Automated Interpretation of Mass Spectra of Complex Mixtures by Matching of Isotope Peak Distributions

ARTICLE *in* RAPID COMMUNICATIONS IN MASS SPECTROMETRY · FEBRUARY 2004

Impact Factor: 2.25 · DOI: 10.1002/rcm.1647 · Source: PubMed

10 AUTHORS, INCLUDING:

Lazaro Betancourt
Center for Genetic Engineering and Biotech...
55 PUBLICATIONS **774** CITATIONS
SEE PROFILE

Vivian Huerta
Center for Genetic Engineering and Biotech...
38 PUBLICATIONS **485** CITATIONS
SEE PROFILE

Vladimir Besada
Center for Genetic Engineering and Biotech...
102 PUBLICATIONS **1,048** CITATIONS
SEE PROFILE

Gabriel Padron
University of Barcelona
103 PUBLICATIONS **1,349** CITATIONS
SEE PROFILE

RAPID COMMUNICATIONS IN MASS SPECTROMETRY
Rapid Commun. Mass Spectrom. 2004; **18**: 2465–2472
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/rcm.1647

Automated interpretation of mass spectra of complex mixtures by matching of isotope peak distributions

Jorge Fernández-de-Cossio[1]*, Luis Javier Gonzalez[1], Yoshinori Satomi[2], Lazaro Betancourt[1], Yassel Ramos[1], Vivian Huerta[1], Vladimir Besada[1], Gabriel Padron[1], Naoto Minamino[3] and Toshifumi Takao[2]**

[1]Center for Genetic Engineering and Biotechnology, P.O. Box 6162, Havana, Cuba
[2]Institute for Protein Research, Osaka University, Yamadaoka 3-2, Suita, Osaka 565-0871, Japan
[3]Department of Pharmacology, National Cardiovascular Center Research Institute, Fujishirodai, Suita, Osaka 565-8565, Japan

Received 13 August 2004; Revised 19 August 2004; Accepted 19 August 2004

Mass spectrometry is now firmly established as a powerful technique for the identification and characterization of proteins when used in conjunction with sequence databases. Various approaches involving stable-isotope labeling have been developed for quantitative comparisons between paired samples in proteomic expression analysis by mass spectrometry. However, interpretation of such mass spectra is far from being fully automated, mainly due to the difficulty of analyzing complex patterns resulting from the overlap of multiple peaks arising from the assortment of natural isotopes. In order to facilitate the interpretation of a complex mass spectrum of such a mixture, such as an MS spectrum of a stable-isotope-enriched ion species, we report on the development of a software application, 'Matching' (web accessible), that enables the automatic matching of theoretical isotope envelopes to multiple ion peaks in a raw spectrum. It is particularly useful for resolving the relative abundances of narrow-split paired peaks caused by enrichment with a stable isotope, such as ^{18}O, ^{13}C, ^{2}H, or ^{15}N. Copyright © 2004 John Wiley & Sons, Ltd.

Along with the recent and accelerated advances in genomics, proteomics, information technologies and instrumentation, research efforts have shifted from the analysis of individual proteins or genes to the overall analysis of complex mixtures of proteins from cell homogenates or tissue lysates. Proteomics research has been heavily influenced by high-throughput technologies based on mass spectrometry (MS), in which protein identification[1–4] by correlating mass spectrometric data with sequence databases, and the *de novo* sequencing of peptides,[5–7] have become routine and indispensable processes. However, the automatic processing of huge amounts of experimental data generated by MS may be prone to significant errors, and mandates validation of the results. A major ambiguity encountered in data processing is caused by the multiplicity of charge states and natural isotopes. Biopolymers such as peptides, proteins, carbohydrates, DNA, etc., are normally composed of carbon, hydrogen, oxygen, nitrogen, sulfur, and phosphorus. Except for phosphorus, these elements contain small fractions of less abundant stable isotopes, with the consequence that mass spectra of compounds exhibit clusters of isotopic ions. Despite the fact that theoretical isotopic distributions are prescribed by elemental compositions, the experimental determination of these distributions is hindered by inaccuracies associated with the m/z measurement, the dynamic range, and the finite resolving power of mass spectrometric measurements, especially in cases of high molecular weight compounds or a mixture of compounds spanning overlapped m/z ranges.

In spite of these difficulties, stable-isotope labeling methods are used to reveal some of the detailed cogent features of biopolymers.[8,9] The mass spectrum of a mixture of stable-isotope-labeled and unlabeled compounds gives distinct masses (m/z values) in which labeled ions are separated from their natural counterparts but labeled and unlabeled ions are observed with the same ionization efficiency. Thus, it is possible to achieve relative quantification analysis between a labeled and unlabeled pair of samples, or absolute quantification, when either a labeled or unlabeled internal standard is used.[10]

Relative quantification by MS represents a promising application in proteomics for revealing the differential expression level of proteins, such as those obtained from distinct stages of tissue and cell growth, etc., in which the precise ratios of stable-isotope-labeled and unlabeled ions can be directly retrieved from a mass spectrum.[11]

Hydrogen/deuterium (H/D) exchange, which is readily monitored by MS, has been used for the study of high-order structural features[12] and the dynamics of a protein or a

*Correspondence to: J. Fernández-de-Cossio, Center for Genetic Engineering and Biotechnology, P.O. Box 6162, Havana, Cuba. E-mail: jorge.cossio@cigb.edu.cu
**Correspondence to: T. Takao, Institute for Protein Research, Osaka University Yamadaoka 3-2, Suita, Osaka 565-0871, Japan. E-mail: tak@protein.osaka-u.ac.jp
Contract/grant sponsors: Center for Genetic Engineering and Biotechnology of Havana; Special Coordination Fund for the Promotion of Science and Technology; Ministry of Education, Culture, Sports, Science and Technology of Japan; Ministry of Health, Labour and Welfare of Japan.

complex of proteins and/or other compounds.[13] Using this technique, changes in the number of exchanged deuterium or back-exchanged hydrogens, as the result of solvent accessibility of regions of a protein that could be driven by a conformational change or interactions with other proteins, ligands, etc., can be precisely determined. The simultaneous determination of H/D exchange rates and the number of hydrogens involved in the exchange could be useful in studies of the dynamics of protein folding, aggregation, or interactions with other compounds.

As a result of the higher mass-resolving power of recently developed mass spectrometers, the abundances of individual isotopic peaks as well as the spacing between two adjacent isotopic peaks, which represents the charge state of an ion, could be obtained for moderate sizes of molecules (<5000 Da) by routine analysis. Thus, the analytical strategies discussed above, based on stable-isotope labeling, are clearly feasible and are currently in common use. Nevertheless, resolving complex patterns of isotope peaks is not an easy task. This is particularly true for partially H/D-exchanged ion peaks, which normally display a complex isotopic ion cluster with a 1 Da mass spacing in a broader mass range. Several stand-alone and web application software programs[14] aid in the calculation of the isotopic distribution of an ion, as well as the theoretical mass, based on an input sequence or a molecular formula. One of these has recently been made available from our sites;[15] this program is particularly useful for calculating and visualizing complex isotope envelopes.[16] Although such software programs are useful for validating the purity of an analyte by visual comparison with the theoretical isotopic distribution, they do not automatically resolve a complex mixture into its constituents when observed within an isotope envelope, such as ^{18}O-labeled compounds, partially H/D-exchanged proteins, etc.

Here we report on a software package, '*Matching*',[17] that is capable of automatically matching the ion peaks in a raw mass spectrum with the theoretical components derived from a peptide/protein, or a mixture of compounds. Stable-isotope enrichment, chemical modification, and metal ion adducts as the constituent components, are all taken into consideration. Furthermore, the isotope envelope containing enriched isotopes in addition to natural ones can be resolved into the constituent species, and their relative abundances can be readily determined. This capability is especially useful for the comparative analysis of the relative quantities of paired samples, based on stable-isotope labeling, in the context of emerging methodologies for quantitative and expression proteomics, in which MS is frequently used. The performance and application of the software is demonstrated for ^{18}O-labeled peptides which typically give a complex mixture of unlabeled, ^{18}O-, and $^{18}O_2$-labeled species, and the electrospray ionization tandem mass spectrum (ESI-MS/MS) of a large peptide, which is comprised of many types of fragment ions observed with various charge states.

EXPERIMENTAL

Samples

Peptides were obtained from Peptide Institutes Inc. (Osaka, Japan). A peptide was esterified by treatment with 0.2 M HCl/50 atom % ^{13}C-methanol and allowed to stand at 25°C for 4 h. ^{18}O-Labeled peptides were generated by the digestion of horse myoglobin (Sigma-Aldrich) with lysylendopeptidase (LEP) at a substrate/enzyme ratio of 10:1 in buffer containing 10, 30, 50, 70, 90, and 97.5 atom % $H_2^{18}O$ (ISOTEC Inc., Miamisburg, OH, USA) at 37°C for 10 h. Human urine samples were separately collected from normal volunteers and subjected to gel-permeation chromatography. Protein fractions were dried, reconstituted in either water or 97.5 atom % $H_2^{18}O$, and digested with LEP.

Nano-flow liquid chromatography (LC)

An UltiMate Nano LC system interfaced with a FAMOS microsampling workstation (LC Packings, Amsterdam, The Netherlands) was used in this study. A C_{18} trapping column (LC Packings) was used (at a flow rate of 30 μL/min) to concentrate the sample prior to separation in an analytical C_{18} column (75 μm i.d. × 15 cm, LC Packings). The mobile phase consisted of 0.05% trifluoroacetic acid (TFA)/H_2O (solvent A) and 0.05% TFA/CH_3CN (solvent B). LEP digests of urinary proteins were separated with a linear gradient of 10–70% of solvent B at a flow rate of 200 nL/min.

Matrix-assisted laser desorption/ionization mass spectrometry (MALDI-MS)

MALDI-MS spectra were obtained with a 4700 proteomics analyzer (Applied Biosystems, Framingham, MA, USA). Ions were generated by irradiating the sample area with a 200 Hz Nd:YAG laser operated at 355 nm. Spectra were obtained by accumulating 900 consecutive laser shots. Solutions (0.5 μL) of peptides (ca. 1 pmol) were mixed with the matrix solution, the supernatant from a 50% acetonitrile solution saturated with α-cyano-4-hydroxycinnamic acid (CHCA), and then air-dried on the flat surface of a stainless steel plate. Calibration was performed using $[M+H]^+$ ions of a mixture of angiotensin I (*m/z* 1296.6), dynorphin (*m/z* 1604.0), ACTH (1-24) (*m/z* 2932.6), and β-endorphin (*m/z* 3463.8).

Electrospray ionization (ESI) mass spectrometry

ESI mass spectra were obtained using a Q-TOF mass spectrometer fitted with a Z-spray™ nanoflow electrospray ion source (Micromass, Manchester, UK). Samples were dissolved in 1% aqueous acetic acid/methanol (1:1, v/v), and loaded into a borosilicate nanoflow tip (Micromass, Manchester, UK). Calibration was performed using cluster ions derived from NaI.

Software

'*Matching*' is a .NET web application, developed using the Microsoft Development Environment Visual Studio .NET, version 7.0 (Copyright © 1987–2001, Microsoft Corp.). '*Matching*' is coded mainly in C++, C# and ASP .NET, using the Microsoft .NET Framework software development kit (SDK) (Copyright © 1987–2001, Microsoft Corp.). '*Isotopica Viewer*' was developed using Borland® Delphi™ Studio Enterprise version 7.0 (Copyright © 1983–2002, Borland Software Corp.).

Algorithm

The isotopic distribution of a compound arises from the natural convolution of the individual isotope abundances of the

elements of which it is comprised. Finite instrumental resolution involves further convolution with the response function inherent to the measuring process and other 'random' events that contribute to the final peak shape.[18] Convolution is more efficiently calculated by multiplication in the Fourier domain followed by inverse Fourier transformation to the original domain, using fast Fourier transform algorithms.[19]

Since the isotopic distribution of a molecule entails the same structure irrespective of the charge state, with only variations in peak spacing and width, in this approach the Fourier transform of a molecular formula is performed only once while the peak shape is multiplied at a different width for each charge state. When the peak shape transform can be expressed analytically as a function of the parameters of the mass domain in the corresponding m/z range, a Fourier transform becomes unnecessary for each charge state. In the present approach we assume a Gaussian peak shape, which remains Gaussian in the Fourier domain, thus saving time at the expense of matching inaccuracy when the Gaussian does not sufficiently approximate the actual peak shape. The calculated isotopic distributions of the compounds in a mixture are combined so as to fit the raw spectrum to within the extent of the noise. The proportions of components are estimated by minimizing the χ^2 measure of the difference between the combination of the calculated isotopic distributions and the experimental raw data, finding the solution with maximum entropy.[20] To calculate the χ^2 difference, the raw experimental data are interpolated at each point of the equally spaced sample interval used for the discrete Fourier transform.

Input format

Amino acid sequences, carbohydrates, nucleotide sequences and molecular formulae can be specified in rows. Negative subindices indicate the removal of elements. For example, deamidation of Asn to Asp in the sequence ATDSNGSR can be specified by 'ATDSNGSR, $[NH_2]$-1 OH' separated by a comma, where one NH_2 group is substituted by one OH group.

Elements with other than natural isotopic distributions are registered on a per-user basis, and are maintained and updated in a custom library. This, combined with the negative indices in a formula, is sufficiently flexible to define a wide-range labeling scheme. For example, an element representing 2.5% ^{16}O and 97.5% ^{18}O atoms, derived from 97.5 atom % $H_2^{18}O$ used in this study, can be specified by an 'O*' symbol via the 'Elements' library. Peptides with a natural isotopic distribution are specified by their sequence of amino acids. ^{18}O-Labeled peptides are specified by the sequence followed by O-1O* indicating the removal of one O (oxygen) and the addition of one O* (user-defined element). The raw mass spectrum is inputted from an ASCII file that obeys the common format of m/z-abundance pairs (m/z and abundance separated by one space) per row.

Visualization of spectral superposition

A file containing the raw spectrum, the calculated theoretical spectrum of each contributing species, and the resulting estimated envelope is temporarily stored in a server. The superposition of the raw and calculated spectra can be visualized with the stand-alone application, *'Isotopica Viewer'*.[15] This application is intended to be previously downloaded and installed locally on a Microsoft Windows PC. A summary of the calculation is displayed in the HTML response page, with a link to the temporal file, the extension of which is associated with the Viewer.

RESULTS AND DISCUSSION

Matching of isotope peak patterns

Isotopic peaks of natural compounds observed in a mass spectrum are a major concern when validating the purity and determining the mass of a compound. Several software programs have been released for the deconvolution of the isotopic peaks to a monoisotopic one. However, such spectral conversion detracts from the reliability of mass determination. A close examination of the original raw spectrum is crucial for the final validation of the compound being analyzed. The present software, *Matching*, can automatically match the theoretical isotopic distribution directly to the original raw spectrum, even one that contains multiple components with close molecular masses within an isotope envelope. Figure 1 shows the screen dump of the output obtained for the superposition of a triply charged ion at m/z 1155.3 with the theoretical natural isotopic distribution of β-endorphin (M: 3462.8 (monoisotopic mass)), resulting in an imperceptible difference between the two. The result clearly shows that the automatic matching was successful, based on the relative abundances of the natural isotopes of the elements involved.

Stable-isotope labeling

Mass spectrometry is not, intrinsically, a quantitative technique, in the sense that the relative abundances of the ions do not necessarily reflect the relative abundances of different analytes in a sample. This is mainly due to differences in ionization efficiency. The use of stable-isotope labeling permits

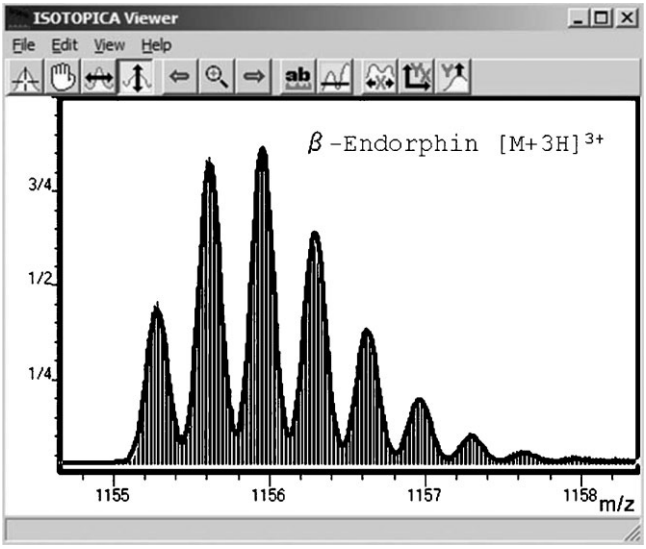


Figure 1. Screen dump from an output of the result for the triply charged ion signal of a 31 amino acid peptide (3463.8 Da) obtained by ESI-MS. The matching result (bold line) is superposed on the raw spectrum.

the comparison and relative quantification of the same analyte under two sets of conditions. However, since a single stable-isotope labeling with ^{13}C, ^{2}H, ^{18}O, ^{15}N, etc., shifts the molecular mass by only 1 or 2 Da, it is usually difficult to estimate the relative abundances of the labeled and unlabeled ion species owing to the overlap of the natural isotopes, and moreover they are resolved with great difficulty in the higher *m/z* range. Although this drawback has been overcome by the use of reagents in which multiple stable isotopes are embedded,[11] or by the *in vivo* incorporation of specific stable-isotope-labeled amino acids into biosynthetically labeled proteins,[21] special labeling protocols or limited production conditions are required. Thus, simple and straightforward labeling techniques that include the use of $H_2^{18}O$,[22] and $(CD_3CO)_2O$,[23] which shift the labeled species by a few mass units, could be useful when the overlapped natural and artificially enriched isotope peaks are precisely resolved into labeled and unlabeled species.

To test the capability of the software in resolving and quantifying species that are separated by 1 Da in molecular mass, the following experiment was conducted. A multiply stable-isotope-labeled peptide was obtained by esterifying five potential γ-carboxyl (Glu) and one β-carboxyl (Asp) groups with 50 atom % ^{13}C-enriched methanol. For an equal probability of esterification occurring at six possible sites, the more favored distribution in abundance of the species with different contents (m) of ^{13}C would follow a binomial distribution. Moreover, the binomial distribution becomes proportional to the binomial coefficients $\binom{6}{m}$ because of the equal proportions of ^{12}C- and ^{13}C-enriched methanol in the esterification media.

The MALDI mass spectrum of the methyl-esterified peptide showed a complex envelope resulting from the superposition of the six isotope distributions that were spaced incrementally from one another by 1 Da (Fig. 2). Each of the observed isotope peak profile(s) were individually deconvoluted into $^{13}C_{0-6}$-containing ion species by the software. Because of the same ionization efficiency expected for each isotopic ion species, the relative abundances obtained for each ion species were in good agreement with the theoretically anticipated values.

^{18}O-Labeling and its application for quantitative analysis of relative protein abundance

In the context of an expression profiling analysis of proteins using stable isotopes, it has become more important to estimate the relative abundances of isotopic ion peaks, especially those observed closely in a narrow *m/z* range. ^{18}O-Labeling in conjunction with enzymatic digestion, which can easily be achieved using a simple protocol, but results in shifts in mass of only 2 or 4 Da, has been used for quantitative comparison(s) between two paired samples. However, in an ordinary proteolytic digestion in $H_2^{18}O$ (97.5 atom % in the present study), a mixture of $^{16}O_2$, ^{16}O ^{18}O and $^{18}O_2$ species appear in peptide-dependent proportions. When mixing the labeled sample with the unlabeled one for a quantitative comparison, the $^{16}O_2$, ^{16}O ^{18}O, and $^{18}O_2$ species overlap in the isotope envelopes, and it becomes difficult to resolve them in the higher *m/z* range, thus rendering the method ambiguous, although, when using full ^{18}O-labeling conditions,[22] the method has been reported to give the correct relative ratios between $^{16}O_2$ and $^{18}O_2$ species.

In order to assess the viability of *Matching* in executing a relative quantification using the ^{18}O-labeling scheme, a MALDI experiment involving two peptides ($[M+H]^+$ at *m/z* 1360.7 and 2859.5) derived from myoglobin was specially designed to allow the ^{18}O-labeling to reach equilibrium in the presence of various concentrations of $H_2^{18}O$ (Experimental). In this regard, the two chemically equivalent oxygen atoms at the C-terminal α-carboxyl groups, produced by enzymatic cleavage, are exchangeable with $^{16}O/^{18}O$ during the course of the digestion by the catalytic effect of the enzyme that can

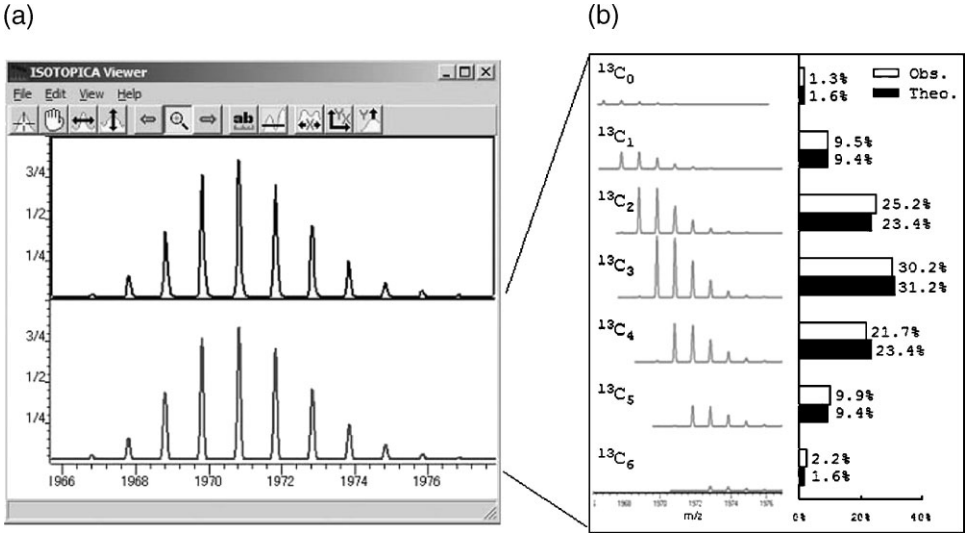


Figure 2. Comparison between the $[M+H]^+$ ion in MALDI-MS of the $^{13}C_{0-6}$-methyl-esterified peptide (AEEETAGGDGRPEPSPRE-NH_2) (upper panel in (a)) and the matching result (lower panel). The resulting peak was resolved into the respective ion peaks with $^{13}C_{0-6}$ (left panel in (b)), and their relative abundances (right panel in (b)) were calculated by *Matching*, based on the observed isotope envelope.

re-associate with the substrates. This condition was proved by the observation of nearly 100% of the 4-Da shifted $^{18}O_2$ species when it was prepared in 97.5 atom % $H_2^{18}O$ (Figs. 3(a) and 3(b)). The contents of ^{18}O for each preparation were simply estimated for the $^{16}O_2$, $^{18}O^{16}O$, and $^{18}O_2$ species according to the proportions of p^2:2p(1−p):$(1-p)^2$, respectively, where p is [^{16}O (atom %)/(^{16}O+^{18}O)] (dashed lines in Fig. 3(a)). Based on the raw spectra, the software generated the peak profile so as to produce the best fit to the spectrum, which was reconstituted by the possible components, in this case $^{16}O_2$-, $^{16}O^{18}O$-, and $^{18}O_2$-labeled ion species, and their relative abundances (Fig. 3(a)). As a result, the best fits for each preparation revealed the relative abundances of these three ion species, which are shown in Fig. 3(b). The proportions of the three distinct isotope components in each preparation were in good agreement with the theoretical plots for each component.

The software, which is capable of resolving isotope peaks even with the narrow-spaced split shown above, readily permits the relative quantities of paired samples to be estimated by comparison using ^{18}O-labeling, which can be prepared by enzymatic digestion either in ordinary water or $H_2^{18}O$. The software provides the following advantages. (1) It does not require a 4 Da shift ($^{18}O_2$-labeled species) for one of the paired protein pools, whereas methodology based on the complete ^{18}O-labeling of the two oxygen atoms at α-carboxyl groups requires special care in sample preparation; the reaction media must be almost pure ^{18}O-labeled water and exhaustive conditions are required for the enzymatic digestion to ensure the steady equilibrium for all peptides, which might render the procedure prone to unspecific cleavage. (2) It permits the estimation of the relative abundances of unlabeled, ^{18}O-, and $^{18}O_2$-labeled species, even when their ions are overlapped owing to their natural isotopes in higher *m*/*z* ranges. This capability has also been demonstrated for a complex mixture of a multiply ^{13}C-methyl-esterified peptide (Fig. 2).

The capability of a combination of the software and ^{18}O labeling as described above was applied to the quantitative comparison of paired protein pools prepared from human urine samples. The urinary protein fractions, isolated from two individuals, were separately digested with lysylendopeptidase in a buffer prepared with either ^{18}O-labeled or normal water. Equal volumes of both samples were then mixed, and separated by a reverse-phase nano-flow LC system. All fractions were directly collected on a MALDI sample plate and subjected to MALDI-MS and MS/MS, the latter of which was performed to identify the peptides. The raw mass spectra were deconvoluted with the software into the constituent ion species, i.e. $^{16}O_2$-, $^{16}O^{18}O$-, and $^{18}O_2$-labeled ions. Figure 4(b) shows a typical example obtained for a peptide from collagen type1-α, which gave the proportions of each ion species as 27, 53, and 20%, respectively. Thus, this protein was deduced to be present in each individual urine sample in the ratio of 1 ($^{16}O_2$):2.7 ($^{16}O^{18}O + {}^{18}O_2$) (column 7 in Fig. 4(a)), based on the proportions obtained above. While the peptides (columns 1–5) showed a nearly equal ratio of unlabeled to ^{18}O-labeled peaks, the above peptide and another one (1:1.7, column 6) gave ratios that were distinct from the others, which can be attributed to differences in the amounts of these proteins in the urine samples used in this experiment. The relevance of such differences in protein secretion levels in urine to the physiological status of the volunteer needs to be clarified.

Overall matching of the MS/MS spectrum to a peptide sequence

The above capability of matching of isotope peak distributions can be applied to the automatic interpretation of a complex mixture of ion peaks, such those derived as MS/MS fragment ions or as a mass spectrum of a protein digest. It allows the prompt validation of the spectrum on the basis of the input components of sequence or fragment ions, leading to the reliable identification of a peptide or protein and the finding of unassigned peaks as the result of various unknown modifications, etc. Figure 5 shows a raw ESI-MS/MS spectrum from a quadruply charged ion at *m*/*z* 866.7 for β-endorphin. The spectrum, comprised of peaks of a variety of fragment ions and charge states of ions, was directly subjected to automatic sequence matching by the stand-alone software, which can automatically produce the theoretical fragments and match them with a whole spectrum, based on the same algorithm of the present software. The assignments of each ion peak and their matching to the sequence were achieved based on an input sequence and the types of fragment ions observed in the MS/MS. As a result, almost all of the observed ion peaks fitted well with the expected ion peaks that had been calculated from the sequence. In addition, the isotope envelopes of each ion peak agreed well with the theoretical isotopic distributions (Fig. 5). It was feasible to fit a MALDI or ESI mass spectrum of a protein digest to the sequence, based on the cleavage sites of the specific enzyme used (data not shown). This overall matching functionality of the software should be useful for the validation of a raw spectrum along with a candidate sequence such as one resulting from a database search.

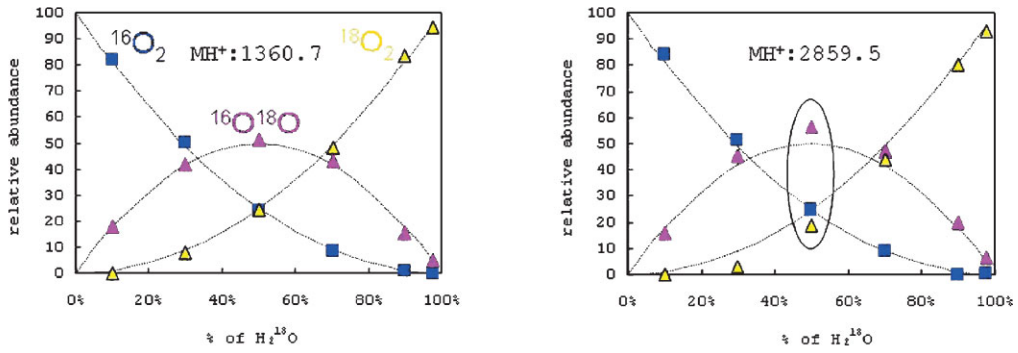
CONCLUSIONS

The developed software allows the automatic superposition of the observed ion peak to the continuum envelope of the theoretical isotopic distribution which is calculated, based on an input sequence or a molecular formula. In addition, it allows the relational analysis of stable-isotope-labeled and unlabeled compounds, even those which only generate a narrow mass spacing when labeled with $^{16}O/^{18}O$ (2 or 4 Da) at the C-terminal carboxyl groups or CH_3CO/CD_3CO (3 Da) at free amino groups,[23] both of which have been proved to be useful for the comparative analysis of the relative abundances of paired protein pools. In particular, ^{18}O-labeling has the advantage of the simple preparation of a labeled sample by the enzymatic digestion of a protein in a buffer containing ^{18}O-labeled water. Despite the complexity arising from the incorporation of one or two ^{18}O atoms into carboxyl groups, the software was able to make a comparative quantification with ^{18}O-labeling more precise and easy, which allows the calculation of the ratio of the sum of ^{18}O- and $^{18}O_2$-labeled species to an unlabeled one.

As shown in the experiment in which $^{13}C_{0-6}$-labeled peptides were used, the software was able to reveal the

RCM

(a)



(b)

% of $H_2{}^{18}O$	10%	30%	50%	70%	90%	97.5%
MH⁺:1360.7						
$^{16}O_2$	82.1	50.4	24.1	8.5	0.8	0.2
$^{16}O^{16}O$	17.9	41.6	51.4	43.1	15.4	5.2
$^{18}O_2$	0.0	8.0	24.5	48.4	83.8	94.6
^{18}O total	17.9	49.7	75.9	91.5	99.2	99.8
MH⁺:2859.5						
$^{16}O_2$	84.3	51.7	24.0	8.7	0.7	0.7
$^{16}O^{16}O$	15.7	45.2	56.2	47.3	19.3	6.5
$^{18}O_2$	0.0	3.1	19.8	44.0	80.0	92.8
^{18}O total	15.7	48.3	76.0	91.2	99.3	99.3

(c)

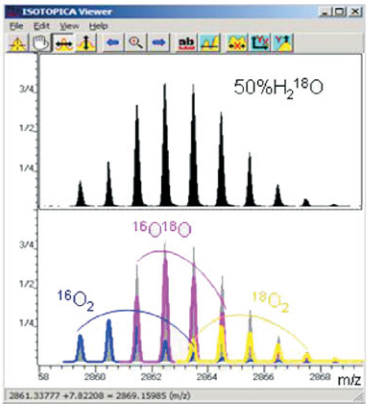
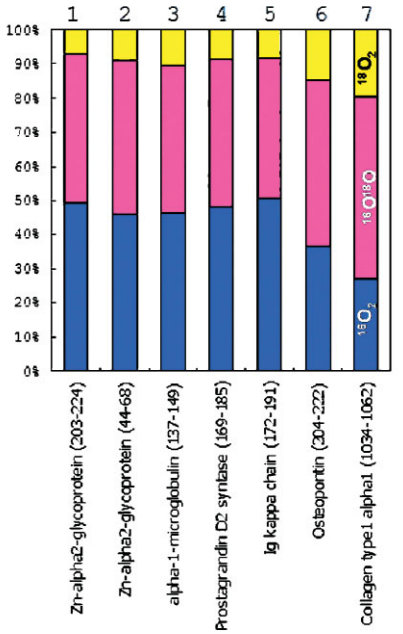


Figure 3. Analysis of the relative abundances of $^{16}O_2$-, $^{16}O^{18}O$-, and $^{18}O_2$-containing ion species in ^{18}O-labeled peptides (MH^+: 1360.7 (ALELFRNDIAAK) and 2859.5 (VEADIAGHGQEVLIRLFTGHPETLEK)), prepared by digesting myoglobin with lysylendopeptidase in a buffer containing 10, 30, 50, 70, 90, and 97.5 atom % $H_2^{18}O$ (a). The respective contents of the above ion species were obtained by matching them with the raw MALDI spectra by the software (b). For example, the points in the ellipse at 50 atom % $H_2^{18}O$ in the right panel of (a) were obtained by deconvolution of the raw spectrum (upper panel in (c)) into each isotope species, colored blue ($^{16}O_2$), pink ($^{16}O^{18}O$), and yellow ($^{18}O_2$) (lower panel). The dashed lines in (a) represent the theoretical proportions of $^{16}O_2$, $^{16}O^{18}O_1$, and $^{18}O_2$ species (see Results and Discussion).

(a)



(b)

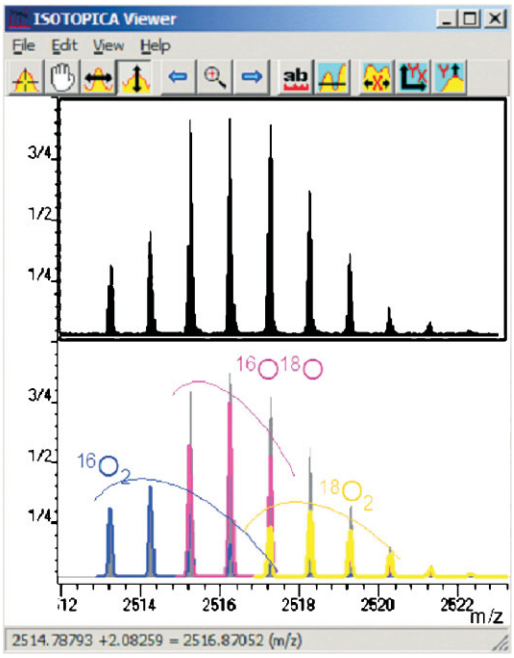


Figure 4. Relative abundances of several urinary proteins obtained from two normal volunteers. The proteins were probed by the peptides whose sequences, shown in parentheses, were analyzed by MALDI-MS/MS (a). The proportions of $^{16}O_2$ (blue), $^{16}O^{18}O$ (pink), and $^{18}O_2$ (yellow) species were obtained by matching them with the raw spectra. For example, the proportions for the peptide from collagen type1-α were obtained by deconvolution of the raw spectrum (upper panel in (b)) into each isotope species, denoted by the above corresponding colors (lower panel).

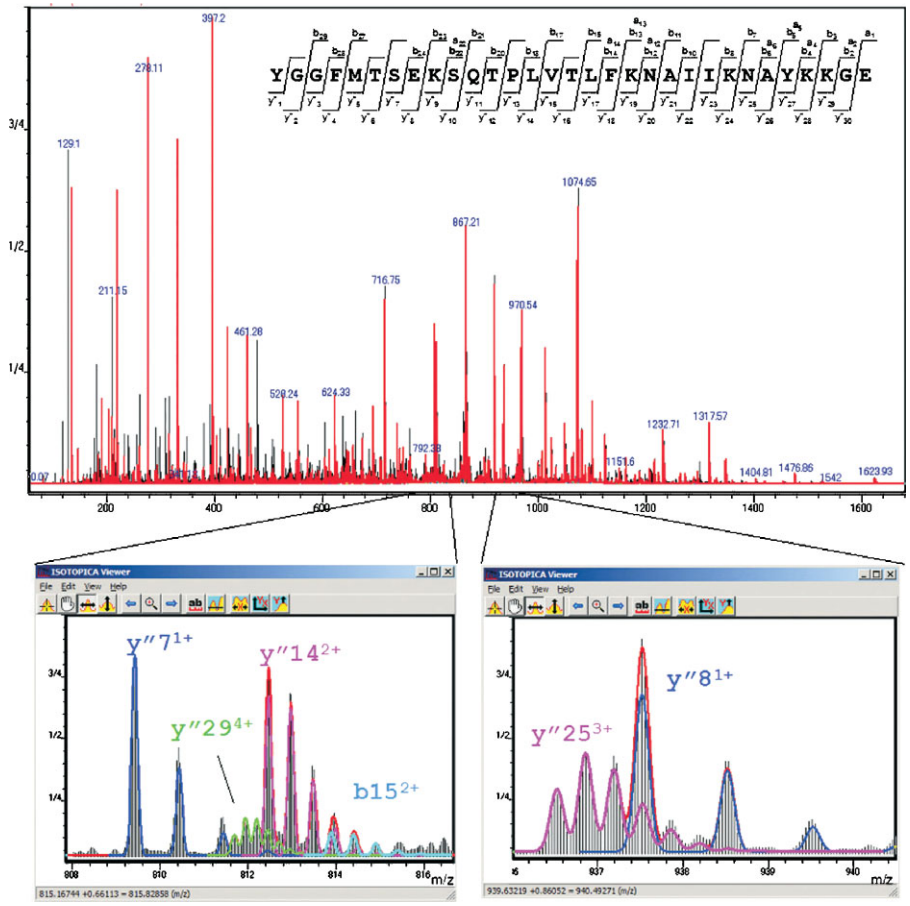


Figure 5. Overall matching of a raw ESI-MS/MS spectrum of the $[M+4H]^{4+}$ precursor at *m/z* 866.7 of β-endorphin with the sequence. The red lines in the upper panel denote peaks that matched the theoretical fragment ions (a, b, y″-series ions), prepared by a prototype of the stand-alone software program which uses the same model and algorithms as are used in the present Web application.[17] The signals shown as black lines, mainly derived from immonium or internal ions, were not used for matching since this version of the software did not contain these ions in the library. The lower two panels are expanded views of the matching results, produced by *Matching*, where each color trace shows the respective fragment ions shown in the sequence.

relative abundances of these seven possible molecular species that were separated by only 1 Da. The results suggest that the software is applicable to the analysis of H/D-exchanged proteins, which generates an array of ion species in MS separated by 1 Da resulting from the incorporation of deuterium at multiple sites on the molecule. The amount of exchanged deuterium, which reflects the accessible surface area of a protein to solvent, can also be estimated by the software.

Finally, the software has been shown to have the capability of overall matching of an MS or MS/MS spectrum of a peptide/protein by a peak-to-peak process, based on the input sequence, the enzyme used, types of fragment ions, etc. This capability will lead to the automated validation of a spectrum, in terms of masses and isotopic distributions of individual peaks, along with the sequence obtained by *de novo* sequencing or the output of a database search. It has proven to be especially powerful for the overall assignments of a variety of protein-derived fragment ions produced by ESI-MS/MS in a Fourier transform ion cyclotron resonance mass spectrometer (data not shown).

Matching[17] has been shown to be an effective tool for the detailed interpretation of a mass spectrum, which has often been omitted in an ordinary analysis in which either an automated protocol or a manual method for data processing is used. The direct comparison of observed ion peaks to the calculated isotopic distributions should lead to a more straightforward and reliable result, which will be quite useful for the final validation of an analyte sample.

Acknowledgements

This study was supported by research and development funds of the Center for Genetic Engineering and Biotechnology of Havana, the Special Coordination Fund for the Promotion of Science and Technology and Grant-in-Aid for Creative Scientific Research (15GS0320 to T. T.) from the Ministry of Education, Culture, Sports, Science and Technology of Japan,

and a project on 'Research on Proteomics' from the Ministry of Health, Labour and Welfare of Japan.

REFERENCES

1. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. *Electrophoresis* 1999; **20**: 3551.
2. Zhang W, Chait BT. *Anal. Chem*. 2000; **72**: 2482.
3. Mann M, Wilm M. *Anal. Chem*. 1994; **66**: 4390.
4. Clauser KR, Baker PR, Burlingame AL. *Anal. Chem*. 1999; **71**: 2871.
5. Taylor JA, Johnson RS. *Rapid Commun. Mass Spectrom*. 1997; **11**: 1067.
6. Fernandez-de-Cossio J, Gonzalez J, Betancourt L, Besada V, Padron G, Shimonishi Y, Takao T. *Rapid Commun. Mass Spectrom*. 1998; **12**: 1867.
7. Ma B, Zhang K, Hendrie C, Liang C, Li M, Doherty-Kirby A, Lajoie G. *Rapid Commun. Mass Spectrom*. 2003; **17**: 2337.
8. Rose K, Savoy LA, Simona MG, Offord RE, Wingfield P. *Biochem. J*. 1988; **15**: 253.
9. Takao T, Hori H, Okamoto K, Harada A, Kamachi M, Shimonishi Y. *Rapid Commun. Mass Spectrom*. 1991; **5**: 312.
10. Mirgorodskaya OA, Kozmin YP, Titov MI, Korner R, Sonksen CP, Roepstorff P. *Rapid Commun. Mass Spectrom*. 2000; **14**: 1226.
11. Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R. *Nat. Biotechnol*. 1999; **17**: 994.
12. Englander JJ, Mar CD, Li W, Englander SW, Kim JS, Stranz DD, Hamuro Y, Woods VL. *Proc. Natl. Acad. Sci. USA* 2003; **100**: 7057.
13. Akashi S, Takio K. *Protein Sci*. 2000; **9**: 2497.
14. http://prospector.ucsf.edu/ucsfhtml4.0/msiso.htm.
15. http://bioinformatica.cigb.edu.cu/isotopica/; http://coco.protein.osaka-u.ac.jp/isotopica/.
16. Fernandez-de-Cossio J, Gonzalez J, Satomi Y, Betancourt L, Ramos Y, Huerta V, Amaro A, Besada V, Padron G, Minamino N, Takao T. *Nucleic Acids Res*. 2004; **32**: W674.
17. http://bioinformatica.cigb.edu.cu/Matching; http://coco.protein.osaka-u.ac.jp/Matching.
18. Rockwood A, Steven L, Orden V, Smith RD. *Anal. Chem*. 1995; **67**: 2699.
19. Press WH, Teukolsky SA, Vetterling WT, Flannery BP. Numerical recipes in C. In *The Art of Scientific Computing* (2nd edn). Cambridge University Press: Cambridge, 1997.
20. Smith DL, Deng Y, Zhang Z. *J. Mass Spectrom*. 1997; **32**: 135.
21. Ong SE, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A, Mann M. *Mol. Cell. Proteomics* 2002; **1**: 376.
22. Yao X, Freas A, Ramirez J, Demirev PA, Fenselau C. *Anal. Chem*. 2001; **73**: 2836.
23. Regnier FE, Riggs L, Zhang R, Xiong L, Liu P, Chakraborty A, Seeley E, Sioma C, Thompson RA. *J. Mass Spectrom*. 2002; **37**: 133.