM in which the recall is chosen identical to PepNovo. [b] Precision−recall values for the whole sequence. [c] HMM without N/C-bit estimation.

when the HMM is retrained if the mass spectrometry conditions change.

The mean runtime of the HMM without grouping is ∼0.1 s per spectrum and 0.9 s for the HMM with grouping on a standard PC. The memory usage is less than 20 MB for the HMM without grouping and less than 70 MB for the HMM with grouping.

To show that the method is able to find posttranslational modifications, we have searched a dataset of spectra from vacuola proteins (*A. thaliana*). The following spectra could be identified as modified peptides: PM[16/Oxidation]EEGLAEAIDDGR and AAHFEESM[16/Oxidation]K.

The underlying (unmodified) sequences can be found in the *Arabidopsis* protein database. For both modified peptides, at least three further peptides belonging to the same protein were found by a database search.

CONCLUSION

A novel method for the analysis of mass spectra in de novo peptide sequencing is presented in this paper. The proposed hidden Markov model emulates the generation process of mass spectra in a fully probabilistic way, which supports a clean separation between signal and noise in the complex mass spectra. The model was tested on a benchmark data set of publicly available mass spectra. The HMM clearly outperforms competing de novo sequencing algorithms in recognition of the parent mass and prediction accuracy and especially in peak localization. The success of NovoHMM demonstrates the flexibility of the machine learning framework for bioinformatics, and soon, we expect substantial progress in discovering posttranslational modifications in complex mass spectra as they arise in proteomics. Furthermore, NovoHMM can also support an advanced database search, and it can provide the user with confidence intervals for sequences identification, a very valuable estimate for statistically sound proteomics.

ACKNOWLEDGMENT

We thank Ari Frank for providing us the sequencing results of PepNovo. Further we thank John Prince and Andy Keller for providing the mass spectra. The software can be downloaded from http://people.inf.ethz.ch/befische/proteomics.

Received for review May 20, 2005. Accepted September 2, 2005.

AC0508853