

Probability Profiles - Novel Approach in Tandem Mass Spectrometry *De Novo* Sequencing

Tema Fridman¹, Robert Day¹, Jane Razumovskaya², Dong Xu² and Andrey Gorin^{1*}

¹Computer Science and Mathematics Division, ²Life Sciences Division, ORNL, Oak Ridge, TN 37830

*Corresponding author agor@ornl.gov

A novel method is proposed for deciphering experimental tandem mass spectra. A large database of previously resolved peptide spectra was used to determine “neighborhood patterns” for each peak category: C- or N-terminus ions, their dehydrated fragments, etc. The established patterns are applied to assign probabilities for new spectra peaks to fit into these categories. A few peaks often could be identified with a fair confidence creating strong “anchor points” for De Novo algorithm assembling sequence subgraphs. Our approach is utilizing all informational content of a given MS experimental data set, including peak intensities, weak and noisy peaks, and unusual fragments. We also discuss ways to provide learning features in our method: adjustments for a specific MS device and user initiated changes in the list of considered peak identities.

1. Introduction

Development of robust computational algorithms for protein identification in mass spectrometry (MS) experiments is a challenging data analysis problem with a typical set of “biological attributes”: abundance of experimental noise, incomplete data, and possibility of wild card factors such as post-translational modifications. Novel and robust MS analysis approaches are crucially important for emerging DOE Genomes-to-Life Program efforts to compile a comprehensive catalog of cellular protein-protein complexes through high-throughput mass spectrometry. The existing methods heavily rely on sequence database lookups and could be prone to many types of the database errors: single residue mutations, unrecognized translation frames, and unidentified modifications. These difficulties grow exponentially if applied to identification of the cross-linked peptide constructs, and such capability is essential for large-scale efforts to get more informative picture of protein-protein interaction using mass-spectrometry methods.

An attractive idea to utilize only spectrum information for peptide identification is known as *De Novo* sequencing approach [1]. In one possible implementation peaks corresponding to ions of the same type, that differ by precisely one amino acid in mass, form adjacent nodes in a sequencing graph. Then possible graph paths represent partial sequences of the parent peptide, and could be used independently of protein sequence database or in conjunction with it [2].

The development of *De Novo* methods is an active research area, but so far substantial progress has been reported mainly for the simulated peptide spectra consisting of calculated sets of few standard ion types [3]. In reality the experimental MS spectra are saturated with peaks from many types of fragments, and the origin for some of them is hard to establish even with knowledge of the parent peptide. Many of the “noise” peaks have a low intensity value, but this observation does not solve all problems. Even in the top intensity bin, containing half of the standard b- and y-ions, those “noble” ions are outnumbered in ratios of 1:8 and 1:5 by peaks from other types of fragments and could not be easily separated from them. Here we formulate a novel approach for peak identification based on all-inclusive analysis of its spectral neighborhoods. The “probabilistic identity” calculated from our analysis strongly correlates with true peak identity, and by our preliminary results could be efficiently utilized for innovative solutions in *De Novo* sequencing.

2. Diversity of peak neighborhoods

Peaks in an experimental MS spectrum can be classified into two large categories: the “noble” b- and y-ions corresponding to N- and C- fragments of the peptide bond breaking and the “fragment noise” ions. While the noble ion mass values are the base for peptide sequence reading, the rest is usually perceived as a nuisance, cluttering spectrum and leading to weird results in many existing *De Novo* algorithms, when they applied to real

life experimental data. The sources of the extra fragments are multiple: b- and y-ions losing certain chemical groups (e.g. water and NH₃); internal fragments, formed when more than one peptide bond is broken; fragments of “intruder” peptides, due to imperfect liquid chromatography separation; chemical contaminants; and possible random physical effects. Further we will only consider single charged ions, and ion positions are given in Da/e (Dalton per electron charge).

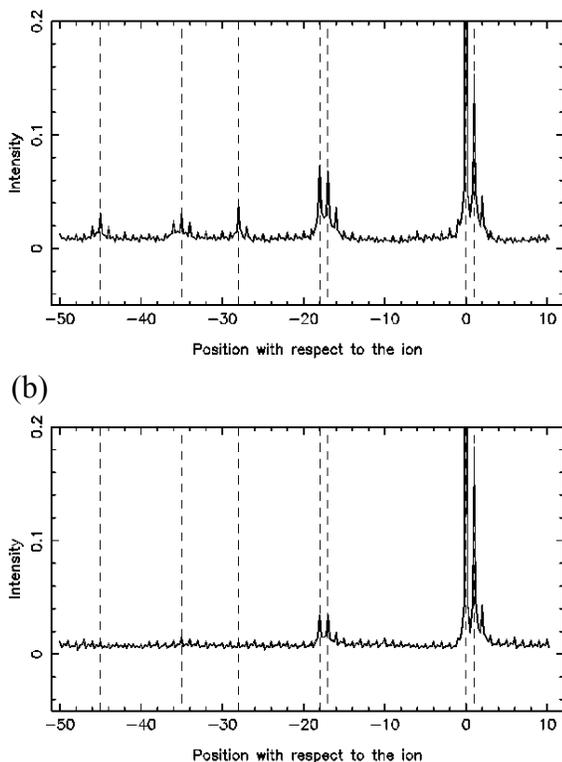


Figure 1. Averaged neighborhood patterns of: (a) b-ion (b) y-ion.

We argue that the detailed consideration of the fragment ion preferences to appear around noble ions could be helpful for the identification problem, as the local spectral neighborhood strongly depends on the peak's category. For example, a b-ion often loses C=O (producing so called a-ion), therefore we expect to see a peak at -28 Da/e to the left of the b-ion peak more often than to the left of y or any other ion. If both b- and y-ions easily lose water, we would often expect to see peak at -18 Da/e to the left of these noble ions, and the “event” of peak observation at this position can be used for discrimination purposes.

The average profiles of b- and y- ions neighborhoods are presented at Figure 1. The data were

collected from 2741 experimental spectra with 16000 b- and 18150 y-ions mixed to 248850 other peaks (when peaks for b- and y-ions have coincided on 276 occasions, they were assigned to the b category). For each considered ion its neighborhood landscape (-45, +10) was normalized by the ion intensity value. First and second isotopes are seen the most clearly. Another signal comes from losing water and ammonia by both ions. Notice that the relative intensities pattern of -18, -17 and -16 peaks differs from that of the isotopic envelope of the reference ion. The rest of statistically significant signals come only from b ion Figure 1a. It is the mentioned above loss of the C = O group at -28 (here the pattern is consistent with the pattern of the isotopic envelope), a peak at -35 - perhaps loss of water and ammonia, and a peak at -45 - likely to be a-ion with a lost ammonia.

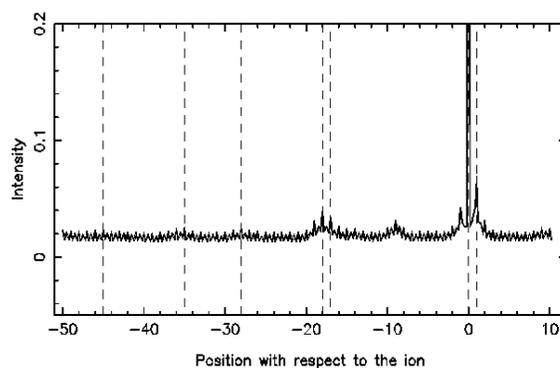


Figure 2. Averaged neighborhood patterns of other ions (not b or y).

Figure 2 demonstrates statistical average neighborhood of a peak that is neither a b- nor a y-ion. It has isotopic peaks at +1, +2 from internal fragment ions and first isotopes of b- and y-ions (isotopes were not excluded for the analysis above), and a peak at -1, coming from isotopes of the internal fragments and second isotopes of b- and y-ions. The isotope presence is responsible also for the weak signal at -18. There is an interesting signal at -9 - echo of double charged b and y ions with lost water or ammonia.

The presented data demonstrate that different peak categories have very distinct neighboring regions. The challenge is to separate the signal from the noisy background and to deal with its statistical nature. In the next section we discuss how to select features describing ion neighborhoods, how to collect statistics for those features, and how to calculate probability that a given peak belongs to a particular category.

3. Probability profiles for peak categories

From the previous analysis we have a good idea about an approach separating different kinds of peaks. Suppose we have 3 classification categories: for example, b-ions (category B), y-ions (category Y), and the rest (category R). First we count frequencies of various features to be present around using positively identified peaks (using large database of deciphered MS spectra) and then use those frequencies to calculate contribution of each feature presence (or absence) into probabilities for a given unknown peak to identify B, Y, or R. This approach retains all the information from the original spectrum, but ranks peaks in quantitative manner. The peak with 90% probability to be R-peak will not be used as the first choice for the sequence reconstruction routine, but still can be used if there are no other choices. Correspondingly the generated solution can be given an objective score of its reliability.

In practice neighborhood features can be implemented in the following way. For each positively identified peak of each category, a small spectrum interval located in its vicinity (for example, at the relative position -18 Da/e, water loss) is examined for presence/absence of another peak. Counting these frequencies one can establish, for example, that B peak is accompanied by “itself minus water” peak with probability of 0.45, Y peak probability is 0.40, and R peak had such “satellite” peak only in 0.27 fraction of all observed R peak cases. Combining together the probabilities p_{IB} , p_{IY} , and p_{IR} with the measured frequencies of the B, Y and R ions in a particular peak bin (f_B , f_Y , and f_R), one can find the probabilities for an unknown peak from the same bin to belong to each of the categories above. They are proportional to $f_B p_{IB}$, $f_Y p_{IY}$, and $f_R p_{IR}$, correspondingly, and could be normalized by $(f_B p_{IB} + f_Y p_{IY} + f_R p_{IR})$. The resulting equation for the peak to belong to B category will be then:

$$f_B p_{IB} / (f_B p_{IB} + f_Y p_{IY} + f_R p_{IR})$$

The formalism is easy to extend to feature sets, and it equally well applies to the cases of negative (peak is absent) event observations.

The described approach contains a number of parameters: (a) size of the matching “precision” interval at the considered relative positions (one reasonable choice is to use precision of the MS experiment); (b) relative intensity of the peak sufficient to record “observation”; (c) vector of peak features. If the first two selections provide a set of flexible learning parameters to adjust for particular set of experiments or a particular experimental device, the last one raises interesting research problem appealing to physical intuition of the experimentalists. From Figures 1 & 2

we can choose at least 8 feature positions (+2, +1, -1, -17, -18, -28, -35 and -45). How should one select the discriminating features from the rich variety described above? How good is the approximation that the features are independent events? Even though events like b-ion losing water and b-ion losing C=O group seem to be independent, feature peaks from noble ions do prefer to appear in a package, so we are bound to introduce conditional probabilities or measure the probabilities of the “double-feature” event directly. In the current work we chose the second approach, and found that the “negative” event at -1 (satellite peak at 1 Da/e position to the left is absent) strongly enhances discriminating power of the “positive” events at positions +1, -17, -18 and -28. Below we describe our results using this probability profile for identification of b-ions in the database of 2741 positively assigned spectra.

4. Sorting out peak identities

The frequencies of the satellite peak observation were collected for positively identified B, Y and R peaks at relative positions +1, -17, -18 and -28. The measurements were conditional on the negative event at -1 position, as we have included in our counts only those peaks from all three categories that do not have another peak at position -1. This condition appears to be valuable for two reasons. First, only 3% of B and 1% of Y peaks were excluded from the analysis due to this condition, but ~23% of the R peaks did not pass and were filtered out. Second, as the condition seems to be efficiently excluding b- and y-ion isotope peaks, the observed differences at four listed above positions become much sharper. This is natural as isotope peaks have the same preferences for fragment formation as the main ions, and the inclusion into frequency measurements dilutes the discriminative power of the selected features.

The discriminative power of four selected features (defined as the ratio between frequencies to observe a satellite peak for B category peak versus R category) falls in the narrow range: from 2.03 for +1 position to 3.3 for -28 position. The theoretical maximum for the full set was the discriminative coefficient of 45. In the reality the weaker power was weaker but still quite impressive: the set of peak identified as having highest probability to have category identity consisted on 62% of true b-ion peaks, 30% of R peaks, and only 8% of true y-ion peaks. This result should be considered against the composition of the initial set of peaks where b-ions were outnumbered 1:8 by R peaks and 2:3 by y-ion peaks. Our procedure resulted in 16-times “b-ion enrichment” of the analyzed set of peaks.

Our preliminary experiments with *De Novo* assembling algorithm have indicated that if the

sequencing graph assembly starts from the true b-ion (something we can now guarantee for most experimental spectra), it is unusually successful even if the probability information is not utilized at the further steps. Sorting out highly reliable y-ions, converting them to b and then doing assembly in the “islands” of highly reliable b-ions seems to be a way to further boost performance of our algorithm.

Acknowledgments

This work was funded in part or in full by the US Department of Energy's Genomes to Life program (www.doe-genomes-to-life.org) under two projects, “Carbon Sequestration in *Synechococcus* Sp.: From Molecular Machines to Hierarchical Modeling” (www.genomes-to-life.org) and “Center for Molecular and Cellular Systems” (www.ornl.gov/GenomesToLife), Oak Ridge National Laboratory is managed under US DOE Contract for No. DE-AC05-00OR22725

[1] V. Dancik, T.A. Addona, K.R. Clauser, J.E. Vath, and P.A. Pevzner. “*De novo* peptide sequencing via tandem mass spectrometry.” *J Comput Biol*, 1999, v 6, pp 327-42.

[2] P.A. Pevzner, V. Dancik V, and C.L. Tang “Mutation-tolerant protein identification by mass spectrometry.” *J Comput Biol*, 2000, v 7, pp 777-87

[3] T. Chen, M. Kao, M. Tepel, J. Rush, and G.M. Church “A dynamic programming approach to *de novo* peptide sequencing via tandem mass spectrometry.” *J Comput Biol*, 2001;v 8, pp 325-3