Probabilistic De Novo Peptide Sequencing with Doubly Charged Ions

Hansruedi Peter, Bernd Fischer, and Joachim M. Buhmann

Institute of Computational Science
ETH Zurich, Switzerland
http:www.ml.inf.ethz.ch

Abstract. Sequencing of peptides by tandem mass spectrometry has matured to the key technology for proteomics. Noise in the measurement process strongly favors statistical models like NovoHMM, a recently published generative approach based on factorial hidden Markov models [1,2]. We extend this hidden Markov model to include information of doubly charged ions since the original model can only cope with singly charged ions. This modification requires a refined discretization of the mass scale and, thereby, it increases its sensitivity and recall performance on a number of datasets to compare favorably with alternative approaches for mass spectra interpretation.

1 Introduction

Proteins control all metabolic processes in biological cells of living organisms. To understand the dynamics and interaction of these processes biologists have to identify all involved proteins, i.e. determine their sequence and their abundance. From a computer science perspective a protein is a string over an alphabet of 20 amino acids. Even for short strings, this combinatorics generates an incredibly high number of possible combinations. The identification of these sequences is becoming increasingly important also for medical research. In biomarker discovery based on gene expression micro-arrays physiologists rely on tissue samples from the affected tissue. Recent research results provide evidence that protein based biomarker discovery can be performed solely on blood samples in the near future [3]. This diagnostics will then be a great research step in early detection of cancer and we might even call it "remote sensing of cancer".

The most promising method of high-throughput protein sequencing is tandem mass spectrometry. The proteins are biologically broken in short sequences by enzymatic digestion. For each peptide a mass spectrum is generated that includes mass measurements of fragments of the peptide. In addition, a rough estimate of the peptide mass is available from liquid chromatography (LC/MS). Peptide sequencing aims at infering the underlying amino acid sequence given the mass spectrum and the mass of the peptide.

Today, the genomes of many organisms have already been sequenced. Given the DNA sequence we can compile a database of protein sequences that can be transcribed and translated from these genes. In a first analysis step biologists will infer the peptide sequences using side information of the protein databases ([4,5]). This procedure, although conceptually very appealing, has some difficulties: (i) the databases are still

K. Franke et al. (Eds.): DAGM 2006, LNCS 4174, pp. 424–433, 2006.

incomplete or have some errors from the sequencing process; (ii) there can be unknown splice variants or even unknown genes in the DNA; (iii) there exist post-translational modifications. Due to combinatorial limits one can not enumerate all possibilities of genes, splice variants and post-translational modifications. Therefore a cascading search strategy is recommended [6], where less and less side-information is used to narrow down the possible amino acid sequences. As a complement to database search, peptide sequences can be infered from mass spectrometry data in a *de novo* fashion and we are following this strategy here. The only information used in *de novo* peptide sequencing is the alphabet of 20 amino acids and the mass spectra as input.

LUTEFISK [7,8] and PEAKS [9] are two widely used *de novo* peptide sequencing methods. LUTEFISK creates a weighted graph using a simple scoring scheme. The sequence is the shortest path in the generated graph. PEAKS creates a similar weighted graph and generates a candidate list of peptides by searching the weighted graph. A refined search is performed afterwards on the candidate list. Dancik [10] proposed the first probabilistic scoring scheme for *de novo* sequencing. It simply estimates the fragmentation pattern of the peptides at a small number of positions. The scoring scheme was improved by adding a noise model and a Bayesian network to model the fragmentation patterns [11]. Fischer *et al.* [1,2] proposed a generative hidden Markov model (`NovoHMM`) of mass spectra. This model can only describe singly charged fragment ions which is a clear shortcoming since about 10-25 percent of the ions are doubly or tripply charged. We will substantially extend `NovoHMM` to include doubly charged fragment ions by refining the discretization of the mass scale.

The next section will summarize the essentials of tandem mass spectrometry as far as our modelling is concerned. The hidden Markov model is presented in section 3. In section 4 we give our extensions of a refined discretization and the inclusion of information from doubly charged fragment ions. The new model is tested on different datasets (sec. 5). It clearly outperforms all its competitors. Furthermore, we show that the model can be applied to triply charged peptide ions.

2 Tandem Mass Spectrometry

The proteomics process pipeline based on mass spectrometry contains the following steps: first, the proteins are digested with an enzyme (typically Trypsin). This chemical process yields a sample of small peptides. A peptide P consists of a short sequence of n amino acid residues $P = a_1a_2a_3, ..., a_n$, with an additional H-atom at the N-terminus and an OH-group at the C-terminus. In figure 1 a small peptide composed of three amino acids is depicted. Typically peptides have 10-20 amino acid residues and they are separated by liquid chromatography.

In a first measurement step the mass of the peptides can be read out from a mass spectrum. Ions (peptides) from a small mass window are selected and they are fragmented in typically two pieces by collision with a noble gas. As shown in figure 1, the most common fragment ions which are denoted as a-,b-,c-,x-,y- and z-ions, are generated by breaking the peptide backbone. The tandem mass spectrum contains peaks at mass positions corresponding to the different fragment ions. The inference of the underlying sequence given a mass spectrum is the goal of peptide sequencing.

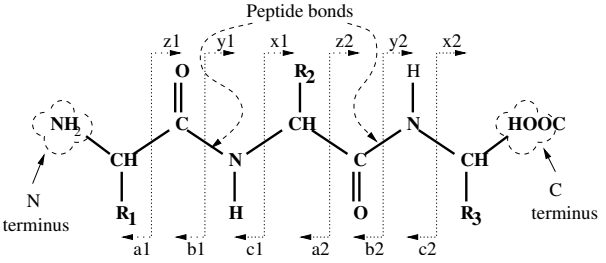

Fig. 1. A simple peptide with three amino acids, an additional H-atom at the N-terminus and an OH-group at the C-terminus. The amino acids are connected by peptide bonds. Depending on the internal link a peptide is broken, the corresponding fragment ions are called a-/b-/c-ions when containing the N-terminus respectively x-/y-/z- ions when containing the C-terminus.

To measure a peptide by mass spectrometry, the peptide has to be charged. Most common the peptide is double positively charged. After the fragmentation most fragment ions are singly charged. In 10–25 percent of the cases the resulting fragment ions are doubly charged. The mass spectra consists of a list of mass-charge ratios m/z and their corresponding intensities. Doubly charged ions appear at half of the position of the corresponding singly charged ions. This shift induces long-range dependencies that are difficult to model for standard hidden Markov models, e.g., the model of Fischer *et al.* [2] can only describe singly charged ions. In this paper we will extend the model for doubly charged ions.

3 Factorial Hidden Markov Model (NovoHMM)

In the hidden Markov model of [2] the (time-)step is 1 elementary mass unit (the mass of a proton or neutron). With the transition probabilities a distribution over all amino acid sequences is modeled. The hidden random variables (representing the peptide sequence) have 2375 states. For each amino acid there are as many states as the amino acid has elementary particles. This chain of states for each amino acid is a counter for the mass. In NovoHMM only peptide sequences with the given peptide mass are considered. This constraint is implemented by introducing a positive and negative end state. After as many time steps as the peptide mass indicates, the model ends in the positive end state for all sequences that obey the peptide mass constraint.

Hypothetical spectra, composed of mass peaks from prefix fragment ions only, can be interpreted in straight forward way: the different possible fragment ions (a-, b-, c-ions) are assigned to specific counter positions on the mass scale. The b-ion is placed at the last counter state of each amino acid. The a-ion (It is a b-ion without a carbon monoxide) is placed 28 counter states before the b-ion. Additional fragments like neutral water loss can be modeled in the same fashion.

There exists, however, a strong dependence between prefix and suffix ions via the total peptide mass. For every prefix fragment ion with mass m there exists with high probability a mass peak of a suffix fragment ion at mass position $M - m$, where M is the total peptide mass. To overcome these long-range dependencies, the Markov model

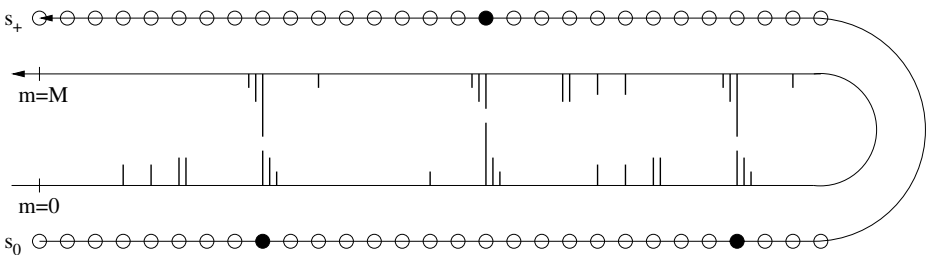

Fig. 2. The internal mirror symmetry of the problem is illustrated by folding the spectrum in the middle. The Markov chain models a sequence with four amino acids. The filled circles correspond to the amino acid boundaries. Around each amino acid boundary a peak pattern is generated once for the *N*-terminal fragments and once for the C-terminal fragments.

is duplicated (see fig. 2). One Markov model starts at the beginning of the sequence and another Markov model starts in the end of the sequence. In the figure a peptide sequence with four amino acids is depicted. The black dots denote the amino acid boundaries. The boundary at mass m generates a peak pattern around mass m and $M - m$. The left part of figure 5 shows the graphical model. Both emission variables x_i and x_{M-i} depend on the sequence variables s_i and s_{M-i}. The new model is still a hidden Markov model but it has a factorial structure [12].

The number of hidden states of the new model is squared compared to the simple model. To reduce the model complexity and the runtime the model is approximated by a mixture model. A second set of hidden variables is introduced: The binary variables B_i decide for each peak x_i if it is generated by a prefix- or a suffix fragment. The emission probability is now a mixture of a prefix-distribution and a suffix-distribution instead of the joint distribution of both.

4 Doubly Charged Fragment Ions

Before the fragmentation process, most peptides are double positively charged. During the collision with the noble gas, the peptide ions are broken in two (or more) fragments as depicted in figure 3. In the figure two possible outcomes of the fragmentation process are shown. In the first case the peptide ion breaks in two parts. Both parts are single positively charged. This can result in two peaks in the spectrum, e.g. at mass-over-charge ratios 418 and 457. In the second case the peptide is broken in the same prefix and suffix parts, but now the prefix is double positively charged and the suffix is not charged. The prefix peak will appear at mass-over-charge ratio 229. Since a mass spectrometer can only measure charged ions, a suffix ion peak does not appear in the spectrum.

In a first step, the discretization of the spectra has to be refined to handle the smaller distance between the doubly charged fragment peaks. In `NovoHMM`, a spectrum is discretized into bins of approximately 1 Dalton width, where the peaks are placed (by definition) in the middle of the bin. Doubly charged fragment ions are spaced with a minimal distance of one half mass-over-charge unit. To consider this effect, we have to discretize the leftmost half of the spectrum in a different way, as shown in figure 4.

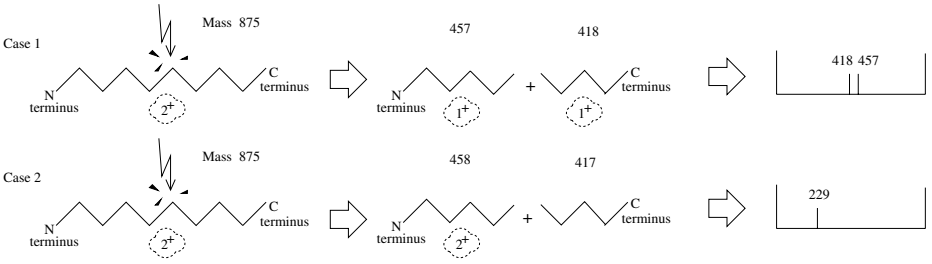

Fig. 3. Cleavage of a peptide by collision with a noble gas. Two possibilities are shown: splitting into two singly charged fragments (Case 1); and into a doubly charged and an uncharged fragment (Case 2). On the right, the corresponding mass spectra are depicted.

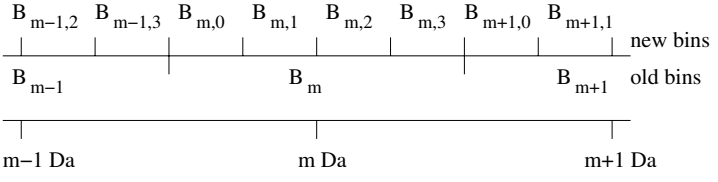

Fig. 4. Discretization of a spectrum for the new model (new bins) and NovoHMM (old bins). In the new discretization, each bin is divided into four sub-bins.

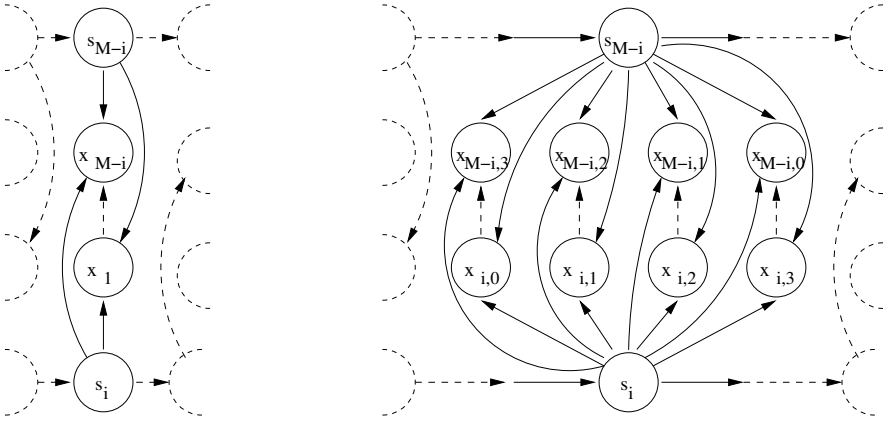

Fig. 5. Part of the dependency structure of NovoHMM (at the left) with the corresponding part in the model with the new discretization. Each emission variable is replaced by four variables for a finer discretization of the spectrum.

In the graphical model (fig. 5) the hidden variables (representing the sequence) are not changed. The emission variables are replaced by four variables each. Note that doubly charged ions can only occur in the first half of the spectrum. Therefore it would be enough to discretize the first half of the spectrum in this way. For reasons of model consistency we decided to discretize the complete spectrum in such a way. As we will see later in the experimental section, the new discretization itself will improve the prediction accuracy of the model.

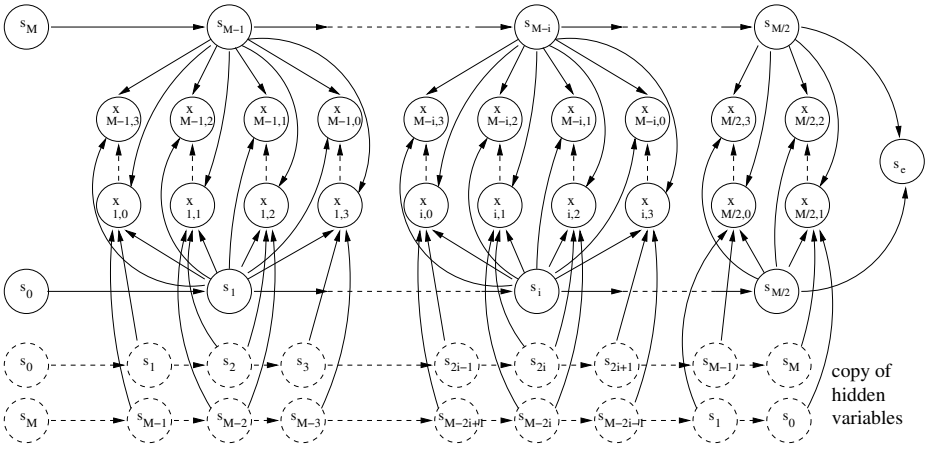

Fig. 6. The dependency structure of the factorial HMM for the model including doubly charged ions. The sequence variables in the last two rows are the same as in the model above. They are only copied for visualization reasons. Each emission variable in the first half of the spectrum depends now on four different sequence variables.

In a second step we can now include the information from doubly charged ions. The emission variables representing the first half of the spectrum have additional dependencies. The variable x_m does additionally depend on s_{2m} and s_{M-2m} which encode the corresponding doubly charged prefix and suffix ion. The upper half of figure 6 shows the graphical model for the new discretized factorial hidden Markov model. To draw the dependencies in a more transparent way, we have copied the sequence variables in the plot. The lower two rows of variables are the same as the sequence variables in the upper model. The problem is again approximated by a mixture model. For each mass bin in the lower half of the spectrum there is now an assignment variable with four different states: 1-charged prefix, 1-charged suffix, 2-charged prefix and 2-charged suffix. The assignment problem is solved by the expectation-maximization algorithm. In the E-step the expectation over all hidden variables (sequence and assignment variables) is computed. Since this is computationally intractable, we decomposed the E-step into two stages. In the first stage, the expectation of the assignment variables is estimated; in the second stage the expectation of the sequence variables is computed by the forward-backward algorithm as proposed in [2].

5 Experiments

To justify our model extension of NovoHMM, we first present experimental results for the model which includes doubly charged fragment ions. Furthermore, the prediction quality of this model (NovoHMM++) is compared to the prediction quality of NovoHMM and other *de novo* sequencing methods on two different datasets. The first data set is composed by Frank and Pevzner [11]. It contains 972 spectra in the training dataset and 280 spectra in the test dataset. The second dataset contains more than 5000 spectra from an unpublished proteomics experiment. We performed 10-fold cross validation on the second dataset and we used the proposed splitting in training and test spectra for the first dataset.

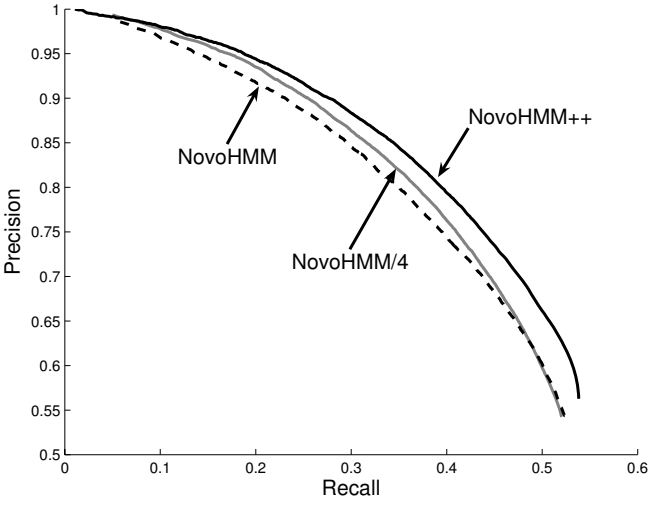

Fig. 7. The precision-recall curves achieved for cross-validation on a the second dataset

When nothing else is mentioned, we consider 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 less than or equal to 2.5 Dalton (see [11]). Furthermore, no distinction has been made between leucine (L) and isoleucine (I), or between lysine (K) and glutamine (Q), as they have almost the same mass and cannot be distinguished by low mass resolution tandem mass spectrometry. The precision value is defined as the number of correct amino acids divided by the number of predicted amino acids. The recall value is defined as the number of correct amino acids divided by the true number of amino acids. In the plots shown later in this section the precision and recall values are varied by changing a threshold on the posterior value computed with the forward-backward algorithm.

To simplify notation, we introduce names for the different versions of our algorithm:

- NovoHMM – The original version of NovoHMM as it was presented in [2].
- NovoHMM/4 – NovoHMM with fine discretization in quarters of a Dalton, but without a model for doubly charged fragment ions.

– NovoHMM++ – NovoHMM with fine discretization in quarters of a Dalton and with an additional model for doubly charged fragment ions.

Figure 7 depicts the precision-recall curves on the second dataset. The readers can easily convince themselves that the new discretization alone improves already the prediction accuracy. The accuracy is further boosted by information from doubly charged ions as they are included in NovoHMM++.

NovoHMM++ was also tested on a dataset with 1020 triply charged peptides from the second dataset. With an average length of 2414.5 Dalton these peptides are clearly longer than the peptides in the datasets of doubly charged peptides. A triply charged peptide is expected to split up into a singly charged fragment ion and a doubly charged fragment ion. We achieved a precision of 0.216 at a recall of 0.200. At a first glance, these precision and recall values may look low, but one has to consider the substantial lengths of the peptides in the dataset which clearly influences the predictive power of the algorithm. For comparison, NovoHMM achieved about 10% precision and recall on this dataset.

NovoHMM++ clearly outperforms all competing algorithms on the first dataset (see table 1). Table 2 presents the relative frequency of correctly labeled subsequences of length at least x. Whereas PepNovo is slightly superior for short subsequences, NovoHMM++ clearly exceeds all its competitors for long peptides. In figure 8, the precision-recall curves of NovoHMM++ and NovoHMM/4 are compared with other *de novo* sequencing methods. The closer the curves approach (1, 1) values, the better is the recall performance. In general, NovoHMM/4 and NovoHMM++ are relatively

Table 1. Comparison of the performance of our algorithms with other *de novo* sequencing methods on the first dataset

Algorithm	NovoHMM++	NovoHMM/4	NovoHMM	PepNovo	Sherenga	Peaks	Lutefisk
Correctly predicted symbols (of 2935)	**2293**	2244	2160	2063	1673	1943	1394
Precision	**0.787**	0.770	0.737	0.727	0.690	0.673	0.566
Recall	**0.781**	0.765	0.736	0.703	0.570	0.662	0.475

Table 2. Percentage of correct subsequences of length at least x

Algorithm	Predictions with correct subsequences of at least							
	$x = 3$	$x = 4$	$x = 5$	$x = 6$	$x = 7$	$x = 8$	$x = 9$	$x = 10$
NovoHMM++	0.943	0.861	0.796	**0.714**	**0.621**	**0.539**	**0.429**	**0.300**
NovoHMM/4	0.943	**0.871**	0.793	0.686	0.607	0.532	0.396	0.264
NovoHMM	0.911	0.829	0.743	0.632	0.546	0.464	0.336	0.229
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

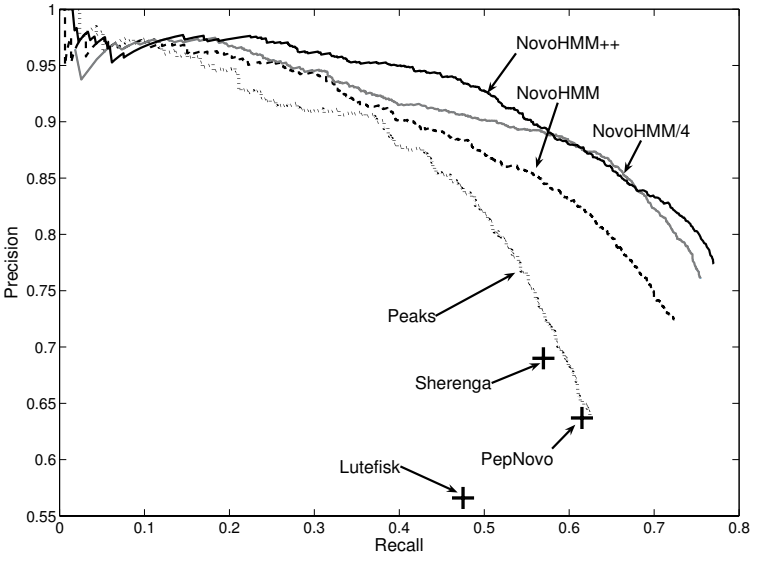

Fig. 8. The precision-recall curves for our algorithms compared with other de novo sequencing methods. Tolerance criterion: exact elementary mass.

close to each other, whereas for high recall values NovoHMM++ seems to be superior to NovoHMM/4.

The model described in the paper includes doubly charged prefix- and suffix ions in a generative hidden Markov model to interprete mass spectra. In addition, we tested the same model including just doubly charged prefix ions or just doubly charged suffix ions. The performance, when including suffix ions only, is almost the same as when we incorporate both ions. On the other hand, with the doubly charged prefix ions only, the model does not demonstrate the performance of NovoHMM++. Therefore we conclude that mainly the doubly charged suffix ions matter are responsible for the performance increase of the model. This observation is biologically plausible, since peptides (and thus suffix ions) digested by Trypsin often end with lysine (K) or arginine (R). These amino acids are known to attract positively charged ions.

6 Conclusion

Factorial Hidden Markov Models provide a flexible framework to explain mass spectra which are gathered from proteomics experiments. An extension of NovoHMM is presented in this paper which contains an additional model for doubly charged fragment ions and a refined discretization. This new model NovoHMM++ increases the accuracy of the predicted sequences by up to 5% in precision and recall on different datasets. On a benchmark test [11], NovoHMM++ substantially and significantly outperform the

most prominent *de novo* sequencing algorithms in terms of prediction accuracy. In addition, NovoHMM++ was shown to reliably explain also mass spectra containing triply charged peptides.

Acknowledgement. We thank Jonas Grossmann, Sacha Baginski and Wilhelm Gruissem (Inst. of Plant Sciences, ETH Zurich) for providing the mass spectrometry data.

References

1. 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. Volume 17., USA, MIT press (2005) 457–464
2. Fischer, B., Roth, V., Roos, F., Grossmann, J., Baginsky, S., Widmayer, P., Gruissem, W., Buhmann, J.M.: NovoHMM: A hidden markov model for de novo peptide sequencing. Analytical Chemistry **77**(22) (2005) 7265–7273
3. Chen, R., Pan, A., Brentnall, T., Aebersold, R.: Proteomic profiling of pancreatic cancer for biomarker discovery. Molecular and Cellular Proteomics **4**(4) (2005) 523–533
4. Eng, J.K., McCormack, A.L., III., J.R.Y.: An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. American Society for Mass Spectrometry **5**(11) (1994) 976–989
5. Hirosawa, M., Hoshida, M., Ishikawa, M., Toya, T.: Masqot: multiple alignment system for protein sequences based on three-way dynamic programming. Computer Applications in the Bioscience **9**(2) (1993) 161–167
6. Sadygov, R.G., Cociorva, D., III., J.R.Y.: Large-scale database searching using tandem mass spectra: Looking up the answer in the back of the book. Nature Methods **1**(3) (2004) 195–202
7. Taylor, J.A., Johnson, R.S.: Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. Rapid Communications in Mass Spectrometry **11** (1997) 1067–1075
8. Taylor, J.A., Johnson, R.S.: Implementation and uses of automated de novo peptide sequencing by tandem mass spectrometry. Analytical Chemistry **73** (2001) 2594–2604
9. Ma, B., Zhang, K., Hendrie, C., Liang, C., Li, M., Doherty-Kirby, A., Lajoie, G.: Peaks: Powerful software for peptide de novo sequencing by tandem mass spectrometry. Rapid Communications in Mass Spectrometry **17**(20) (2003) 2337–2342
10. Dancik, V., Addona, T.A., Clauser, K.R., Vath, J.E.: De novo peptide sequencing via tandem mass spectrometry. Journal of Computational Biology **6** (1999) 327–342
11. Frank, A., Pevzner, P.: Pepnovo: De novo peptide sequencing via probabilistic network modeling. Analytical Chemistry **77**(4) (2005) 964–973
12. Zoubin Ghahramani, M.I.J.: Factorial hidden markov models. Machine Learning **29**(2–3) (1997) 245 – 273