

PPM-Chain – *De novo* Peptide Identification Program Comparable in Performance to Sequest

Robert M. Day, Andrey Borziak, Andrey Gorin*

*Computer Science and Mathematics Division, Computational Biology Institute,
Oak Ridge National Laboratory, Oak Ridge, TN 37830-6114*

**agor@ornl.gov*

Abstract

Recently, we introduced Probability Profile Method (PPM), which utilizes neutral-loss neighborhoods around each peak in MS/MS spectrum to “label” it: to assign a probability that the peak in question belongs to one of the specific categories (such as b- or y-ion peaks). Here we present the PPM-chain program – a PPM-based tool for *de novo* protein tag identification. *De novo* peptide identification involves finding a connected sequence of ion peaks separated by amino acid mass intervals, corresponding to a tag - partial peptide of the source protein. In the existing approaches the number of the possible connected sequences may run into hundreds of thousands, and it increases exponentially with the desired length of the tag. PPM can be used to locate high probability islands, containing very pure sets of b- and y-ion peaks, thereby reducing computational complexity and sharply increasing precision of tagging. In addition, the obtained tags can be reliably ranked using PPM-derived probabilities assigned to the connected peaks. The value of peptide tags was demonstrated on a large database of ~20,000 spectra. With the additional flanking mass constraints PPM-chain shows precision and coverage similar to the field industrial-power standard – Sequest program, while providing a set of novel unique capabilities and significantly outperforming Sequest in speed.

1. Introduction

Tandem mass spectrometry has become the major tool in proteomics for the identification of proteins. In the past decade several algorithms and software tools have been created to open an age of automated, high-throughput use of MS/MS. The most powerful and popular (such as Sequest [1] and Mascot [2]) use a

database of proteins to create theoretical mass spectra, and then directly compare the uninterrupted experimental spectra to find the best corresponding peptide in the database. This has an immediate disadvantage when attempting to identify spectra from proteins outside of the database, or proteins that may have been modified. Common modifications can be dealt with, but in a limited capacity. This and other issues have led to the push for *de novo*-based approaches, which use only the information from the spectrum to identify peptides.

In recent years, many automated *de novo* algorithms have been devised. The approach is generally made from a graph-theoretical stand-point (Lutefisk [3], SHERENGA [4]), in which ion peaks are nodes that are connected by edges corresponding to the masses of amino acid combinations. Results have generally been limited by the complexity of the problem, due to such factors as noise, unknown identities of ion peaks, incomplete peptide fragmentation, and high connectivity between peaks from different categories (e.g. abundance of cases when b-ion can be connected by precise amino acid mass to y-ion). Furthermore, the latest methods (SeqMS [5], Popitam [6], PEAKS [7]) are generally intended for very high-resolution MS/MS data.

The major problems with automated *de novo* of low-resolution data are the misidentification of ion peak types and the abundance of high intensity peaks (isotopes and neutral-loss peaks) that can distract the *de novo* process. The recently developed Probability Profile Method (PPM) can address these problems. This method uses the neutral-loss neighborhood around an ion peak to assign probabilities for belonging to several peak categories. It has been shown not only to reliably separate the b-ion and y-ion peaks from the rest of the spectrum, but to even discriminate between the ion types themselves. These probabilities reliably reflect actual probabilities, and

can be effectively used to filter most of the nonessential peaks, leaving the true b-ion and y-ion peaks for *de novo* searching.

We have developed a new *de novo* peptide identification approach based on PPM – PPM-Chain. The new approach opens new opportunities in peptide identification. It has very strong self-verification features built-in and may be especially powerful when the target protein database contains various types of modifications.

2. Method

The set of tandem mass spectra was produced from experiments during the characterization of the 54 ribosomal proteins of the *R. palustris* bacteria [8]. A 2D LC-MS-MS experiment was performed, with a total of ~20,000 MS/MS spectra produced from runs using 14 different mixture injections. Sequest was used to produce the initial identifications, using the entire *R. palustris* proteome as the database. A non-specific digest was assumed, and other Sequest parameters were left as default. For the reported results we have used only spectra identified as tryptic charge +2 peptides.

The PPM model used for this data incorporates the major neutral losses (H_2O , NH_3 , and CO), isotopic peaks, and complementary ion information to compute probabilities for three categories: b-ions - B, y-ions- Y, and R category the rest of the peaks. Probabilities (P_B , P_Y , and P_R) were computed for the most intense peaks in the 6 groups based on the estimated length of the parent ion. Lower intensity peaks were not predicted. A data set of 1585 spectra obtained on a standard protein mixture [9] was used as PPM training set.

3. Algorithm

In the first stage, each spectrum is processed in the following manner, in order to reduce complexity of the *de novo* process. All ion peaks with P_R above certain threshold (0.6 for our test) are removed from consideration, so that only peaks most likely to be b-ion or y-ion peaks are retained. This threshold can be increased if more confidence in peak selection is desired. The remaining peaks are then examined for the existence of a complementary peak, based on the parent ion mass. If a complement does not exist for a given peak, it is *mirrored* (i.e., a new peak is created), with probabilities assigned from the original peak (exchanging P_B and P_Y scores). This step allows the algorithm to reduce the effect of missing or low

intensity ion peaks. This resulting set of peaks is then passed to the second stage of the *de novo* algorithm.

In the second stage, we create a “chain” (graph), connecting peaks whose mass difference is approximately equal (within 0.4 Da for ion trap data) to a single amino acid or an amino acid pair. The addition of amino acid pairs in the construction of the graph further reduces the effect of missing ion peaks. Using a modified depth first search, all paths and sub-paths are extracted from the graph. These peak *chains* are assigned scores for the three peak categories based on the product of the probability scores of the peaks that compose the chain. Finally, the chains are sorted by length, and within equal length are sorted by P_B scores. This list of proposed b-ion chains is the list of potential partial peptides – peptide *de novo* tags.

For the database search, the chains are processed starting with the longest (and highest-scoring). Each chain is converted to a *mass-sequence-mass motif tag* construct (Fig 1).



Figure 1: In addition to the sequence tag, a candidate peptide must match the left (M_L) and right (M_R) flanking masses of the tag.

The *left* mass corresponds to the flanking amino acid residues in positions *before* the peak chain, and the *right* mass to positions *after* the sequence motif. The database is searched for all peptides (with options for specific or non-specific digestion) within specific tolerances for *left* and *right* masses. If no site satisfies all three conditions for the current chain, the search is continued with the next. If a chain matches more than one peptide in the database, the peptides are sorted by the number of ions matched in the spectrum.

Several other parameter choices are available for the database search. Two principal ones are threshold of the P_R probabilities and the minimum tag length (e.g., length 3 tags require connecting at least 4 peaks). It should be noted that performance of the algorithm is affected by these parameters. When we allow more peaks to be considered, the calculation times and rate of mistakes grow, but lower quality spectra can be used for *de novo* tag construction. The effects of decreasing the minimum tag length are similar.

4. Results and Discussion

We start the with a general discussion of the properties and advantages of *de novo* identification

methods based on our experience with PPM-chain, report preliminary results of PPM-chain identifications in comparison to Sequest and conclude with a summary that also reflects our future development plans.

First, *de novo* solutions have very strong “self-verification” properties. Consider, for example, a *de novo* peptide tag of 5 amino acids. Not only the tag sequence should be matched to the database, but also left and right masses should be satisfied (Fig. 1). Yet for tags of such length a sequence match may be a unique position in the bacterial genome. Under such condition left and right masses represent *independent* constraints; the mass conditions are independently supporting or rejecting already uniquely identified position.

Second, *de novo* identification has inherent flexibility in regard to the results, which is not possible in database look-up programs. For a given spectrum and given specifications for a *de novo* chain (e.g. 3 residues are set as a minimum length) PPM-Chain has three possible outcomes: (1) “no chain” – no satisfactory *de novo* tag could be constructed for the spectrum; (2) “no answer” – there are good *de novo* tags, but they do not conform to the available database; (3) “answer” – a satisfactory *de novo* tag is found and mapped to a protein in the database. In contrast, database look-up programs return the best match with an attached score, which slowly decrease from the confidently identified spectra toward definite identification failures. For these cases, the bad quality of the match (e.g., due to database error or inadequacy) is hard to distinguish from the mediocre informational content of the spectrum (e.g., due to poor fragmentation). This leads to “grey” area situations, where valuable information – often about unusual or interesting cases – can be irrecoverably lost.

Third, some of the algorithmic coups could be relatively straightforward for *de novo* methods, but are virtually impossible for the traditional approaches. Consider identification of the peptide containing an *unknown* post-translational modification. In the traditional way, one would have to test for 200 or more feasible modification compounds – hardly a productive approach as the sequence search space would expand on several orders of magnitude. In a *de novo* framework the unknown PTM could be potentially located using the “constraint excess” property of the constructed tags.

Consider a peptide containing an unexpected PTM (Fig. 2), where the constructed *de novo* tag does not include the PTM site. An attempt to match this tag to

the database likely will not return answer, as the computations of the “left” mass will not include PTM correction. But we still can use the tag itself and “right” mass to narrow down list of possible locations. Depending on the tag length, it could be even a unique solution. Then the investigation of the left mass “deficit” at those few locations may provide a good guess for the unexpected PTM, especially if the given PTM is used more than once in the investigated proteome.



Figure 2: The asterisk represents a site of an unknown PTM. Matching the tag and right mass may lead to identification of such sites.

Finally, computational performance of the *de novo* identification methods is governed by principles that are very different from the traditional identification algorithms. The laborious comparison between theoretical spectra and experimental spectra is the heart of the database look-up algorithms, and the performance typically scales linearly with the size of the search space. The disadvantage of this is that the search space grows exponentially in many situations (e.g., with the number of PTMs considered for each peptide). In a *de novo* approach, almost all work is done initially on the experimental spectra: peak labeling, forming the chains, chain scoring, etc. The need for the database comes very late in the process, involves extremely simple procedures, and could be skipped all together for spectra with too little (no chain) or too much (direct *de novo* identification) informational content.

We have explored results of *de novo* peptide identification for three separate spectral sets separated by the initial Sequest identification: high confidence (>3.2 X-correlation value), medium (between 2.2 and 3.2) and low (<2.2). For the high confidence subset, “no chain” outcome was obtained only for 21 spectra (1.4%) and out of 1263 “answer” spectra Sequest identification was confirmed for 1262 (99.9% precision of Sequest high confidence result reproduction). “No answer” outcome was observed for 216 cases (14%), and this fraction increases as we move to medium and low confidence subsets: 38% of “no answer” cases were recorded for both subsets. At the same time “answer” outcomes continued to be excellently aligned with Sequest identifications. The corresponding precision numbers were 99% and 96% for these two data subsets, respectively. The fraction of “no chain” cases grows sharply: 18% for medium

and 57% for low confidence sets, reflecting an absence of the differentiating information in many spectra belonging to these two categories.

As an example of the worth of the “no answer” results, our data set contained four spectra, which were matched by Sequest to the charge +2 peptide NNIHIVDLTQTVPLLHR from the 30S subunit of the ribosome. In two of the spectra, however, PPM-Chain discover a b-ion chain of 9 peaks that was mass-shifted by +1 Da. Due to the mass shift, these spectra were not matched to any protein, but they were marked for review. Under manual examination it became clear that the spectra contained a modification of asparagine to aspartic acid.

In summary, we conclude:

- With the existing technology, PPM-Chain tags can be constructed for a large majority of MS/MS spectra – and virtually for all high quality spectra.
- When *de novo* solution is compatible with the database, it is almost always the same as provided by Sequest. This conclusion confirms the high quality of the Sequest identifications in the cases when the expected peptides are present in the protein database.
- There is a significant fraction of spectra (~33% for medium X-correlation values, ~50% low X-correlation values) where PPM-Chain finds good *de novo* tags not compatible with anything in the target database. Some of these tags definitely reflect complex and interesting cases, where PTMs and point mutations are blocking the possibility of finding the right answers in the “plain vanilla” database searches.
- *De novo* identification methods have unique capabilities and far more potential for further algorithmic development than mature database look-up programs. In coming years *de novo* approaches will play an increasingly important role in the large-scale proteomic efforts.

5. Acknowledgments

Gregory Hurst and Michael Strader have provided all *R. palustris* experimental data used for this study. We also acknowledge many valuable and fruitful discussions with them and with Hayes McDonald, David Tabb, and Tema Fridman.

This work was funded by two US Department of Energy's Genomics: GTL programs: Sandia-ORNL “Carbon Sequestration in Synechococcus Sp.: From Molecular Machines to Hierarchical Modeling” (<http://www.genomes-to-life.org>) and ORNL-PNNL

Center for Molecular and Cellular Systems”. ORNL Laboratory Directed Research and Development Fund has supported our work on the PTM identification. ORNL is operated for DOE by UT-Battelle under contract number DE-AC05-00OR22725.

6. References

- [1] J.K. Eng, A.L. McCormack, and J.R. Yates, “An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in the Protein Database”, *J Am Soc Mass Spectrom* 1994, 5:976-989.
- [2] D.N. Perkins, D.J. Pappin, D.M. Creasy, and J.S. Cottrell, “Probability-based Protein Identification by Searching Sequence Databases using Mass Spectrometry Data”, *Electrophoresis*, 1999, 20:3551-3567.
- [3] J.A. Taylor and R.S. Johnson, “Sequence Database Searches via *de Novo* Peptide Sequencing by Tandem Mass Spectrometry”, *Rapid Comm. Mass Spect.*, 1997, 11:1067-1075.
- [4] V. Dancík, T.A. Addona, K.R. Clauser, J.E. Vath, and P.A. Pevzner, “*De Novo* Peptide Sequencing via Tandem Mass Spectrometry”, *J. of Comp. Bio.*, 1999, 6:327-342.
- [5] J. Fernandez-de-Cossio, J. Gonzalez, Y. Satomi, T. Shima, N. Okumura, V. Besada, L. Betancourt, G. Padron, Y. Shimonishi, and T. Takao, “Automated Interpretation of Low-energy Collision-induced Dissociation Spectra by SeqMS, a Software Aid for *De Novo* Sequencing by Tandem Mass Spectrometry”, *Electrophoresis*, 2000, 21:1694-1699.
- [6] P. Hernandez, R. Gras, J. Frey, and R.D. Appel. “Popitam: Towards New Heuristic Strategies to Improve Protein Identification from Tandem Mass Spectrometry Data”, *Proteomics* 2003, 3:870-878.
- [7] B. Ma, K. Zhang, C. Hendrie, C. Liang, M. Li, A. Doherty-Kirby, G. Lajoie, “PEAKS: Powerful Software for Peptide *De Novo* Sequencing by MS/MS”, *Rapid Comm. Mass Spec.*, 2003, 17:2337-2342.
- [8] M.B. Strader, N.C. VerBerkmoes, D.L. Tabb, H.M. Connelly, J.W. Barton, B.D. Bruce, D.A. Pelletier, B.H. Davison, R.L. Hettich, F.W. Larimer, and G.B. Hurst, “Characterization of the 70S Ribosome from *Rhodospseudomonas palustris* using an Integrated ‘Top-Down’ and ‘Bottom-Up’ Mass Spectrometric Approach”, accepted for publication in *J. of Proteome Research*, 2004.
- [9] A. Keller, S. Purvine, A.I. Nesvizhskii, S. Stolyar, D.R. Goodlett, E. and Kolker, “Short Communication: Experimental Protein Mixture for Validating Tandem Mass Spectral Analysis”, *OMICS*, 2002, 6:207-212.