AN ITERATIVE ALGORITHM TO QUANTIFY THE FACTORS INFLUENCING PEPTIDE FRAGMENTATION FOR MS/MS SPECTRUM

Chungong Yu†, Yu Lin†, Shiwei Sun, Jinjin Cai, Jingfen Zhang
Institute of Computing Technology,
Chinese Academy of Sciences, Beijing 100080, China

Zhuo Zhang, Runsheng Chen*
Institute of Biophysics,
Chinese Academy of Sciences, Beijing 100035, China

Dongbo Bu *
David R. Cheriton School of Computer Science University of Waterloo
Waterloo, Ontario, Canada N2L 3G1
Email: bdb@ict.ac.cn

In protein identification through MS/MS spectrum, it is critical to accurately predict theoretical spectrum from a peptide sequence, which heavily depends on a quantitative understanding of the fragmentation process. To date, widely used database searching methods adopted a simple statistical model to predict theoretical spectrum, yielding a spectrum deviating significantly from the practical spectrum for some peptides and therefore preventing automated positive identification. Here, in order to derive an improved predicting model, we proposed a novel method to automatically learn the factors influencing fragmentation from a training set of MS/MS spectra. In this method, the determining of factors is converted into an optimization problem to minimize an objective function that measures the distance between experimental spectrum and theoretical one. Then, an iterative algorithm was proposed to minimize the non-linear objective function. We implemented the methods and tested them on experimental data. The examination of 1451 spectra is in good agreement with some known knowledge about peptide fragmentation, such as the tendency of cleavage towards the middle of peptide, and Pro's preference of N-terminal cleavage. Moreover, on a testing set containing 1425 spectra, comparison between predicted and practical spectra generates a median correlation of 0.759, showing this method's ability to predict a "realistic" spectrum. The results in this paper help to an accurate identification of protein through both database searching and *de novo* methods.

1. INTRODUCTION

A major goal of proteomics is to study biological processes comprehensively through the identification, characterization, and quantification of expressed proteins in a cell or a tissue. Tandem mass spectrometry(MS/MS) has emerged as a powerful tool for sensitive high-throughput identification of proteins[1, 2]. In an experiment, proteins of interest are selected, and digested by enzyme such as trypsin, and then the resultant peptides are separated in the mass analyzer according to their mass to charge ratio ($m/z - value$). In a single experiment, multiple copies of the same peptide are fragmented into many charged fragments, and the fragments retaining the ionizing charge after CID have their $m/z - value$ measured, the aggregate of which forms MS/MS spectrum [16].

Predicting theoretical spectrum accurately from a peptide sequence lies at the core of protein identification, especially for the database searching methods. Most database searching methods start with constructing a theoretical spectrum for each peptide in a protein database, followed by a comparison of theoretical spectrum with experimental one using an effective scoring functions[1, 2, 11, 12, 14–16, 28]. The peptides with the highest score would be reported as potential solutions. Lacking a complete understanding of the fragmentation process, the widely used algorithms, such as Sequest[10] and Mascot[13], adopted a simple statistical model to predict theoretical spectrum, which assumes that cleavage will occur at peptide bonds in a uniform manner, regard-

**To whom the correspondence should be addressed. † These two authors contributed equally to this paper.

less of some important influencing factors such as position of amino acids, types of bond, etc. Though succeeds in general cases, this simple model produces theoretical spectrum deviated significantly from experimental one for some peptides, leading to low or insignificant scores, and thus preventing positive protein identification.

Furthermore, the *de novo* identification approaches could also benefit from an accurate prediction of the theoretical spectrum. Many studies haver been conducted to identify protein without dependence of a protein sequence database. Sakurai adopted a method to enumerate all possible sequences and compare each one with the spectrum[5], and *prefix pruning* technique was proposed to speed up the search[6, 2]. An alternative strategy is *spectrum graph*, which formulates the spectrum into a graph, and attempts find the longest path in the graph[28, 15, 4, 8]. To overcome the shortcomings of *spectrum graph* arisen by missing and mixed peaks, PEAKS employs a sophisticated dynamic programming method[9]. In addition, Zhongqi Zhang proposed a method to combine a divide-and-conquer algorithm with a spectrum simulation[7]. Typically, a *de novo* sequencing method computes candidate sequences first, and then evaluates them by comparing the experimental spectrum with the predicted spectra. Hence, an accurate prediction of theoretical spectrum is not only useful to the database search approach, but also useful to *de novo* methods.

To accurately predict theoretical spectrum, it depends on a quantitative understanding of the fragmentation process occurring in mass spectrometry, which remains a challenge because of the following reasons: First, fragmentation is a stochastic process governed by complicated physical and chemical rules and affected by many factors such as position and identity of amino acids, types of bonds, etc. Moreover, it's also unclear that to what extent each factor affects the fragmentation process. Secondly, isotopic atoms, neutral losses, post-transcription modification and measuring error always result in a peak deviation from its expected position. This paper addresses an attempt to quantify the factors influencing fragmentation.

1.1. Related Work

To predict theoretical spectrum, except for the promising chemical kinetic model to simulate fragmentation process[24], several studies have been conducted to develop a statistical predicting model. Dancik et al. introduced an automatic tool-*offset frequency function*- to learn the ion types tendency and intensity threshold from the experimental spectra[15, 16]. J.R. Yates III et al. attempted to identify statistical trend in spectrum peak intensities and put them into the chemical context. F. P. Roth and S.P. Gygi applied probability decision tree approach to distinguish the important factors from a total 63 peptide and fragmentation attributes[29]. Another interesting method to determine the factors influencing fragmentation is a linear model proposed by F.Schutz[21]. In this method, F.Schutz fitted a linear model to spectrum, in which the influence of some specific amino acid and their position in the peptide are reflected. Moreover, the linear model also shows ability to accurately predict theoretical spectrum.

The linear model has some difficulties. In this model, the preference for cleavage at a bond is represented as the sum of the influence of C-terminal residue and of N-terminal residue. This assumption is strict since it implies that Xaa-Pro bond has an enhanced cleavage than any Xaa-Yaa bond regardless of which amino acid Xaa is, which is inconsistent with the observation that Xaa-Pro bond's cleavage is hindered when Xaa is Gly or Pro[21]. Hence, it is more reasonable to consider the cleavage preference in bond's manner rather than the sum of residues influence. In this paper, we present a novel model to overcome these difficulties.

1.2. Our Contribution

Our contributions within this paper are as follows:

1. We introduce a novel statistical model to determine the important factors that influence the global fragmentation. Following the well-known "mobile proton" hypothesis, our model accounts for influence of amino acids position, cleavage preference for a bond in a more reasonable manner.

2. We used this model to predict theoretical spectrum for a test set and made comparison with practical ones. Using the derived quantitative pa-

rameters, theoretical spectrum could be generated by simulating the tendency of cleavage towards middle, preference for N-terminal or C-terminal cleavage for a specific bond. Experimental results show that this model could predict a more 'realistic' spectrum.

We implemented these algorithms into an open source package PI (Peptide Identifier, downloadable freely from http://www.bioinfo.org.cn/MSMS/) and trained PI on several sets of spectra from ISB[18]. As a result, we rediscovered some known knowledge about peptide fragmentation, such as the tendency of cleavage towards the middle of peptide, and Pro's preference of N-terminal cleavage. Moreover, PI could predict accurate theoretical mass spectrum from a peptide sequence.

2. METHODS

2.1. Fragmentation Model

"Mobile proton" hypothesis is one of the widely accepted tenets of the peptide fragmentation. In this model, the ionizing protons on the peptide migrate to an amide carbonyl oxygen along the peptide backbone, resulting in the cleavage of its N-terminal peptide bond and the production of a b-ion or y-ion depending on N-terminus or C-terminus retains the charge, respectively. Occasionally, an a-ion is generated from a b-ion by losing of carbon monoxide. Other possible backbone ions, such as c, x, z ions, are not typically generated under low energy collision-induced dissociation conditions[22, 23, 26].

Several factors have significant affection on the fragmentation process since fragmentation in spectrometry is a stochastic process governed by the physical and chemical properties of a peptide and the collision dynamics. Some of the factors are listed as follows: First, there is a tight relationship between peak intensity and the relative position of cleavage site, that is, fragmentation occurs more often in the middle of peptide than that at ends[17, 22, 25]. Second, individual amino acid has different preference for which of the two adjacent amide bonds(N-terminal or C-terminal) may break. For example, it was reported that Pro has a strong C-bias cleavage[22]. Other factors, such as excitation method, charge state of ions, etc, also have influence on the fragmentation process[26]. Hence, identifying the significant factors is important to improve theoretical spectrum predicting. This paper attempts to quantify the factors influencing fragmentation process under the "mobile proton" hypothesis.

2.2. Influence of Cleavage Site and Peptide Bonds

In the mobile proton fragmentation model, it was reported that the proton attachment depends partly on the relative affinities and the position of amino acids[31]. Let $A(a_i)$ denote the relative proton affinities of amino acids a_i, $f(j)$ denote the influence of an amino acid at position j on proton affinities. For a peptide bond $< a_i, a_j >$, let $B(a_i, a_j)$ denote the relative possibility that the bond breaks when a proton migrates onto a_j. Thus, for peptide $P^{(i)} = p_1^{(i)} p_2^{(i)} ... p_L^{(i)}$, the number of cleavage events of the $j-th$ bond, denoted as C_{ij}, can be estimated to be proportional to $f_j * A(p_{j-1}^{(i)}) * B(p_{j-1}^{(i)}, p_j^{(i)})$. Hence, minimizing the difference between the actual value C_{ij} and its estimation will assign reasonable value for $A(a_i), f(j)$ and $B(a_i, a_j)$. For the sake of simplicity, we define $C(a_i, a_j) = A(a_j) * B(a_i, a_j)$, which is a measurement of the relative possibility that a proton migrates onto a_j and therefore results in the cleavage at the bond $< a_i, a_j >$. Hence, all the above parameters could be determined through solving the following non-linear programming problem on a training peptide set $P^{(1)}, P^{(2)}, ..., P^{(K)}$ with same length $|P^{(1)}| = |P^{(2)}| = ... = |P^{(K)}| = L$

$$min \sum_{i=1}^{K} \sum_{j=2}^{L} ((\alpha_i * f_j * C(p_{j-1}^{(i)}, p_j^{(i)}) - C_{ij})^2 \quad (1)$$

$$\begin{aligned} s.t. \quad & \sum f(i) = 1, f(i) \geq 0, \\ & \sum_{j=2}^{L} \alpha_i * f_j * C(p_{j-1}^{(i)}, p_j^{(i)}) = 1. \\ & \alpha_i \geq 0, C(a_i, a_j) \geq 0 \end{aligned}$$

Here, α_i is an auxiliary variable, a scale factor to meet $\sum_{j=2}^{L} \alpha_i * f_j * C(p_{j-1}^{(i)}, p_j^{(i)}) = 1$. The objective function is the sum of square of the difference between the theoretical and experimental intensities.

We tried some classical non-linear programming methods but failed to find optimal solution in reasonable time for the high rank of restriction formulas.

Here, an iterative method was adopted to solve this problem. The method is based on the fact that the above formula could be reduced into a least square problem if two of the three types of variables, e.g., α_i and $C(a_i, a_j)$, were fixed, while only one types, e.g., $f(j)$ was chosen as variables.

At first, all the variables are assigned with random initial value. Each iteration loop contains three steps corresponding to one of $f(j), \alpha_i$ and $C(a_i, a_j)$ was chosen as variables. For example, if $f(j)$ was chosen as variable while $C(a_i, a_j)$ and α_i was fixed with the current value, a classical optimization algorithm is called to solve the least square problem over variable $f(j)$. So do α_i and $C(a_i, a_j)$. The iteration loop is repeated until the value of the objective function does not change.

It can be easily proved that the iterative algorithm must converge at last. The proof is based on the fact that the value of the objective function is non-negative and decreases monotonously at each step. In practice, the algorithm always converges to a fixed point after no more than 10 iteration loops, and experiments reach the same fixed point on different random initializations. In addition, the characteristic of the formula guarantees that only positive solution would be found.

The algorithm to minimize the distance function is given in Fig. 1.

3. EXPERIMENTAL RESULTS

3.1. Datasets

A public online spectrum set, AB_IP, from ISB [18] was used to test our algorithm, which contains spectra generated through shotgun analysis of proteins from human K562 cells. We restricted our analysis to doubly charged, 'mobile' peptide for this proof-of-concept experiment. The spectrum set are randomly divided set into two parts, a training set (1451 matches) and a testing set (1425 matches). (See supplementary material http://www.bioinfo.org.cn/MSMS/).

3.2. Position and Bond's Influence on Cleavage

Relationship between fragmentation probability and cleavage site The training set was categorized into several subsets with respect to the length of peptide. On some subsets containing peptides with same length $L = 7, 8, 9, 10, 11, 12, 13, 14, 15$, the relationship between amino acids position and affinity ability are calculated. Fig. 2 shows the cases where $L = 9, 11, 13, 15$. Fig. 2 demonstrates that fragmentation occurs more often towards the middle of a peptide than at its ends, which is consistent with the observation given previously[21, 29, 17]. Moreover, Figure 1 shows that the shorter the peptide, the more asymmetric the curve, which supports the observation that fragmentation near the N-terminus differs significantly from that at other sites[26].

Cleavage Preference of Peptide Bonds

The statistical results of the preference of fragmentation at all the 400 peptide bonds are calculated (See supplementary material at http://www.bioinfo.org.cn/MSMS/).

To justify the motivation of this work, we compared the cleavage preference of Xaa-Pro bonds with that of Xaa-Trp bonds(See Figure 3a). Figure 3a shows that in general cases a Xaa-Pro bond has a higher tendency to cleavage than the corresponding Xaa-Trp bond; however, a Xaa-Pro bond is relatively hard to cleavage when Xaa is Gly or Pro since cleavage is hindered in these two cases[21]. This phenomenon cannot be reflected correctly if simply measuring the cleavage preference as the sum of residue influence. Hence, it is more reasonable to consider the cleavage preference in bond's manner.

Examination of the preference data is in good agreement with knowledge already known to mass spectrum experts. First, some amino acids prefer cleavage at N-terminus over C-terminus bond. For example, it is well-known that cleavage at Pro's N-terminus is preferred than that at C-terminus because attacking of the adjacent carbonyl oxygen at the electropositive carbon is hindered due to the molecular structure of Pro[22]. Figure 3b shows that Xaa-Pro always has a higher possibility of fragmentation than the counterpart Pro-Xaa, supporting that Pro tends to cleavage at its N-terminal than C-terminal bond[21]. It was also reported that cleavage at His-Xaa bonds are much more often than others[23, 22]. As an example, Figure 3c shows a comparison between His-Xaa bonds and Asn-Xaa bonds, which is consistent with the above observation. Second, fragmentation of the Xaa-Pro bond is encouraged when Xaa is Ile(0.020), His(0.016) or Trp(0.014), while is hindered when Xaa is Gly(0.0018) or Pro(0.0009) (See Figure 3a). In conclusion, the above results have strongly supports from the"mobile proton model", i.e., the more basic the residue, the more large the affinity of proton, and then the more facile the fragmentation.

3.3. Predicting Theoretical Spectrum

For a given peptide P, theoretical spectrum could be predicted by simulating the fragmentation process following the mobile proton model. That is, the number of cleavage events of the $j-th$ bond can be estimated to be proportional to $f_j * A(p_{j-1}) * B(p_{j-1}, p_j)$. Here, we roughly assumed that a b ion or y ion would be formed by a cleavage event with equal probability since the 'effective' temperature is unknown[26].

Two examples were shown in Fig. 4, one for "DPLLLAIIPK" containing Pro since Pro has a unique fragmentation preference, the other for "DAGTIAGINVMR". Each predicted spectrum was plotted on the lower axis below, showing reasonable similarity to their experimental counterpart (Correlation coefficient are 0.80 and 0.81, respectively).

On the test set containing 1425 pairs of spectra and its corresponding peptide, theoretical spectrum were predicted and compared with experimental spectra. The median correlation between predicted and the practical spectra is 0.759, showing that this method could predict a "realistic" spectrum.

4. CONCLUSION AND DISCUSSION

Prediction of theoretical spectrum accurately is important to database searching methods, however, this prediction is in great need of a quantitative understanding of the fragmentation process. Here, we proposed a non-linear programming methods to estimate the factors influencing the fragmentation. We applied this algorithm to real data, and successfully obtained many biological features which also supported by some known rules of fragmentation, demonstrating the efficiency of the methods. And our simulated mass spectra are reasonably similar to their experimental counterparts. Currently, we have not taken charge +3 and non-mobile peptides into account, and adopted a rough assumption that b-ion and y-ion are produced with equal probability by a cleavage event. The influence of distant amino acids on fragmentation is also not considered in our model. How to incorporate those factors in PI remains an open problem.

ACKNOWLEDGMENT

This work was supported by National Sciences Foundation of China under grants 60496320, 30500104 and 30570393, National Key Basic Research and Development Program under grants 2002CB713805 and 2003CB715900, and opening task of Shanghai Key Laboratory of Intelligent Information Processing Fudan University No. IIPL-04-001. Dongbo Bu was partly supported by NSERC operating grant OGP0046506.

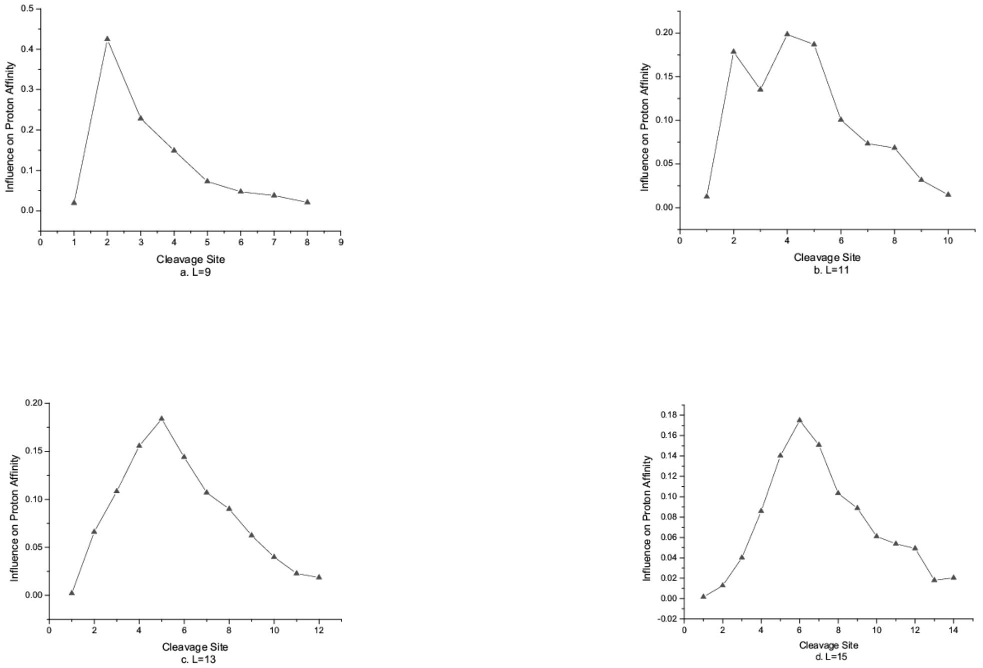
References

1. Zhu, H.; Bilgin, M.; Snyder, *M. Annu Rev Biochem 2003*, 72, 783-812.
2. Yates, J. R., *3rd J Mass Spectrom 1998*, 33, 1-19.
3. Aebersold, R.; Goodlett, *D. R. Chem Rev 2001*, 101, 269-295.
4. Hines, W.M., Falick, A.M., Burlingame, A.L., and Gibson, B.W. *J. Am. Sco. Mass. Spectrom. 1992* 3,326-336.
5. Sakurai, T., Matsuo, T., Matsuda, H., and Katakuse, I. *Biomed. Mass Spectrum 1984* 11(8), 396-399.
6. Siegel, M.M., and Bauman, N. *Biomed. Environ. Mass Spectrum 1988* 15, 333-343.
7. Zhongqi Zhang. *Anal. Chem. 2004* 76,6374-6383
8. Bartels, C. *Biomed. Environ. Mass Spectrim 1990* 19, 363-368.
9. Bin Ma, Kaizhong Zhang, and Chengzhi Liang. *Journal of Computer and System Sciences 2005* 70: 418-430.
10. Yates, J. R., 3rd; Eng, J. K.; McCormack, *A. L. J Am Soc Mass Spectrom 1994*, 5, 976-989.
11. *Sonar http://65.219.84.5/ProteinId.html.*
12. *MOWSE http://www.hgmp.mrc.ac.uk/Bioinformatics/Webapp/mowse/mowsedoc.html.*
13. Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, *J. S. Electrophoresis 1999*, 20, 3551-3567.
14. Zhang, N.; Aebersold, R.; Schwikowski, *B. Proteomics 2002*, 2, 1406-1412.
15. Dancik, V.; Addona, T. A.; Clauser, K. R.; Vath, J. E.; Pevzner, *P. A. J Comput Biol 1999*, 6, 327-342.

16. Bafna, V.; Edwards, *N. Bioinformatics 2001*, 17 Suppl 1, S13-21.
17. Havilio, M.; Haddad, Y.; Smilansky, *Z. Anal Chem 2003*, 75, 435-444.
18. Resing et al. 2004, *Anal Chem.* 2004 Jul 1;76(13):3556-68.
19. Pevzner, P. A.; Mulyukov, Z.; Dancik, V.; Tang, *C. L. Genome Res 2001*, 11, 290-299.
20. Keller, A.; Purvine, S.; Nesvizhskii, A. I.; Stolyar, S.; Goodlett, D. R.; Kolker, *E. Omics 2002*, 6, 207-212.
21. Schutz, F.; Kapp, E. A.; Simpson, R. J.; Speed, *T. P. Biochem Soc Trans 2003*, 31, 1479-1483.
22. Tabb, D. L.; Smith, L. L.; Breci, L. A.; Wysocki, V. H.; Lin, D.; Yates, J. R., *3rd Anal Chem 2003*, 75, 1155-1163.
23. Wysocki, V. H.; Tsaprailis, G.; Smith, L. L.; Breci, *L. A. J Mass Spectrom 2000*, 35, 1399-1406.
24. Zhang. Z. *Proc. 50th Am So. Mass Spectrom. Orlando. Fl. 2002.* Paper TPE-126
25. O'Hair, *R. A. J Mass Spectrom 2000*, 35, 1377-1381.
26. Paizs, B.; Suhai, *S. Mass Spectrom Rev 2004* 5, 103-113
27. Chen, T.; Kao, M. Y.; Tepel, M.; Rush, J.; Church, *G. M. J Comput Biol 2001*, 8, 325-337.
28. Yunhu Wan, Ting Chen, *RECOMB 2005, pp. 342-356, 2005*
29. J. E. Elias, F.D. Gibbons, O. D. King, F. P. Roth, S. P. Gygi, *Nature BioTechnology 2004* Vol. 2 Num. 2, Feb. 2004
30. Huang, Y. Wysocki, V.H. Tabb, D.L and Yates, J.R. *J. Am Soc Mass Spectrom 2002* 219, 233-244
31. M. J. Nold, B. A. Cerda, and C. Wesdemiotis *J. Am Soc Mass Spectrom 1999* 10,1-8

```
Algorithm to Minimize Distance Function
Input: $K$ pairs of peptides and tandem mass spectra $\{(P_1, S_1), (P_2, S_2), ..., (P_K, S_K)\}$, $|P^{(1)}| = |P^{(2)}| = ... = |P^{(K)}| = L$.
Output: Bond's preference of cleavage $C(a_i, a_j)$ for each bond $< a_i, a_j >$, Position's influence on cleavage $f(j), j = 1, 2, ..., L$ ;
1. Initializing $C(a_i, a_j)$ and $f(j)$ randomly;
2. Optimize formula (1) over $\alpha_i$ with $C(a_i, a_j)$ and $f(j)$ holding the current value;
3. Optimize formula (1) over $C(a_i, a_j)$ with $\alpha_i$ and $f(j)$ holding the current value;
4. Optimize formula (1) over $f(j)$ with $C(a_i, a_j)$ and $f(j)$ holding the current value;
5. Repeat step 2-4 until the objective function value converges.
6. Output $C(a_i, a_j)$ and $f(j)$.
```

Fig. 1. Algorithm to Minimize Distance Function



a. L=9

b. L=11

c. L=13

d. L=15

Fig. 2. Relationship between Proton Affinity and Cleavage Site. (L=9,11,13,15)

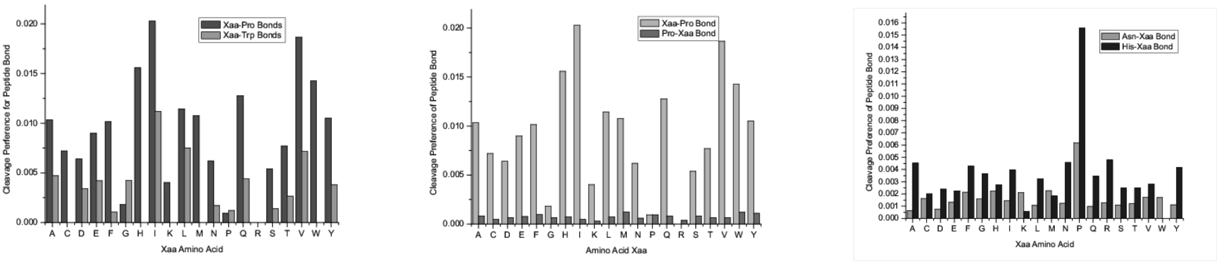


Fig 3a

Fig 3b

Fig 3c

Fig. 3. Bonds Preference for Proton Affinity and Cleavage

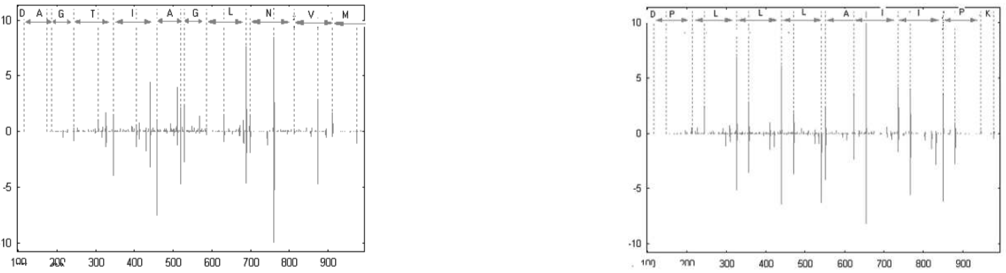


Fig. 4. Simulated and Experimental Spectra for 'DAGTIAGINVMR' and 'DPLLLAIIPK'