Contents lists available at ScienceDirect

International Journal of Mass Spectrometry

journal homepage: www.elsevier.com/locate/ijms

eCRAM computer algorithm for implementation of the charge ratio analysis method to deconvolute electrospray ionization mass spectra

Simin D. Maleknia[a,*], David C. Green[b]

[a] School of Biological, Earth & Environmental Sciences, University of New South Wales, Sydney, NSW 2052, Australia
[b] Information Technology Services, University of Queensland, Brisbane, QLD 4072, Australia

ARTICLE INFO

Article history:
Received 29 July 2009
Received in revised form 6 October 2009
Accepted 6 October 2009
Available online 20 November 2009

Keywords:
Charge state
ESI
Isotopic composition
Deconvolution
FT-ICR

ABSTRACT

A computer program (eCRAM) has been developed for automated processing of electrospray mass spectra based on the charge ratio analysis method. The eCRAM algorithm deconvolutes electrospray mass spectra solely from the ratio of mass-to-charge (m/z) values of multiply charged ions. The program first determines the ion charge by correlating the ratio of m/z values for any two (i.e., consecutive or non-consecutive) multiply charged ions to the unique ratios of two integers. The mass, and subsequently the identity of the charge carrying species, is further determined from m/z values and charge states of any two ions. For the interpretation of high-resolution electrospray mass spectra, eCRAM correlates isotopic peaks that share the same isotopic compositions. This process is also performed through charge ratio analysis after correcting the multiply charged ions to their lowest common ion charge. The application of eCRAM algorithm has been demonstrated with theoretical mass-to-charge ratios for proteins lysozyme and carbonic anhydrase, as well as experimental data for both low and high-resolution FT-ICR electrospray mass spectra of a range of proteins (ubiquitin, cytochrome *c*, transthyretin, lysozyme and calmodulin). This also included the simulated data for mixtures by combining experimental data for ubiquitin, cytochrome *c* and transthyretin.

1. Introduction

Biopolymers are routinely analyzed by electrospray ionization (ESI) with high-mass accuracy when combined with Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR MS) [1,2]. Computer algorithms have been developed to expedite the analysis of electrospray mass spectra and to accurately translate mass-to-charge (m/z) ratios of highly charged ions to zero-charge molecular mass values. For low-resolution mass spectra, where isotope peaks of multiply charged ions are not resolved, algorithms were developed by assuming the nature of the charge carrying species or considering only a limited set of charge carrying species [3,4]. These algorithms had other limitations of producing artefact peaks that were greatly reduced by incorporating maximum-entropy based or multiplicative correlation algorithms [5,6].

With the high-resolving power of ion cyclotron resonance (ICR) mass analyzers, multiply charged ions generated by ESI could be resolved to their isotopic compositions affording mass accuracies in the part-per-million (ppm) range, provided both the ion charge (z) and the assignment of an ion's isotopic composition was accurately determined. In the case of high-resolution electrospray mass spectra, the ion charge can be derived directly from the reciprocal of the mass-to-charge separation between adjacent isotopic peaks ($1/\Delta m/z$) for any multiply charged ion [7,8]. Although the isotope spacing method is direct, the complexity of overlapping isotope peaks for mixtures and the addition of spectral noise may result in inaccurate ion charge determination. Furthermore, for high charge states ions, distinguishing $1/z$ and $1/(z+1)$ would require mass accuracies of a few ppm that is not routinely possible. To overcome some of these limitations, pattern recognition techniques were combined with the isotope spacing method in Zscore [9] and THRASH [10] algorithms for automated charge state determination of FT-ICR mass spectra. For example, the THRASH algorithm incorporated matching the experimental abundances with theoretical isotopic distributions based on the model amino acid *averagine* ($C_{4.938}H_{7.7583}N_{1.3577}O_{1.4773}S_{0.0417}$) [8], which restricts its application to a specific group of compounds and elemental compositions.

Other algorithms developed in recent years, AID-MS [11] and PTFT [12], work on the basis of subtractive peak finding routines to locate possible isotopic clusters in the spectrum. The former algorithm shares similarities with the THRASH and relies on the isotope spacing method as well as correlating the experimental isotopic distributions with theoretical patterns (i.e., known elemental compositions). The AID-MS algorithm [11] requires the use of *averagine* for modeling isotopic distributions, while the PTFT subtracts the peak intensities in the frequency domain [12].

[*] Corresponding author. Tel.: +61 2 9385 8213.
E-mail address: s.maleknia@unsw.edu.au (S.D. Maleknia).

1387-3806/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.ijms.2009.10.005

We have developed the charge ratio analysis method (CRAM) for ESI analysis of biopolymers [13,14]. The unique feature of the CRAM is that charge states of ions are identified *purely* from the ratios of m/z values of different multiply charged ions and by correlating these values to unique ratios of two integers that subsequently identify charge of ions. This approach is therefore *independent* of the charge carrying species and as such the nature of charge carrying species is not kept *uniform* for all ions in a mass spectrum. For high-resolution data, the CRAM process could also correlate the isotopic peaks of different multiply charged ions that share the same isotopic compositions [14]. This isotopic correlation step can subsequently result in a more accurate mass determination, particularly for sample mixtures with overlapping cluster peaks and for low signal-to-noise data. In addition, the application of isotopic peak correlation within the CRAM can be valuable for mass spectrometric quantification approaches that rely on correlation of accurate isotopic peaks [15–19]. Note that the CRAM does not require a prior knowledge of the elemental composition of a molecule, and as such does not rely at all on correlating experimental isotopic patterns with the theoretical patterns (i.e., known compositions), and therefore CRAM could be applied to mass spectral data for a range of compounds (i.e., including unspecified compositions). The application of the CRAM algorithm (eCRAM) for analysis of both the low and high-resolution ESI data is described, and the algorithm is demonstrated by processing data for several proteins.

2. Experimental

Electrospray mass spectra were recorded on a 4.7 Tesla magnet (APEX, Bruker Daltonics, Billerica, MA, USA) mass spectrometer in the positive ion mode and spectra were processed with 512k or 1M data points. Theoretical isotopic distributions were produced from the isotopic distribution utility of the Xmass software (Bruker Daltonics, Billerica, MA, USA). Protein samples were obtained from Sigma Chemicals (St. Louis, MO, USA), and were used without further purification, and solutions were prepared at a concentration range of 1–5 μM in 50:50 water and methanol or acetonitrile containing 2–4% acetic acid or 0.1% TFA. The sample of transthyretin (TTR) was provided in a buffer of 400 mM sodium phosphate from the Scripps Research Institute (La Jolla, CA, USA) [20], and a C_{18} Sep-Pak (Waters Corporation, MA, USA) was used for desalting. Solutions were infused at a rate of between 3 and 5 μL/min with an electrospray needle voltage in the range of 4.2–4.5 kV. Spectra were mass calibrated with the most abundant isotopic peak of angiotensin-1 (2+ and 3+ ions) and ubiquitin (7+ to 13+) as external or internal calibration ions.

3. Results and discussions

3.1. Theoretical basis of the CRAM

The theoretical basis of the CRAM was fully described earlier [11,12], and is briefly explained here. The mass-to-charge (m/z) values for two multiply charged ions, i and j, originating from the same compound with a molecular mass (M) can be represented as $(R_z)_i$ and $(R_z)_j$, respectively. These multiply charged ions of charge (z) would correspond to the addition or abstraction of a charge carrying species (m_A), where $R_z = (M \pm zm_A)/z$. The ratio of m/z values for two ions, i and j, can then be represented by Eq. (1):

$$\frac{(R_z)_i}{(R_z)_j} = \frac{z_j(M \pm z_i m_A)}{z_i(M \pm z_j m_A)} \tag{1}$$

The CRAM approach makes an assumption that $M > z_i m_A$ or $z_j m_A$, and therefore Eq. (1) is simplified to Eq. (2):

$$\frac{(R_z)_i}{(R_z)_j} = \frac{z_j}{z_i} \tag{2}$$

Thus from the ratio of m/z values of any two multiply charged ions, the inverse ratio of their two ion charges can be calculated. The unique property of the ratio of two integers is the basis of the CRAM where by correlating the ratio of m/z values of two multiply charged ions to the unique ratio of two integers, the charge states of ions are identified without *a priori* knowledge or assumption of the nature of the charge carrying species. The purely mathematical basis of the CRAM makes it applicable to mass spectral analyses of proteins, oligonucleotides or carbohydrates with a wide range of compositions (i.e., proteins and non-proteins).

The mass (m_A) and subsequent identity of the charge carrying species is determined by the CRAM according to Eq. (3):

$$((R_z)_i z_i - (R_z)_j z_j) = \pm m_A(z_i - z_j) \tag{3}$$

For high-resolution ESI data, in order to correlate two isotopic peaks across two different multiply charged ion distributions of charge z_i and z_j where $z_i > z_j$ and $i = j + n$, the m/z value for $(R_z)_i$ in Eqs. (1) to (3) need to be corrected for the additional mass of the charge carrying species by substituting it with $(R_z)_i - n(m_A/z_i)$ [14]. This correction factor is important for high-resolution accurate mass determination, where Eq. (4) can be used to correlate isotopic peaks that share a common isotopic composition:

$$\frac{(R_z)_i - n(m_A/z_i)}{(R_z)_j} = \frac{z_j}{z_i} \tag{4}$$

The value for n represents the difference in charge of the multiply charged ions that are being correlated ($n = i - j$). For ions of charge 13 and 10, the value for n is 3.

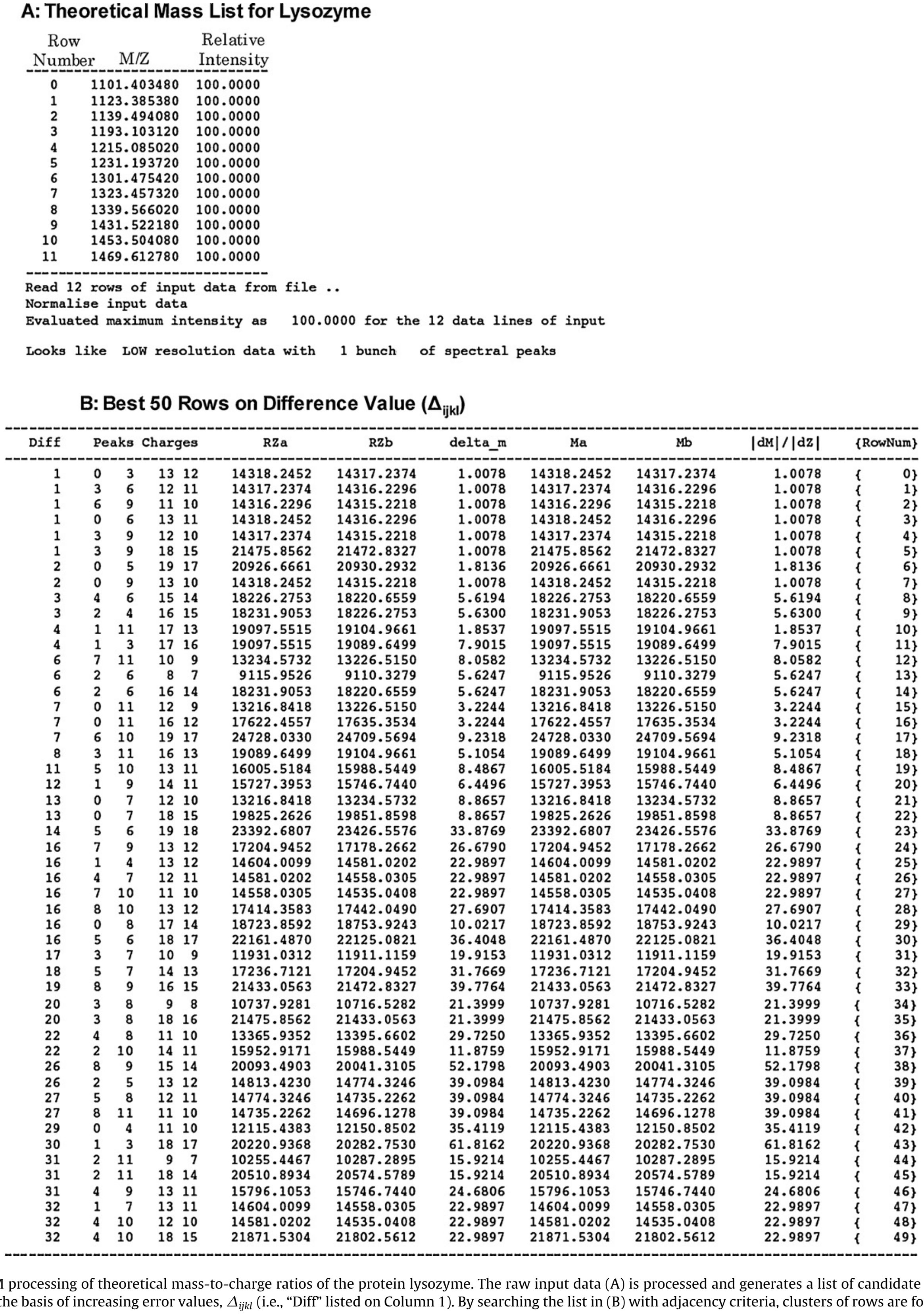
The CRAM processes electrospray mass spectra without a theoretical upper mass limit, and was previously demonstrated to calculate charge states more accurately, on the order of 100–1000 fold, in comparison to the isotope spacing method [14]. Consider for example, human carbonic anhydrase II (Protein Data Bank entry 1ca2) with the elemental composition of $C_{1324}H_{2019}N_{356}O_{383}S_2$, a monoisotopic mass of 29097.88961 Da and m/z values 2910.79679, 2646.27052 and 2425.83196 corresponding to additions of +10 to +12 protons (i.e., charges). The ratio of 2910.79679/2646.27052 is 1.09996 that is within 4×10^{-5} of 1.1 derived from dividing 11 by 10, or the ratio of 2910.79679/2425.83196 is 1.19992, which is within 8×10^{-5} of 1.2 derived from dividing 12 by 10. After applying the correction factor (Eq. (4)), and adjusting the m/z values to a common charge state of 10 (i.e., 2910.79679, 2646.27052 and 2425.83196), the ratio of 2910.79679/2646.27052 is 1.1 that is the same as dividing 11 by 10.

3.2. Computational basis of eCRAM algorithm

The CRAM algorithm (eCRAM) has been implemented using the Perl programming language for flexibility, in combination with Unix command line utilities for efficient sorting and stream editing. The program accepts, as input a comma separated file containing peak positions (m/z) and relative peak intensities. Spectral peaks are sorted automatically into increasing m/z values.

The program approaches the analysis by first populating the space of all possible solutions and then efficiently identifying physically realistic solutions within that space. The solution space is four-dimensional since each pair of spectral peaks (i and j) is combined with candidate charge values (k and l). The term “row” will be used to refer to each unique combination of (i, j, k, and l) in the solution space. The physically realistic solutions can be drawn from the best rows (i.e., those with the smallest errors between their ratio of m/z values and their ratio of charge values).

For each row, that is a pair of spectral peaks, i and j, and candidate charges, k, and l, the program computes the experimental ratio of m/z values for the two peaks, as well as with the theoretical charge ratio of the candidate charge values. The amount of discrepancy between experimental and theoretical charge ratios can be



A: Theoretical Mass List for Lysozyme

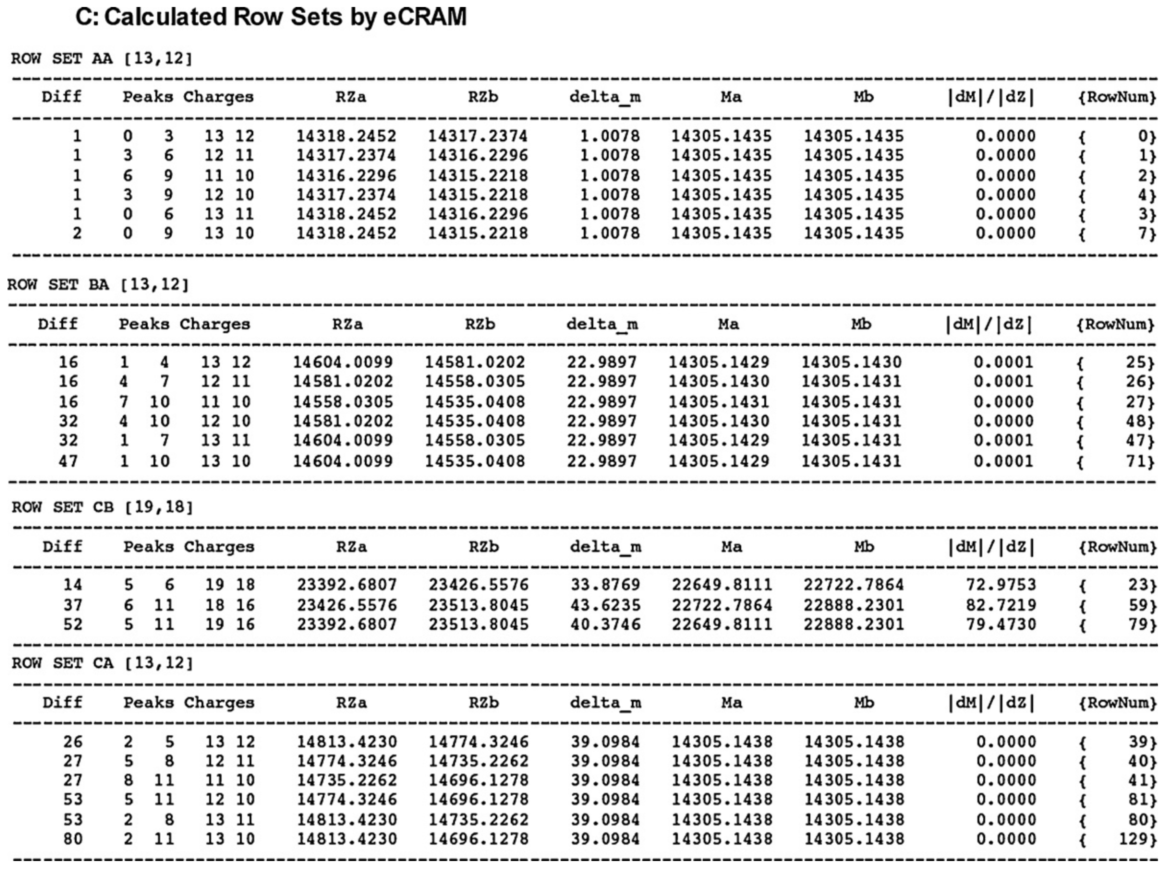
```
 Row                  Relative
Number    M/Z         Intensity
-------------------------------
   0    1101.403480  100.0000
   1    1123.385380  100.0000
   2    1139.494080  100.0000
   3    1193.103120  100.0000
   4    1215.085020  100.0000
   5    1231.193720  100.0000
   6    1301.475420  100.0000
   7    1323.457320  100.0000
   8    1339.566020  100.0000
   9    1431.522180  100.0000
  10    1453.504080  100.0000
  11    1469.612780  100.0000
-------------------------------
Read 12 rows of input data from file ..
Normalise input data
Evaluated maximum intensity as   100.0000 for the 12 data lines of input

Looks like  LOW resolution data with   1 bunch   of spectral peaks
```

B: Best 50 Rows on Difference Value (Δ_{ijkl})

```
-----------------------------------------------------------------------------------------------------------
   Diff    Peaks Charges      RZa         RZb       delta_m       Ma          Mb         |dM|/|dZ|   {RowNum}
-----------------------------------------------------------------------------------------------------------
      1    0   3   13 12   14318.2452  14317.2374    1.0078   14318.2452  14317.2374       1.0078    {     0}
      1    3   6   12 11   14317.2374  14316.2296    1.0078   14317.2374  14316.2296       1.0078    {     1}
      1    6   9   11 10   14316.2296  14315.2218    1.0078   14316.2296  14315.2218       1.0078    {     2}
      1    0   6   13 11   14318.2452  14316.2296    1.0078   14318.2452  14316.2296       1.0078    {     3}
      1    3   9   12 10   14317.2374  14315.2218    1.0078   14317.2374  14315.2218       1.0078    {     4}
      1    3   9   18 15   21475.8562  21472.8327    1.0078   21475.8562  21472.8327       1.0078    {     5}
      2    0   5   19 17   20926.6661  20930.2932    1.8136   20926.6661  20930.2932       1.8136    {     6}
      2    0   9   13 10   14318.2452  14315.2218    1.0078   14318.2452  14315.2218       1.0078    {     7}
      3    4   6   15 14   18226.2753  18220.6559    5.6194   18226.2753  18220.6559       5.6194    {     8}
      3    2   4   16 15   18231.9053  18226.2753    5.6300   18231.9053  18226.2753       5.6300    {     9}
      4    1  11   17 13   19097.5515  19104.9661    1.8537   19097.5515  19104.9661       1.8537    {    10}
      4    1   3   17 16   19097.5515  19089.6499    7.9015   19097.5515  19089.6499       7.9015    {    11}
      6    7  11   10  9   13234.5732  13226.5150    8.0582   13234.5732  13226.5150       8.0582    {    12}
      6    2   6    8  7    9115.9526   9110.3279    5.6247    9115.9526   9110.3279       5.6247    {    13}
      6    2   6   16 14   18231.9053  18220.6559    5.6247   18231.9053  18220.6559       5.6247    {    14}
      7    0  11   12  9   13216.8418  13226.5150    3.2244   13216.8418  13226.5150       3.2244    {    15}
      7    0  11   16 12   17622.4557  17635.3534    3.2244   17622.4557  17635.3534       3.2244    {    16}
      7    6  10   19 17   24728.0330  24709.5694    9.2318   24728.0330  24709.5694       9.2318    {    17}
      8    3  11   16 13   19089.6499  19104.9661    5.1054   19089.6499  19104.9661       5.1054    {    18}
     11    5  10   13 11   16005.5184  15988.5449    8.4867   16005.5184  15988.5449       8.4867    {    19}
     12    1   9   14 11   15727.3953  15746.7440    6.4496   15727.3953  15746.7440       6.4496    {    20}
     13    0   7   12 10   13216.8418  13234.5732    8.8657   13216.8418  13234.5732       8.8657    {    21}
     13    0   7   18 15   19825.2626  19851.8598    8.8657   19825.2626  19851.8598       8.8657    {    22}
     14    5   6   19 18   23392.6807  23426.5576   33.8769   23392.6807  23426.5576      33.8769    {    23}
     16    7   9   13 12   17204.9452  17178.2662   26.6790   17204.9452  17178.2662      26.6790    {    24}
     16    1   4   13 12   14604.0099  14581.0202   22.9897   14604.0099  14581.0202      22.9897    {    25}
     16    4   7   12 11   14581.0202  14558.0305   22.9897   14581.0202  14558.0305      22.9897    {    26}
     16    7  10   11 10   14558.0305  14535.0408   22.9897   14558.0305  14535.0408      22.9897    {    27}
     16    8  10   13 12   17414.3583  17442.0490   27.6907   17414.3583  17442.0490      27.6907    {    28}
     16    0   8   17 14   18723.8592  18753.9243   10.0217   18723.8592  18753.9243      10.0217    {    29}
     16    5   6   18 17   22161.4870  22125.0821   36.4048   22161.4870  22125.0821      36.4048    {    30}
     17    3   7   10  9   11931.0312  11911.1159   19.9153   11931.0312  11911.1159      19.9153    {    31}
     18    5   7   14 13   17236.7121  17204.9452   31.7669   17236.7121  17204.9452      31.7669    {    32}
     19    8   9   16 15   21433.0563  21472.8327   39.7764   21433.0563  21472.8327      39.7764    {    33}
     20    3   8    9  8   10737.9281  10716.5282   21.3999   10737.9281  10716.5282      21.3999    {    34}
     20    3   8   18 16   21475.8562  21433.0563   21.3999   21475.8562  21433.0563      21.3999    {    35}
     22    4   8   11 10   13365.9352  13395.6602   29.7250   13365.9352  13395.6602      29.7250    {    36}
     22    2  10   14 11   15952.9171  15988.5449   11.8759   15952.9171  15988.5449      11.8759    {    37}
     26    8   9   15 14   20093.4903  20041.3105   52.1798   20093.4903  20041.3105      52.1798    {    38}
     26    2   5   13 12   14813.4230  14774.3246   39.0984   14813.4230  14774.3246      39.0984    {    39}
     27    5   8   12 11   14774.3246  14735.2262   39.0984   14774.3246  14735.2262      39.0984    {    40}
     27    8  11   11 10   14735.2262  14696.1278   39.0984   14735.2262  14696.1278      39.0984    {    41}
     29    0   4   11 10   12115.4383  12150.8502   35.4119   12115.4383  12150.8502      35.4119    {    42}
     30    1   3   18 17   20220.9368  20282.7530   61.8162   20220.9368  20282.7530      61.8162    {    43}
     31    2  11    9  7   10255.4467  10287.2895   15.9214   10255.4467  10287.2895      15.9214    {    44}
     31    2  11   18 14   20510.8934  20574.5789   15.9214   20510.8934  20574.5789      15.9214    {    45}
     31    4   9   13 11   15796.1053  15746.7440   24.6806   15796.1053  15746.7440      24.6806    {    46}
     32    1   7   13 11   14604.0099  14558.0305   22.9897   14604.0099  14558.0305      22.9897    {    47}
     32    4  10   12 10   14581.0202  14535.0408   22.9897   14581.0202  14535.0408      22.9897    {    48}
     32    4  10   18 15   21871.5304  21802.5612   22.9897   21871.5304  21802.5612      22.9897    {    49}
-----------------------------------------------------------------------------------------------------------
```

Fig. 1. eCRAM processing of theoretical mass-to-charge ratios of the protein lysozyme. The raw input data (A) is processed and generates a list of candidate charge values (B) sorted on the basis of increasing error values, Δ_{ijkl} (i.e., "Diff" listed on Column 1). By searching the list in (B) with adjacency criteria, clusters of rows are found that form the "rowsets" for each charge carrying species illustrated in (C).



C: Calculated Row Sets by eCRAM

```
ROW SET AA [13,12]
-------------------------------------------------------------------------------------------------------
   Diff     Peaks Charges       RZa          RZb       delta_m        Ma           Mb       |dM|/|dZ|    {RowNum}
-------------------------------------------------------------------------------------------------------
      1     0   3   13 12   14318.2452   14317.2374    1.0078   14305.1435   14305.1435       0.0000    {      0}
      1     3   6   12 11   14317.2374   14316.2296    1.0078   14305.1435   14305.1435       0.0000    {      1}
      1     6   9   11 10   14316.2296   14315.2218    1.0078   14305.1435   14305.1435       0.0000    {      2}
      1     3   9   12 10   14317.2374   14315.2218    1.0078   14305.1435   14305.1435       0.0000    {      4}
      1     0   6   13 11   14318.2452   14316.2296    1.0078   14305.1435   14305.1435       0.0000    {      3}
      2     0   9   13 10   14318.2452   14315.2218    1.0078   14305.1435   14305.1435       0.0000    {      7}
-------------------------------------------------------------------------------------------------------
ROW SET BA [13,12]
-------------------------------------------------------------------------------------------------------
   Diff     Peaks Charges       RZa          RZb       delta_m        Ma           Mb       |dM|/|dZ|    {RowNum}
-------------------------------------------------------------------------------------------------------
     16     1   4   13 12   14604.0099   14581.0202   22.9897   14305.1429   14305.1430       0.0001    {     25}
     16     4   7   12 11   14581.0202   14558.0305   22.9897   14305.1430   14305.1431       0.0001    {     26}
     16     7  10   11 10   14558.0305   14535.0408   22.9897   14305.1431   14305.1431       0.0000    {     27}
     32     4  10   12 10   14581.0202   14535.0408   22.9897   14305.1430   14305.1431       0.0000    {     48}
     32     1   7   13 11   14604.0099   14558.0305   22.9897   14305.1429   14305.1431       0.0001    {     47}
     47     1  10   13 10   14604.0099   14535.0408   22.9897   14305.1429   14305.1431       0.0001    {     71}
-------------------------------------------------------------------------------------------------------
ROW SET CB [19,18]
-------------------------------------------------------------------------------------------------------
   Diff     Peaks Charges       RZa          RZb       delta_m        Ma           Mb       |dM|/|dZ|    {RowNum}
-------------------------------------------------------------------------------------------------------
     14     5   6   19 18   23392.6807   23426.5576   33.8769   22649.8111   22722.7864      72.9753    {     23}
     37     6  11   18 16   23426.5576   23513.8045   43.6235   22722.7864   22888.2301      82.7219    {     59}
     52     5  11   19 16   23392.6807   23513.8045   40.3746   22649.8111   22888.2301      79.4730    {     79}
-------------------------------------------------------------------------------------------------------
ROW SET CA [13,12]
-------------------------------------------------------------------------------------------------------
   Diff     Peaks Charges       RZa          RZb       delta_m        Ma           Mb       |dM|/|dZ|    {RowNum}
-------------------------------------------------------------------------------------------------------
     26     2   5   13 12   14813.4230   14774.3246   39.0984   14305.1438   14305.1438       0.0000    {     39}
     27     5   8   12 11   14774.3246   14735.2262   39.0984   14305.1438   14305.1438       0.0000    {     40}
     27     8  11   11 10   14735.2262   14696.1278   39.0984   14305.1438   14305.1438       0.0000    {     41}
     53     5  11   12 10   14774.3246   14696.1278   39.0984   14305.1438   14305.1438       0.0000    {     81}
     53     2   8   13 11   14813.4230   14735.2262   39.0984   14305.1438   14305.1438       0.0000    {     80}
     80     2  11   13 10   14813.4230   14696.1278   39.0984   14305.1438   14305.1438       0.0000    {    129}
-------------------------------------------------------------------------------------------------------
```

Fig. 1. (*Continued*).

represented by the variable Δ_{ijkl}:

$$\Delta_{ijkl} = \left| \left(\left[\frac{(m/z)_i}{(m/z)_j} \right] \cdot \left[\frac{k}{l} \right] - 1 \right) \right| \qquad (5)$$

For an ideal spectrum, the value for Δ_{ijkl} will be zero for the correct charge values k and l and their corresponding peaks i and j. However, this criterion cannot be applied alone. Uncertainties in the experimental measurements and/or spectral peak selection could result in physically realistic solutions having non-zero Δ_{ijkl} values. Conversely, non-physical solutions could have low values of Δ_{ijkl}. This is apparent when one recognizes that the ratios of integer charges are not all unique. For instance, the ratio of charges 18 and 15 is the same as the ratio of charges 12 and 10.

A valid combination (i, j, k, l) will always occur amongst a set of other related combinations (rows). In order to eliminate non-physical solutions, adjacency criteria are applied. The nature of this adjacency depends on the resolution of the original spectral data. In a low-resolution spectrum, the m/z peak positions occur in a set corresponding to a sequence of charge values for the *same* mass value. On the other hand, in a high-resolution spectrum, the peaks in a spectral cluster correspond to different isotopes of the *same* pair of charge values.

Consider two "rows", A and B, with their attributes denoted thus:

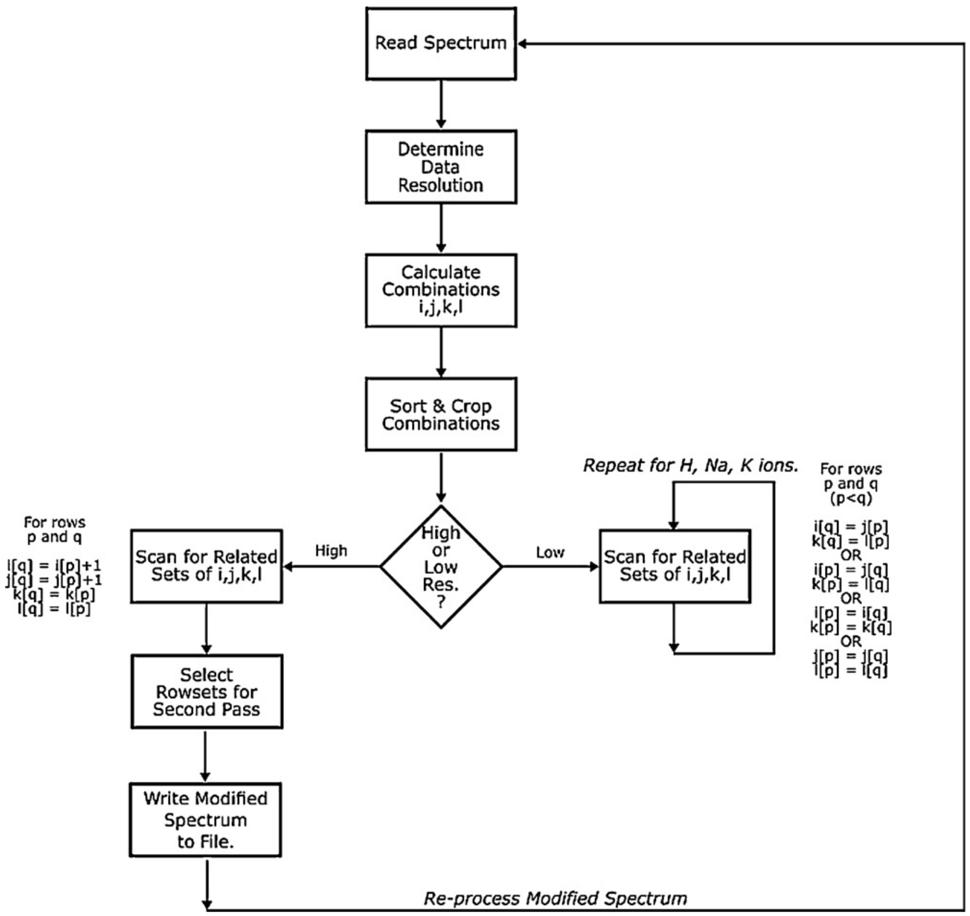
$$\begin{array}{ll} i_A, j_A, k_A, l_A, & [\Delta_{ijkl}]_{\mathbf{A}} \\ i_B, j_B, k_B, l_B, & [\Delta_{ijkl}]_{\mathbf{B}} \end{array}$$

These two "rows" (that is, rows of the printout such as in Fig. 1B) are part of a "row set" if they satisfy certain adjacency criteria (see below). A "row set" is a coherent clump of peak identifiers and associated charge values. For low resolution, a row set will be comprised of multiple charge combinations. For high-resolution data, a row set will be comprised of isotope peaks for the same charge combination. A valid row set represents a physically realistic solution that is internally consistent. In the analysis of the low-resolution spectra shown in Fig. 1C, the row set labelled AA is comprised of spectral peaks labelled 0 (1101.403480), 3 (1193.103120), 6 (1301.475420) and 9 (1431.522180) and their assigned charge values of 13, 12, 11 and 10. The consistency of the mass difference values (delta_m) and the Diff column adds further weight to the argument that this "row set" represents a physically realistic solution.

Scanning for combinations of peaks and charge values that form part of a "row set" involves scanning the combinations (i, j, k, l) that satisfy the adjacency conditions. In a low-resolution spectra, the adjacency condition requires one of the following to be true for the pair: ($i_B = j_A$ and $k_B = l_A$), ($i_A = j_B$ and $k_A = l_B$), ($i_A = i_B$ and $k_A = k_B$), or ($j_A = j_B$ and $l_A = l_B$). For the high-resolution mode, the adjacency condition requires $k_B = k_A$ and $l_B = l_A$.

There is no line drawn between "row set" boundaries. The rows cluster together into row sets based purely on their charge ratios, the adjacency criteria and their charge carrying species. The maximum difference values that are considered can be adjusted when the program is run. For both the low and high-resolution spectra, the best "row set" contains all the related charge values. Realistic solutions are observed to form into clusters corresponding to related combinations of charges. For example, a combination of charges 13 with 11 would be expected along with the combinations 13 with 12, and, 12 with 11. The row sets corresponding to unrealistic solutions may still be present in the final output and are discounted automatically or by the user input.

To analyze the isotopic effects in high-resolution spectra, the m/z values of the row sets of interest are corrected for the mass of the charge carrier and then the corrected spectrum is



Scheme 1. Flowchart for eCRAM processing of low and high-resolution mass spectral data.

completely re-processed. During the second pass, the program prompts the user to select a subset of the row sets to display as the final output. The molecular weights of the proteins (allowing for isotopic variations) are computed and printed out so as to highlight the correspondence of spectral peaks and associated protein mass.

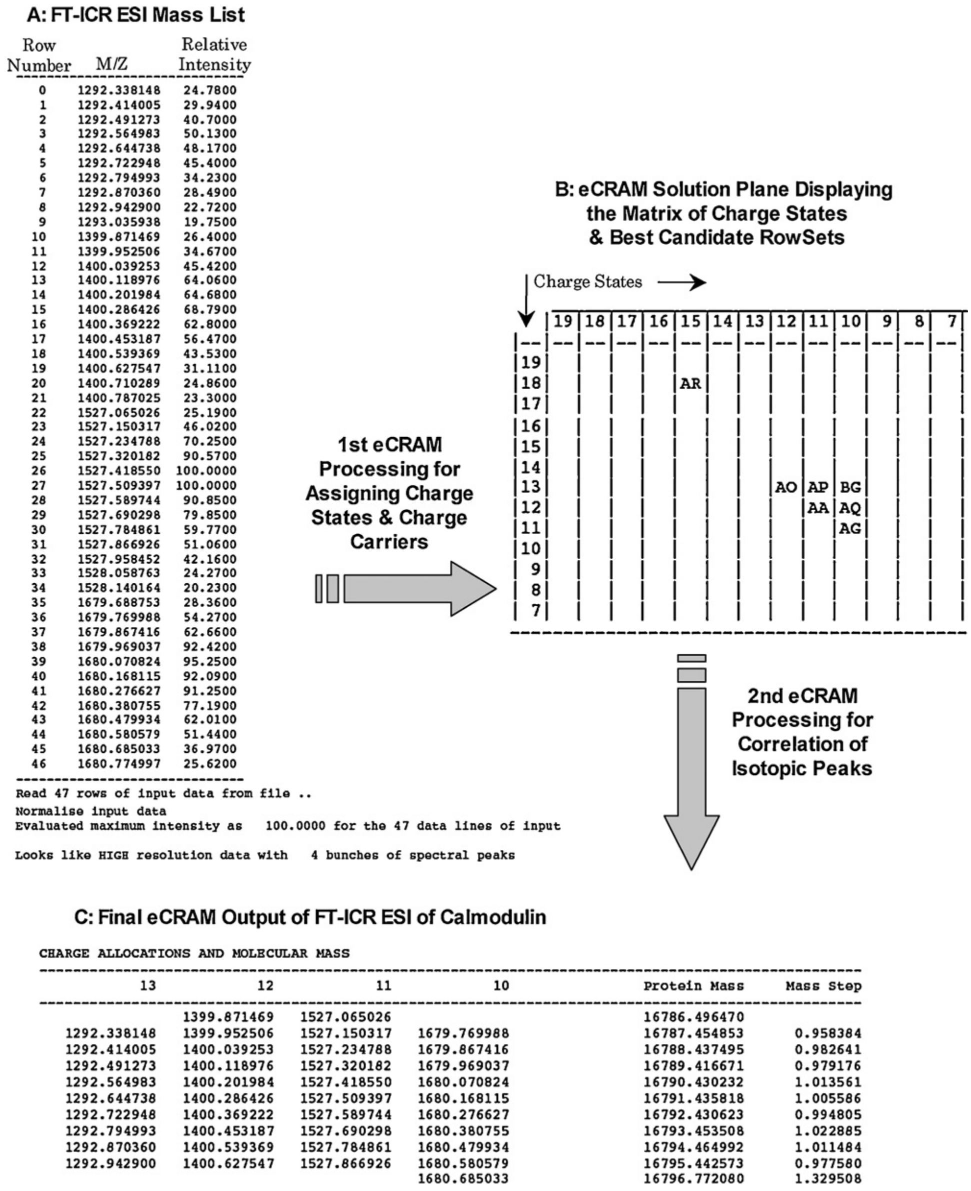
The major processing steps performed by the eCRAM are shown in Scheme 1 and described below:

- *Step 1*. Load spectrum, normalize intensities, and determine data resolution (low versus high resolution).
- *Step 2*. Compute error values, Δ_{ijkl}, for all combinations of spectral peaks (i and j) and candidate charge values (k and l). These combinations of peaks and charges are referred to as rows (a user-defined lower and upper range can be used for the candidate charge values to minimize computing time). The error values, Δ_{ijkl}, are referred to as "Diff" for difference values in the program output and are scaled to a suitable range for display.
- *Step 3*. Sort the combinations (i, j, k, l) on increasing Δ_{ijkl}. A user-defined error range is typically selected (i.e., 0.03 for low resolution and 0.0001 for high-resolution data).
- *Step 4*. For each combination of i, j, k, and l, commencing at the smallest Δ_{ijkl} values, search for other combinations in the sorted list that form part of the same "row set". The criteria for adjacency depend on the resolution of the spectral data.
- *Step 5*. For low-resolution spectra, display the best "row sets" for each candidate charge species. For high-resolution spectra, on the first pass, display the best "row sets" and allow for input of the row sets for subsequent re-processing for isotopic effects (Eq. (4)). For high-resolution spectra, on the second pass, display the best "row sets" and allow for input of the row sets for final display.
- *Step 6*. For high-resolution spectra, on the second pass, calculate and display the peak registrations and corresponding protein mass values.

3.3. *eCRAM processing of low-resolution electrospray mass spectral data*

The example provided in Fig. 1 illustrates the application of eCRAM utilizing theoretical mass-to-charge ratios of the protein lysozyme [11]. This data set is for low-resolution ESI mass spectra and contains protonated as well as adducts of sodium and potassium ions. The list shown in Fig. 1A has 12 m/z values in ascending order with their corresponding row numbers. For example, row zero has a corresponding m/z value of 1101.4034. The intensity of all ions in this example is set to an arbitrary value of 100 to represent the theoretically calculated m/z values. The program recognizes that the set is for low-resolution data with one bunch of spectral peaks (i.e., no isotopic peaks), and computes m/z ratio values along with candidate charge ratio values. In this example, the lower and upper values of the matrix were 7 and 19 (i.e., a user-defined range minimizes the computing time).

The output shown in Fig. 1B shows the best 50 rows sorted on the basis of increasing error values, Δ_{ijkl} (i.e., "Diff" listed on Column 1). The mass of charge carrying species listed as "delta_m" on Column 8 of Fig. 1B is calculated from Eq. (3). The correction factor (Eq. (4)) for isotopic correlation of high-resolution data is represented as $|dM|/|dZ|$ under Column 11, which is the same value



A: FT-ICR ESI Mass List

```
   Row                  Relative
 Number    M/Z          Intensity
 ------------------------------
    0    1292.338148    24.7800
    1    1292.414005    29.9400
    2    1292.491273    40.7000
    3    1292.564983    50.1300
    4    1292.644738    48.1700
    5    1292.722948    45.4000
    6    1292.794993    34.2300
    7    1292.870360    28.4900
    8    1292.942900    22.7200
    9    1293.035938    19.7500
   10    1399.871469    26.4000
   11    1399.952506    34.6700
   12    1400.039253    45.4200
   13    1400.118976    64.0600
   14    1400.201984    64.6800
   15    1400.286426    68.7900
   16    1400.369222    62.8000
   17    1400.453187    56.4700
   18    1400.539369    43.5300
   19    1400.627547    31.1100
   20    1400.710289    24.8600
   21    1400.787025    23.3000
   22    1527.065026    25.1900
   23    1527.150317    46.0200
   24    1527.234788    70.2500
   25    1527.320182    90.5700
   26    1527.418550   100.0000
   27    1527.509397   100.0000
   28    1527.589744    90.8500
   29    1527.690298    79.8500
   30    1527.784861    59.7700
   31    1527.866926    51.0600
   32    1527.958452    42.1600
   33    1528.058763    24.2700
   34    1528.140164    20.2300
   35    1679.688753    28.3600
   36    1679.769988    54.2700
   37    1679.867416    62.6600
   38    1679.969037    92.4200
   39    1680.070824    95.2500
   40    1680.168115    92.0900
   41    1680.276627    91.2500
   42    1680.380755    77.1900
   43    1680.479934    62.0100
   44    1680.580579    51.4400
   45    1680.685033    36.9700
   46    1680.774997    25.6200
------------------------------
Read 47 rows of input data from file ..
Normalise input data
Evaluated maximum intensity as   100.0000 for the 47 data lines of input

Looks like HIGH resolution data with   4 bunches of spectral peaks
```

B: eCRAM Solution Plane Displaying the Matrix of Charge States & Best Candidate RowSets

```
| Charge States  ——>
↓  |19|18|17|16|15|14|13|12|11|10| 9| 8| 7|
|--|--|--|--|--|--|--|--|--|--|--|--|--|--|
|19|  |  |  |  |  |  |  |  |  |  |  |  |  |
|18|  |  |  |  |AR|  |  |  |  |  |  |  |  |
|17|  |  |  |  |  |  |  |  |  |  |  |  |  |
|16|  |  |  |  |  |  |  |  |  |  |  |  |  |
|15|  |  |  |  |  |  |  |  |  |  |  |  |  |
|14|  |  |  |  |  |  |  |  |  |  |  |  |  |
|13|  |  |  |  |  |  |  |AO|AP|BG|  |  |  |
|12|  |  |  |  |  |  |  |  |AA|AQ|  |  |  |
|11|  |  |  |  |  |  |  |  |  |AG|  |  |  |
|10|  |  |  |  |  |  |  |  |  |  |  |  |  |
| 9|  |  |  |  |  |  |  |  |  |  |  |  |  |
| 8|  |  |  |  |  |  |  |  |  |  |  |  |  |
| 7|  |  |  |  |  |  |  |  |  |  |  |  |  |
-------------------------------------------
```

C: Final eCRAM Output of FT-ICR ESI of Calmodulin

```
CHARGE ALLOCATIONS AND MOLECULAR MASS
-------------------------------------------------------------------------------------------
          13           12           11           10              Protein Mass    Mass Step
-------------------------------------------------------------------------------------------
                  1399.871469  1527.065026                       16786.496470
   1292.338148   1399.952506  1527.150317  1679.769988           16787.454853     0.958384
   1292.414005   1400.039253  1527.234788  1679.867416           16788.437495     0.982641
   1292.491273   1400.118976  1527.320182  1679.969037           16789.416671     0.979176
   1292.564983   1400.201984  1527.418550  1680.070824           16790.430232     1.013561
   1292.644738   1400.286426  1527.509397  1680.168115           16791.435818     1.005586
   1292.722948   1400.369222  1527.589744  1680.276627           16792.430623     0.994805
   1292.794993   1400.453187  1527.690298  1680.380755           16793.453508     1.022885
   1292.870360   1400.539369  1527.784861  1680.479934           16794.464992     1.011484
   1292.942900   1400.627547  1527.866926  1680.580579           16795.442573     0.977580
                                           1680.685033           16796.772080     1.329508
```

Fig. 2. eCRAM processing of high-resolution FT-ICR electrospray mass spectrum of calmodulin. The rowsets within the triangular region bounded by (AO, BG and AG) are selected for second stage processing.

as “delta_m” in this example of low-resolution data processing (Fig. 1B).

The program assembles the row sets in the following way. By starting with the best row (smallest Diff value), the eCRAM algorithm searches down the list for other rows that satisfy the adjacency criteria required to join the current “row set”. The program then returns to the top of the list and selects the next best, non-assigned row and repeats the process. A number of filtering and merging steps are performed to obtain the rowset depicted in Fig. 1C. The assembly of the row set AA depicted in Fig. 1C, can be explained by reading down the “RowNum” in Fig. 1B (last column):

- RowNum 0 – the row (0, 3, 13, 12) is the nucleation point for this “row set” as it occurs first in the row list (best difference value).
- RowNum 1 – the row (3, 6, 12, 11) satisfies the adjacency criteria with RowNum 0 (i.e., 3 = 3 and 12 = 12), and also RowNum 2 and 3.
- RowNum 2 – the row (6, 9, 11, 10) satisfies the adjacency criteria with RowNum 1 (i.e., 6 = 6 and 11 = 11), and also RowNum 3 and 4.
- RowNum 3 – the row (0, 6, 13, 11) satisfies the adjacency criteria with RowNum 2 (i.e., 6 = 6 and 11 = 11), and also RowNum 1 and 7.
- RowNum 4 – the row (3, 9, 12, 10) satisfies the adjacency criteria with RowNum 0 (i.e., 3 = 3 and 12 = 12), and also RowNum 2 and 7.
- RowNum 5 – the row (3, 9, 18, 15) is not adjacent to any row set members and appears to be an integer multiple of the charge ratio 12:10.
- RowNum 6 – the row (0, 5, 19, 17) is not adjacent to any row set members and the delta_m is not an expected value.
- RowNum 7 – the row (0, 9, 13, 10) satisfies the adjacency criteria with RowNum 3 (i.e., 0 = 0 and 13 = 13) and RowNum 4 (i.e., 9 = 9 and 10 = 10).

eCRAM has matched all protonated species as part of the row set AA. When looking for sodium species (22.9897), the nucleation point is the row (1, 4, 13, 12), and the adjacency criteria result in the rows forming the row set BA (Fig. 1C) from peaks 1, 4, 7, 10 with charges of 13, 12, 11 and 10. Similarly for potassium (39.0984), the row set CA (Fig. 1C) is comprised of peaks 2, 5, 8 and 11 with corresponding charges of 13, 12, 11 and 10. Note that although the row set CB satisfies the adjacency criteria, inspection of the delta_m values discounts it as a physically unrealistic solution.

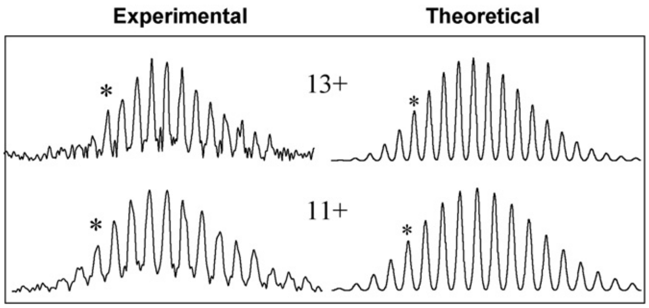


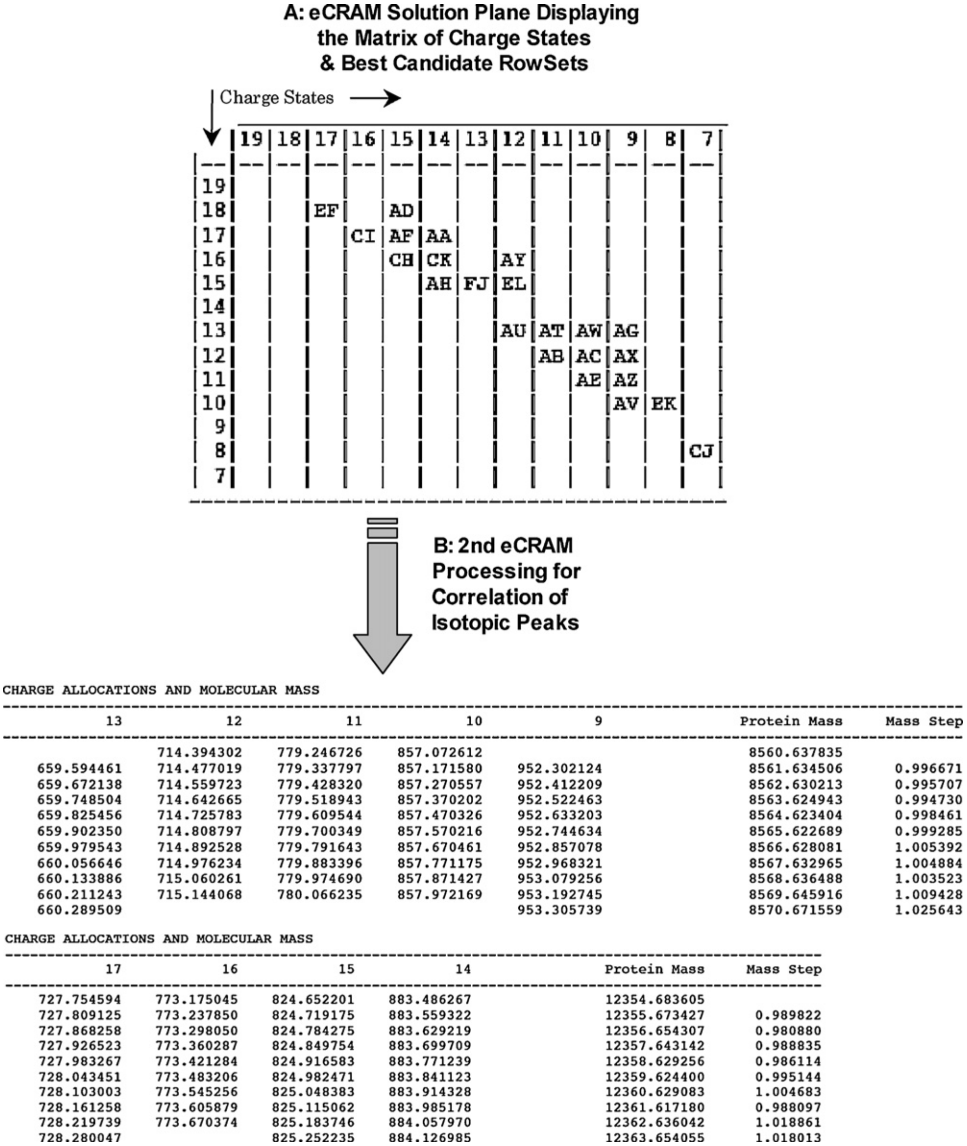
Fig. 3. Correlation of the experimental and theoretical isotopic peaks for +13 and +11 ions of calmodulin by the eCRAM.

Note that the values of charge carrying species (i.e., proton, sodium and potassium) are consistent throughout various row sets. For instance, the charge carrying species calculated for all six sets of charges of row set AA is 1.0078 (proton). Similarly, for the row set BA, the charge carrying species is 22.9897 (sodium) and for the row set CA, a value of 39.0984 (potassium) is calculated. This process of identifying the charge carrying species, purely from the ratios of charge states, is a unique feature of the CRAM. Values for Δ_{ijkl} (Eq. (5)) listed under "Diff" are larger for row sets BA and CA when compared to the row set AA. This is expected and is due to the increase in the mass of charge carrying species and the assumption of Eq. (3) ($M > z_i m_A$ or $z_j m_A$).

As described above in the description of the eCRAM algorithm, non-realistic solutions could also be found (for example, the ratio of charges 18 and 15 is the same as the ratio of charges 12 and 10). For the lysozyme data, Fig. 1C shows the row set CB with acceptable error values. This set matches data for charge states 19 and 18, 18 and 16, and 19 and 16, while missing other key charge states of 18 and 17, or 19 and 17. Additionally, the value of charge carrying species is not consistent for each set of charges. These two features identify that the row set CB is not a valid realistic solution and could be removed.

3.4. eCRAM processing of high-resolution electrospray mass spectral data

Processing of high-resolution mass spectral data occurs in two steps [12]. In the first instance, the charge state values of ions


Charge States

A: eCRAM Solution Plane Displaying the Matrix of Charge States & Best Candidate RowSets

```
     |19|18|17|16|15|14|13|12|11|10| 9| 8| 7|
|--|--|--|--|--|--|--|--|--|--|--|--|--|--|
|19|  |  |  |  |  |  |  |  |  |  |  |  |  |
|18|  |  |EF|  |AD|  |  |  |  |  |  |  |  |
|17|  |  |  |CI|AF|AA|  |  |  |  |  |  |  |
|16|  |  |  |  |CH|CK|  |AY|  |  |  |  |  |
|15|  |  |  |  |  |AH|FJ|EL|  |  |  |  |  |
|14|  |  |  |  |  |  |  |  |  |  |  |  |  |
|13|  |  |  |  |  |  |  |AU|AT|AW|AG|  |  |
|12|  |  |  |  |  |  |  |  |AB|AC|AX|  |  |
|11|  |  |  |  |  |  |  |  |  |AE|AZ|  |  |
|10|  |  |  |  |  |  |  |  |  |  |AV|EK|  |
| 9|  |  |  |  |  |  |  |  |  |  |  |  |  |
| 8|  |  |  |  |  |  |  |  |  |  |  |  |CJ|
| 7|  |  |  |  |  |  |  |  |  |  |  |  |  |
--------------------------------------------
```

B: 2nd eCRAM Processing for Correlation of Isotopic Peaks

```
CHARGE ALLOCATIONS AND MOLECULAR MASS
--------------------------------------------------------------------------------------------------------
         13            12            11            10             9               Protein Mass   Mass Step
--------------------------------------------------------------------------------------------------------
                 714.394302    779.246726    857.072612                            8560.637835
 659.594461      714.477019    779.337797    857.171580    952.302124              8561.634506    0.996671
 659.672138      714.559723    779.428320    857.270557    952.412209              8562.630213    0.995707
 659.748504      714.642665    779.518943    857.370202    952.522463              8563.624943    0.994730
 659.825456      714.725783    779.609544    857.470326    952.633203              8564.623404    0.998461
 659.902350      714.808797    779.700349    857.570216    952.744634              8565.622689    0.999285
 659.979543      714.892528    779.791643    857.670461    952.857078              8566.628081    1.005392
 660.056646      714.976234    779.883396    857.771175    952.968321              8567.632965    1.004884
 660.133886      715.060261    779.974690    857.871427    953.079256              8568.636488    1.003523
 660.211243      715.144068    780.066235    857.972169    953.192745              8569.645916    1.009428
 660.289509                                                953.305739              8570.671559    1.025643
```

```
CHARGE ALLOCATIONS AND MOLECULAR MASS
--------------------------------------------------------------------------------------
         17            16            15            14               Protein Mass   Mass Step
--------------------------------------------------------------------------------------
 727.754594      773.175045    824.652201    883.486267             12354.683605
 727.809125      773.237850    824.719175    883.559322             12355.673427    0.989822
 727.868258      773.298050    824.784275    883.629219             12356.654307    0.980880
 727.926523      773.360287    824.849754    883.699709             12357.643142    0.988835
 727.983267      773.421284    824.916583    883.771239             12358.629256    0.986114
 728.043451      773.483206    824.982471    883.841123             12359.624400    0.995144
 728.103003      773.545256    825.048383    883.914328             12360.629083    1.004683
 728.161258      773.605879    825.115062    883.985178             12361.617180    0.988097
 728.219739      773.670374    825.183746    884.057970             12362.636042    1.018861
 728.280047                    825.252235    884.126985             12363.654055    1.018013
```

Fig. 4. eCRAM processing of the combined high-resolution FT-ICR electrospray mass spectra of ubiquitin and cytochrome *c* as an example of mixtures. Two distinct rowsets (i.e., mixture of two proteins) within the triangular regions bounded by (AU, AG and AV) and (CI, AA, AH) are selected for second stage processing.

are calculated along with the mass (m_A) and subsequent identification of the charge carrying species (Eq. (3)). In the following steps, the isotopic peaks of the multiply charged ions that share a common isotopic composition are correlated (Eq. (4)). These are illustrated with high-resolution mass spectra of the protein calmodulin (Figs. 2 and 3).

The list in Fig. 2A contains 47 rows of m/z and intensity values. The program first recognizes that the list represents a set of high-resolution data with four bunches of spectral lines. The row sets are then identified from the ratios of m/z values and by comparing those to ratios of integers. The matrix in Fig. 2B shows seven possible row sets, all within an acceptable error value, $\Delta_{ijkl} = 0.0005$ (Eq. (5)). The row set AR (corresponding to charge state 18, 15) does not correlate with other Row Sets, and is eliminated at this stage. The other six row sets (i.e., AO, AP, BG, AA, AQ, AG) represent correlation of ion charges 13, 12, 11 and 10. The charge carrying species is identified as a “proton” (as described above for low-resolution data).

For the correlation of the isotopic peaks, m/z values are adjusted and eCRAM generates a new list of m/z values based on the correction applied from Eq. (4) (i.e., the mass and the numbers of charge carrying species associated with each ion). eCRAM applies the algorithm to the new m/z list and generates the final output (Fig. 2C) that correlates isotopic peaks for all ions. Note in this example, the first isotope peak of charge 13 (m/z 1292.338148), the second isotope peaks of charge 12 and 11 (i.e., m/z 1399.952506 and 1527.150317) and the first isotope peak of charge 10 (m/z 1679.769988) are correlated. The protein mass (i.e., neutral mass) is calculated from the average of m/z values multiplied by their corresponding charges and subtracting the mass of charge carrying species (i.e., proton in this example). The protein mass based on the lightest isotopic peak is 16786.496470 as listed on the first row of Column 5 (Fig. 2C).

The eCRAM correlation of isotopic peaks for high-resolution ESI data is also helpful for matching experimental and theoretical isotopic patterns [14]. In the case of compounds where the elemental composition is not known, an elemental composition is assumed ($C_aH_bO_cN_dS_e$). Correlation of the experimental and theoretical isotopic peaks for +13 and +11 ions of calmodulin by the eCRAM is shown in Fig. 3. Similar results for processing of both low and high-resolution ESI data of five proteins, ubiquitin, lysozyme, cytochrome *c*, transthyretin (TTR) and calmodulin were obtained by the eCRAM. In all these cases, the eCRAM algorithm correctly identified the ion charge and correlated isotopic peaks.

3.5. eCRAM processing of high-resolution electrospray mass spectral data for mixtures

To evaluate the eCRAM algorithm for processing of complex sample mixtures, mass spectral data for two proteins, ubiquitin and cytochrome *c* were combined. The simulated data for mixtures contained high-resolution m/z values for charge states 9 to 13 of ubiquitin, and charge states 14 to 17 for cytochrome *c*. The eCRAM algorithm successfully processed the mixture data by correctly calculating the charge states of the two proteins, and subsequently correlated their charge states with m/z values of their isotopic peaks. The eCRAM output with charge allocations and molecular mass of ubiquitin and cytochrome *c*, from processing the proteins as a mixture, are presented in Fig. 4. Note in this case, the matrix (Fig. 4A) displays two distinct triangular regions bounded by (AU, AG and AV) and (CI, AA, AH) that represent the charge states and their associated row sets for the two proteins.

The eCRAM algorithm was evaluated further by including an impurity in the form of an overlapping peak (data not shown). Mass spectral data for +18 charge state of transthyretin (TTR) were added to the mass list of ubiquitin and cytochrome *c* mixture (described above). In this instance, 10 extra peaks associated with TTR ranging from m/z 772.599554 to 773.099633, partially overlapped with +16 peaks of cytochrome *c* ranging from m/z 773.114083 to 773.795056. The eCRAM algorithm once again identified and correlated correct charge states with m/z values associated with ubiquitin and cytochrome *c*, and the presence of the impurity did not disturb the processing of the data.

4. Conclusions

The eCRAM computer program and the processing of both low and high-resolution data were described. In addition, methodologies for the selection of real solutions and elimination of non-realistic solutions were provided. Mass spectral data for a range of proteins including mixtures, were successfully processed by the eCRAM algorithm, and examples were provided for lysozyme, calmodulin, and a mixture containing ubiquitin and cytochrome *c*. The eCRAM demonstrated that both the ion charge and the mass of charge carrying species are calculated without *a priori* knowledge and assumption of the nature of charge carrying species, and more importantly, the charge carrying species is not assumed to be *uniform* across an electrospray mass spectrum. The isotope correlation feature of the eCRAM is especially useful for quantitative applications where cross-correlation of mass spectral data with theoretical isotopic distribution is required, particularly for low signal-to-noise data and where elemental compositions are not known. For accurate processing of the high-resolution mass spectral data by eCRAM some user-defined input for selection of a set of solutions is recommended. The eCRAM algorithm is currently being optimized for availability on the World Wide Web. Options for a standalone version are also being explored with provisions to provide the program code to non-commercial users.

Acknowledgements

This work was commenced while the authors were based in the Faculty of Science and Research Computing Services (respectively) at Griffith University, QLD, Australia.

References

[1] J.B. Fenn, M. Mann, C.K. Meng, S.F. Wong, G.M. Whitehouse, Science 246 (1989) 64.
[2] F.W. McLafferty, Acc. Chem. Res. 27 (1994) 379.
[3] T.R. Covey, R.F. Bonner, B.I. Shushan, J. Henion, R.K. Boyd, Rapid Commun. Mass Spectrom. 11 (2) (1988) 249.
[4] M. Mann, C.K. Meng, J.B. Fenn, Anal. Chem. 61 (1989) 1702.
[5] B.B. Reinhold, V.N. Reinhold, J. Am. Soc. Mass Spectrom. 3 (1992) 207.
[6] J.J. Hagen, C.A. Monnig, Anal. Chem. 66 (1994) 1877.
[7] M.W. Senko, S.C. Beu, F.W. McLafferty, J. Am. Soc. Mass Spectrom. 6 (1995) 52.
[8] M.W. Senko, S.C. Beu, F.W. McLafferty, J. Am. Soc. Mass Spectrom. 6 (1995) 229.
[9] Z. Zhang, A.G. Marshall, J. Am. Soc. Mass Spectrom. 9 (1998) 225.
[10] D.M. Horn, R.A. Zubarev, F.W. McLafferty, J. Am. Soc. Mass Spectrom. 11 (4) (2000) 320.
[11] L. Chen, S.K. Sze, H. Yang, Anal. Chem. 78 (14) (2006) 5006.
[12] L. Chen, H. Yang, J. Am. Soc. Mass Spectrom. 19 (2008) 46.
[13] S.D. Maleknia, K.M. Downard, Anal. Chem. 77 (2005) 111.
[14] S.D. Maleknia, K.M. Downard, Int. J. Mass Spectrom. 246 (2005) 1.
[15] M.T. Olson, A.L. Yergey, J. Am. Soc. Mass Spectrom. 20 (2009) 295.
[16] J. Meija, Anal. Bioanal. Chem. 385 (2006) 486.
[17] F. Former, L.J. Foster, S. Toppo, Curr. Bioinform. 2 (2007) 63.
[18] M. Wehofsky, R. Hoffmann, J. Mass Spectrom. 37 (2002) 223.
[19] J. Fernandez-de-Cossio, L.J. Gonzales, Y. Satomi, L. Betancourt, Y. Ramos, V. Huerta, V. Besada, G. Padron, N. Minamino, T. Takao, Rapid Commun. Mass Spectrom. 18 (2004) 2465.
[20] S.D. Maleknia, N. Reixach, J.N. Buxbaum, FEBS J. 273 (2006) 5400.