Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Environmental Sciences 8 (2011) 483 – 491

Procedia Environmental Sciences

ICESB 2011: 25-26 November 2011, Maldives

The Prediction of Peptide Charge States for Electrospray Ionization in Mass Spectrometry

Hui Liu[a,b], Jiyang Zhang[a,b], Hanchang Sun[a,b], Changming Xu[a,b], Yunping Zhu[b], Hongwei Xie[a*]

[a]*College of Mechatronic Engineering and Automation, National University of Defense Technology, Changsha 410073, China*
[b]*State Key Laboratory of Proteomics, Beijing Proteome Research Center, Beijing Institute of Radiation Medicine, Beijing 102206, China*

Abstract

Electrospray ionization in proteomic mass spectrometry is one of important methods of soft ionization or vaporization for the widest range of polar biomolecules. The number of charges attached onto peptides in ESI can extend the detection limit of mass spectrometer and can be used to estimate the location of mass signal of peptides in the mass spectrum. We present an approach to predict the charge state of peptides according to the composition of peptides, and it has been tested and verified on different datasets. It shows sufficient performance.

Keywords: prediction; peptide charge state; electrospray ionization; mass spectrometry;

1. Introduction

Mass spectrometry (MS) is the most important and comprehensive tool in proteomics research. MS based proteomics has become an indispensable experimental method and have application in molecular and cellular biology, even system biology [1-2]. The rapid growth of protein MS in the past two decades has benefited from the important developments in experimental methods, instrumentation, and data analysis approaches. Many of the improvements in MS are the direct consequence of the introduction of the soft ionization technologies that make it possible to analyze polar and thermally labile biomolecules without prior derivatization. Two soft ionization techniques, matrix-assisted laser desorption ionization

* Corresponding author. Tel.: 86-731-84576311
E-mail address: xhwei65@nudt.edu.cn

1878-0296
doi:10.1016/j.proenv.2011.10.076

(MALDI) and electrospray ionization (ESI), explored the way for the modern MS proteomics. Developments in soft ionization methods have driven MS manufactures to build instruments with increased mass range, higher routine resolution, mass accuracy and lower cost [3-4].

ESI technique is operated by high voltage, 2-6kV, to form a strong electric field between the emitter at the end of the separation pipeline and the inlet of the mass spectrometer. An excess of charge will be created at the tip of a capillary containing the analyte solution by means of the electrostatic field, and charged droplets will be emited from the capillary as a spray and travel at atmospheric pressure. Gas phase ions are then transported through different vacuum stages to the mass analyzer, and at last the detector [5].

As early as in 1910s, John Zeleny had observed the electrospray dispersion of liquid with the electroscope[6-7], and Chapman first investigated the ions evaporation from the surface of water at atmospheric pressure in 1937/1938[8]. There are a number of reports about the details of the electrochemical nature of the electrospray process [5, 9-11]. Several physical models explore the formation of ESI ions [12-14].

As for the application of ESI-MS, two models, the charged residue mechanism (CRM) and the ion evaporation mechanism (IEM), were developed to describe the actual mechanism which makes the gas ions ultimately detected by the mass spectrometry. Dole et al.[15-16] proposed the CRM that evaporation of solvent from a droplet would increase the surface charge density until it reached the Rayleigh limit[17] at which Coulombic repulsion and surface tension become comparable, then the instability would emerge and lead to an emission of the parent droplet into a plurality of smaller offspring droplets that would continue to evaporate. A series of such explosions would ultimately give rise to droplets so small that each one just contains a single solute molecule. But the interpretation of their results remains somewhat amphibolous because of problems in the method of mass analysis they used [18]. Iribarne and Thomson [12] proposed their IEM model for ion formation. The IEM holds that a sequence of evaporation and Coulombic explosions lead to droplets with small radii and charge high densities that the electrostatic filed at a droplet surface would be sufficiently intense to lift solute ions into the gas or vapor.

J. Fernandez de la Mora [19] compared the maximum charge zmax on diverse eletrospray ions of globular proteins keeping their native structure to the Rayleigh charge zR, and zmax was between 65% and 110% of zR that strongly support Dole's CRM. The protein is charged by the excess positive ions such as NH4+, and the number of positive ions can be predicted on the basis of CRM. The proteins can hold all of the protons provided that must have a sufficient number of basic side chains located at the surface of the protein. It is found that most proteins have sufficient basic sites to retain the charge [14]. But there is an observed difference in charge distribution for molecules of the same species of different configuration, and the difference depends on whether in a folded molecule the charges have access to the sites [18].

Mass analysis of ions by ESI source indicates ESI can substantially extend the mass range of mass analyzer [20], and the number of charges of the ions is an important parameter that decides whether or not the ions could be detected by mass spectrometer, especially for biological macromolecules. The number of charges of the ions in the identification of peptide in MS-based proteomic experiment is affected by the experimental conditions, like instruments, the voltage applied, the concentration and the flow rate of the solution etc., and obviously, it will be strongly controlled by the properties of peptides in the process of ESI, such as the count and species of amino acids and the conformation of peptides.

Here we propose a model that estimates the number of charges held by peptides in MS-based protein identification. Based on this, we could predict that whether or not one peptide could be in the mass range of mass spectrometer and measured in the experiment, and it is helpful that determine the m/z ratio in the mass spectrum and assist to validate the result of database research.

2. Material and Method

2.1. Peptides dataset

We use two datasets in this work; the details of datasets are described as follows:

Dataset 1

The first dataset used in this work is available from state key laboratory of proteomics in Beijing and Beijing Proteome Research Center (BPRC) that contains recorded MS/MS spectra using LTQ MS platform with ESI-RPLC. MS/MS database searches are carried out using SEQUEST[21]. The dataset 1 were extracted from the database search results that include 215321 reliably identified peptide sequences of human liver proteins, and the corresponding charge state of each peptide. There are 3 charge states in the dataset, and the distribution of charge states in the peptides dataset is shown in Table 1:

Table 1. The distribution of peptides for different charge states in the dataset1

Charge state	1	2	3	Total
Amount	8604	189867	16850	215321
Percentage%	3.9959	88.1786	7.8255	100

Because of the repeated occurrence of the same peptide sequence in the database search, it is necessary to get rid of the redundant data from the dataset. The number of distinct peptide sequences in the dataset after the redundant is filtered is 13273. The distribution of the nonredundant data in the dataset 1 is shown in Table 2:

Table 2. The distribution of peptides for different charge states in the nonredundant dataset

Charge state	1	2	3	Total
Amount	969	12277	2142	15388
Percentage%	6.298	79.783	13.92	100

Note: The 15388 (more than 13273) entries include some peptides with more than one charge state，and these peptides are counted repeatedly.

In addition, there is a trait in the dataset that one peptide may correspond to more than one charge state because of the interaction of peptides and the experimental settings and accidental factors. That means a peptide will hold two or more charges, and these data will influence the classification of the dataset. The statistic is shown in Table3:

Table 3. The number of peptides that hold two or more charges for the identical peptide

Charge state	1 and 2	2 and 3	1, 2 and 3	Total
Amount	834	1209	36	2079

Dataset 2

The dataset 2 used in this work is available from Peptide Atlas (http://www.peptideatlas.org/repository/) and the sample accession is PAe000324[22] that includes 57963 redundant and 4987 distinct peptide identifications of yeast protein using LTQ MS platform with ESI-LC. The number of peptides that have multiple charge states is 798. The details of the dataset 2 are shown in Table 4, 5 and 6.

Table 4. The distribution of peptides for different charge states in the dataset2

Charge state	1	2	3	Total
Amount	2992	43747	11224	57963
Percentage	0.05162	0.75474	0.19364	100

Table 5. The distribution of peptides for different charge states in the nonredundant dataset

Charge state	1	2	3	Total
Amount	499	4039	1260	5798
Percentage %	8.61	69.66	21.73	100

Note: The 5798 (more than 4987) entries include some peptides with more than one charge state，and these peptides are counted repeatedly.

Table 6. The number of peptides that hold two or more charges for the identical peptide

Charge state	1 and 2	2 and 3	1，2 and 3	Total
Amount	278	507	13	798

2.2. Model descriptions

The model is used for the estimation of the charge state for the peptides in the process of ESI. As described above [14], the charge observed in the mass spectrum depends not only on the charging according to the CRM mechanism, but also on the ability of the biomolecule to maintain the charge through the process of ESI. That means the property of the component of biomolecule, for example, peptides, is an important factor to determine the charge states observed in the mass spectrum, especially the basic residues of the peptide [23-24], thus we use a multivariate linear regression to model the relationship between the amino acid sequence of a peptide (i.e. explanatory variables) and the charge state of the peptide (i.e. response variable). We can use this model to fit the observed data. The model is described as the expression (1):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \cdots \beta_{20} X_{20} + \varepsilon \quad (1)$$

Where: β_0 is a regression constant and $\beta_1, \cdots\cdots \beta_{20}$ are the coefficients respectively for 20 kinds of amino acids that constitute a peptide; $X_1, X_2, \cdots\cdots X_{20}$ are the explanatory variables that denote the number of 20 amino acids within a peptide respectively; ε is an error term; Y is the charge state of a peptide.

Because some peptides may involve more than one charge state, this will make a certain amount of overlaps between the data of the different charge state that originated from the same peptide sequence. This introduces a question that how to estimate the probability which charge state of a peptide from the overlapped data could be classified correctly. Here we use multiply normal distribution to fit the response value of multivariate linear regression for each charge state data, thus we could get the probability of the response variable that belongs to the specific charge state. The expression of multiple normal distributions is shown as in (2):

$$P(X = x \mid j) = \sum_{i=1}^{3} \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}} \quad (2)$$

Where:

μ_i and σ_i are the mean and the variance of distribution i, and i =1, 2, 3; j is the charge state, and j=1, 2, 3; x is the response value of different charge state.

$P(X = x \mid j)$ is just a conditional probability under the condition of the known charge state, if we know the response value from the multiple linear regression we can deduce the probability based on the conditional probability $P(X = x \mid j)$ and the probability of different charge state of the peptide in the dataset. The expression is shown as in (3):

$$P(j \mid X = x) = \frac{P(X = x \mid j) * P(j)}{\sum_{i=1}^{3} P(X = x \mid i) * P(i)} \quad (3)$$

Here x is the response of the regression as above, and $P(j)$ is the probability for the presence of different charge states of the peptides in the dataset.

3. Results and Discussion

3.1. 5 folds cross-validation (CV)

A 5-fold CV study is performed to determine the ability of the model to predict the charge state of the peptide within the dataset 1. The dataset 1 is divided into 5 subsets or folds of roughly equal size. Among the 5 subsets, a single subset is retained as the validation set for testing the model, and the other 4 subsets are used as the training set for fitting the model. The CV process is then repeated 5 times, with each of 5 subsets used exactly once as the validation set. The 5 results from 5 rounds of CV are then averaged to produce a single estimation. The advantage of this method over repeated random sub-sampling is that all observations are used for both training and validation, and each observation is used for validation once. We introduce the fitting of the model.

Firstly, the multivariate linear regression is used to model the training set, but there are the overlapped response values due to the fact that the same peptide sequence may hold different charge states. The distribution of the response values for different charge states is shown in Figure 1(up). The overlapped response values make it difficult to decide the charge state of a peptide if the response value of this peptide sequence lies within the overlapped region. The multiple normal distribution is used to model the distribution of the response values of different charge states and calculate the probability of different charge states of the peptide sequence. The result of the multiple normal distribution fitting is shows in Figure 1(below).

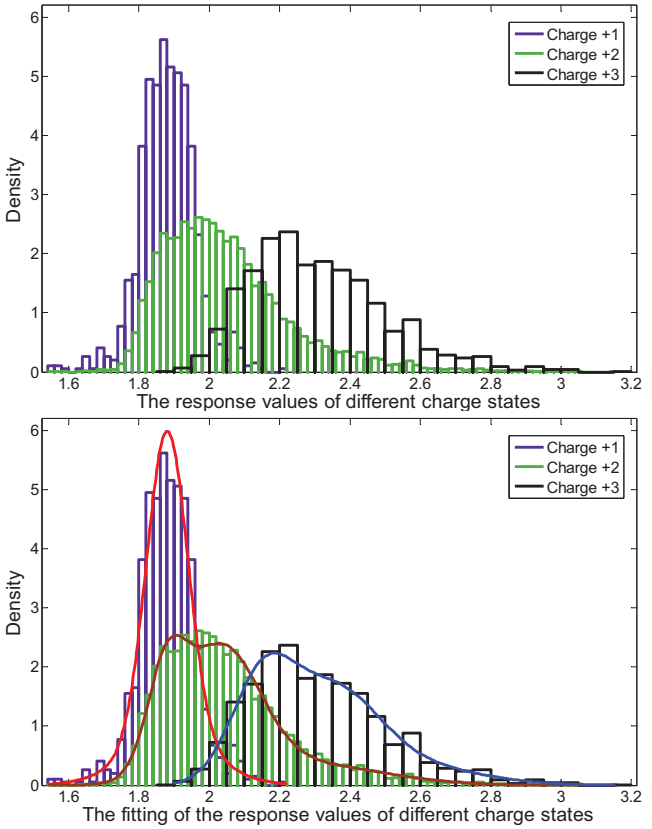

Fig. 1. The distribution of the response values for different charge states (up) and the fitting of response values by the multiple normal distribution (below).

The prediction results of model using 5 folds CV is shown as blow in Table 7. The average accuracy of the prediction is 96.89%, error rate is just 3.11%. The results demonstrate the performance of model is very well for the prediction of the charge state of the peptide in the dataset 1.

Table 7. The results of 5 folds cross-validation

Test set	Test1	Test2	Test3	Test4	Test5	Average
Peptide amount	2991	2985	2998	2982	2997	2990
Correct	2901	2907	2893	2895	2891	2897
Accuracy %	97	97.39	96.5	97.08	96.46	96.89
Error	90	78	105	87	106	93
Error rate%	3	2.61	3.5	2.92	3.54	3.11

3.2. Validation across different species datasets

In this section, we validate the model across species datasets. The model generated from the identified

peptides data of human liver protein is used to predict the charge state of the peptide from yeast protein, and vice versa. The result of predictive accuracy is shown in Table 8. Despite of datasets with different peptide composition, the method has a good prediction accuracy. The sample size maybe is an important factor to get a constant accuracy, and a large sample size is beneficial to the reliability and stability of the accuracy.

Table 8. The predictive accuracy within or across species dataset

Training species	Test species	
	Human liver	Yeast
Human liver	96.08%	90.92%
Yeast	87.89%	89.91%

3.3. The length and basic sites of the peptide for different charge states

The basicity of the peptide residues is one of major factors that the peptides hold the charge(s) during the process of ESI [14, 23-24]. In our work, the coefficients of three basic residues (Arg, Lys, and His) are larger than the coefficients of other amino acids, and the basic residues of the peptide play an important role in the ionization of peptides. The number of basic sites of the peptides varies with the charge state of the peptide and it's shown in Table 9. In addition, we observed that there is a positive correlation between the length of peptides and the number of charges in two datasets, the correlation coefficient is 0.4953 and 0.6316 respectively. The number of the charges will increase as the length of the peptides grows. Figure 2 shows the relationship between the length of peptides and the number of peptides for each charge state in the dataset.

Table 9. The length and basic sites of the peptides

Charge state	1	2	3
Length distribution	5-17	5-49	9-49
Average length	13	16	22
Basic sites per peptide	1.05	1.3	2.02

The peptides will hold more charges as the length of peptides gets longer because there is more space to arrange the charges on the surface of the peptides.

However, we take into account the length of the peptide as one of the input variables of the model, the prediction accuracy increases slightly, 96.5% and 91.2% respectively, for both of human liver and yeast identified peptide data. This means the length of the peptide is not a sufficient but necessary factor for the charged peptide.

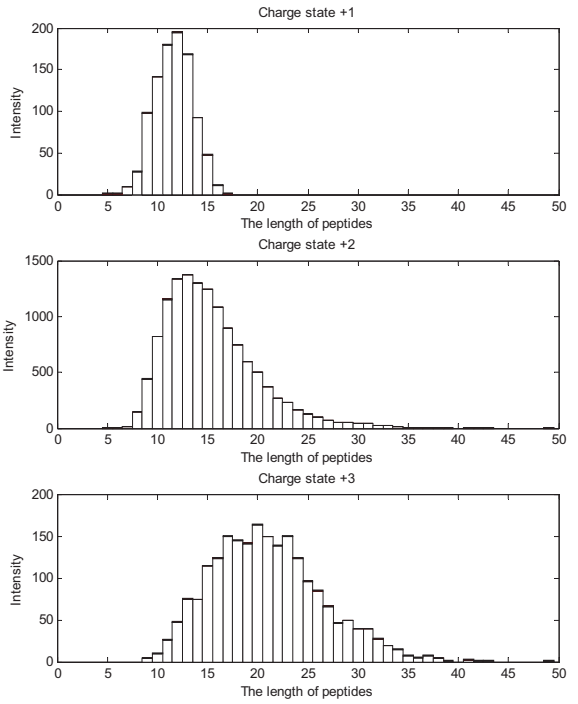

Fig.2. The relationship between the length of peptides and the number of peptides for each charge state in each subgraph

4. Conclusion

We present an approach based on the multivariate linear regression and the multiple normal distribution that predict the probability of peptide charge state in the process of ESI. The approach has been tested and verified on the different species datasets of LTQ, and it offers a good performance to predict the charge state of peptides in proteomic experiment based on MS.

Acknowledgements

This work was supported by the Chinese National Key Program of Basic Research (2006CB910803, 2006CB910706, 2010CB912700), the National High Technology Research and Development Program of China (2006AA02A312), National S&T Major Project (2008ZX10002-016, 2009ZX09301-002) and State Key Laboratory of Proteomics (grant No. SKLP-Y200811).

References

[1] R. Aebersold and M. Mann, Mass spectrometry-based proteomics. *Nature*, 2003. 422(6928): pp.198-207.

[2] X. Han, A. Aslanian, and J.R. Yates, 3rd, Mass spectrometry for proteomics. *Curr Opin Chem Biol*, 2008. 12(5): pp.483-90.

[3] W.J. Griffiths, A. P. Jonsson, S. Liu, D. K. Rai and Y. Wang, Electrospray and tandem mass spectrometry in biochemistry. *Biochem J*, 2001. 355(Pt 3): pp. 545-561.

[4] J.R. Yates, C.I. Ruse, and A. Nakorchevsky, Proteomics by mass spectrometry: approaches, advances, and applications. *Annu Rev Biomed Eng*, 2009. 11: pp.49-79.

[5] Manisali, D.D.Y. Chen, and B.B. Schneider, Electrospray ionization source geometry for mass spectrometry: past, present, and future. *Trends in Analytical Chemistry*, 2006. 25(3): pp.243-256.

[6] J. Zeleny, A Lecture Electroscope for Radioactivity and Other Ionization Experiments. *Phys. Rev. (Series I)*, 1911. 32(6): pp.581-584.

[7] J. Zeleny, Instability of Electrified Liquid Surfaces. *Phys. Rev.*, 1917. 10(1): pp.1-10.

[8] S. Chapman, Carrier Mobility Spectra of Spray Electrified Liquids. *Physical Review*, 1937. 52(3): pp.184-190.

[9] A.T. Blades, M.G. Ikonomou, and P. Kebarle, Mechanism of electrospray mass spectrometry. Electrospray as an electrolysis cell. *Anal. Chem.*, 1991. 63(19): pp.2109-2114.

[10] G.J.V. Berkel, S.A. McLuckey, and G.L. Glish, Electrochemical Origin of Radical Cations Observed in Electrospray Ionization Mass Spectra. *Anal. Chem.*, 1992. 64(14): pp.1586-1593.

[11] G.J.V. Berkel and F. Zhou, Characterization of an Electrospray Ion Source as a Controlled-Current Electrolytic Cell. *Anal. Chem.*, 1995. 67(17): pp.2916-2923.

[12] B.A. Thomson and J.V. lribarne, Field induced ion evaporation from liquid surfaces at atmospheric pressure. *J. Chern. Phys.*, 1979. 71(11): pp.4451-4463.

[13] J.V. Iribarne and B.A. Thomson, On the evaporation of small ions from charged droplets. *The Journal of Chemical Physics*, 1976. 64(6): pp.2287-2294.

[14] N. Felitsyn, M. Peschke, and P. Kebarle, Origin and number of charges observed on multiply-protonated native proteins produced by ESI. *International Journal of Mass Spectrometry,* 2002. 219: pp.39-62.

[15] M. Dole, L. L. Mack, R. L. Hines, R. C. Mobley, L. D. Ferguson and M. B. Alice, Molecular Beams of Macroions. *J. Chem. Phys*, 1968. 49(5): pp. 2240-2249.

[16] Mack, L.L., et al., Molecular Beams of Macroions. II. *J. Chem. Phys.* , 1970. 52(10): pp.4977-4986.

[17] L. Rayleigh, On the equilibrium of liquid conducting masses charged with electricity. *Philosophical Magazine*, 1882. 14(87): pp.184-186.

[18] J.B. Fenn, Ion Formation from Charged Droplets: Roles of Geometry, Energy, and Time. *J. Am. Soc. Mass Spectrom*, 1993(4): pp.524-535.

[19] J. Fernandez de la Mora, Electrospray ionization of large multiply charged species proceeds via Dole's charged residue mechanism. *Analytica chimica acta*, 2000(1): pp.93-104.

[20] S. F. Wong, C.K.M., and J. B. Fenn, Multlple Charging in Electrospray Ionization of Poly(ethy1ene glycols). *J. Phys. Chem.*, 1988. 92: pp.546-550.

[21] J.K. Enga, A.L. McCormacka, and J.R.Yates III, An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry*, 1994. 5(11): pp.976-989.

[22] B.D. Piening, P.Wang, C. S. Bangur, J. Whiteaker, H. Zhang, L. C. Feng, J. F. Keane, J. K. Eng, H. Tang, A. Prakash, M. W. McIntosh and A. Paulovich, Quality control metrics for LC-MS feature detection tools demonstrated on Saccharomyces cerevisiae proteomic profiles. *J Proteome Res*, 2006. 5(7): pp.1527-1534.

[23] P.D. Schniera, D.S. Grossa, and E.R. Williams, On the maximum charge state and proton transfer reactivity of peptide and protein ions formed by electrospray ionization. *Journal of the American Society for Mass Spectrometry*, 1995. 6(11): pp.1086-1097

[24] G.A. Pallante, and C.J. Cassady, Effects of peptide chain length on the gas-phase proton transfer properties of doubly-protonated ions from bradykinin and its N-terminal fragment peptides. *International Journal of Mass Spectrometry*, 2002. 219(1): pp.115-131