Sequence tag identification of intact proteins by matching tandem mass spectral data against sequence data bases

EIVIND MORTZ*, PETER B. O’CONNOR‡, PETER ROEPSTORFF*, NEIL L. KELLEHER‡, TROY D. WOOD‡, FRED W. MCLAFFERTY‡, and MATTHIAS MANN†

*Department of Molecular Biology, Odense University, Campusvej 55, DK-5230 Odense M, Denmark; ‡European Molecular Biology Laboratory, Meyerhofstrasse 1, D-69012 Heidelberg, Germany; and †Baker Laboratory, Department of Chemistry, Cornell University, Ithaca, NY 14853-1301

Contributed by Fred W. McLafferty, May 7, 1996

ABSTRACT  Molecular and fragment ion data of intact 8- to 43-kDa proteins from electrospray Fourier-transform tandem mass spectrometry are matched against the corresponding data in sequence data bases. Extending the sequence tag concept of Mann and Wilm for matching peptides, a partial amino acid sequence in the unknown is first identified from the mass differences of a series of fragment ions, and the mass position of this sequence is defined from molecular weight and the fragment ion masses. For three studied proteins, a single sequence tag retrieved only the correct protein from the data base: a fourth protein required the input of two sequence tags. However, three of the data base proteins differed by having an extra methionine or by missing an acetyl or heme substitution. The positions of these modifications in the protein examined were greatly restricted by the mass differences of its molecular and fragment ions versus those of the data base. To characterize the primary structure of an unknown represented in the data base, this method is fast and specific and does not require prior enzymatic or chemical degradation.

With the new genomic data bases of model species (1-3), such as Escherichia coli, Saccharomyces cerevisiae, mouse, and human, the sequences of many proteins of biological interest will in principle be known, and the problem of characterizing a protein primary structure will be reduced to identifying it in the data base and localizing differences between the putative data base sequence and the actual protein. Analysis at the protein level should be especially relevant for characterization of the products of genes expressed in a given cell line (4) and of proteins with variable mRNA processing or with posttranslational modifications, variations that cannot be identified at the DNA level. Such characterization will be important for correlation of genes with protein function and elucidation of biosynthetic and cell signaling pathways.

Mass spectrometry (MS) has recently become increasingly relevant to the analysis of biomolecules. Ionization techniques such as matrix-assisted laser desorption (5, 6) and electrospray ionization (ESI) (7, 8) have greatly extended the mass range for analysis of proteins and nucleotides. ESI combined with Fourier-transform (FT) MS (9, 10) appears to have unusually superior attributes for proteins up to 67 kDa (11), with 10^5 resolving power and <10^15 mol sample requirements (12-15). This can even be achieved for tandem MS (MS/MS) that can provide extensive sequence information by dissociating the multiply charged molecular ion with a variety of techniques (16-19) in the FTMS.

Extensive FTMS studies of proteins such as ubiquitin (8.6 kDa) (17, 19), carbonic anhydrase (29 kDa) (20), thiaminase I (42 kDa) (21), and creatine kinase (43 kDa) (22) show that most protein product ions are formed by amide bond cleavage to yield multiply charged ions containing the N terminus (B-ions) and/or C terminus (Y"-ions), with these dissociating further at higher energies to form internal ions (23). A pair of B or of Y" fragment ions whose masses differ by that of an amino acid thus indicates its presence in the protein and its mass separation from a terminus; a series of such fragments provides a partial amino acid sequence in the protein.

Within the past few years several research groups have demonstrated how MS can be used for identification of proteins in sequence data bases (24-33). One approach is to cleave the protein with a sequence-specific proteolytic enzyme, measure molecular weight values for the resulting peptide mixture by mass spectrometry, and search a sequence data base for proteins that should yield these values. Search algorithms have also been implemented recently that utilize low resolution tandem mass spectra of selected peptides (<3 kDa) from the protein degradation. Yates and coworkers (32) compared the MS/MS sequence data to the sequences predicted for each of the peptides that would be generated from each protein in the data base. In the PEPTIDESSEARCH sequence tag approach of Mann and Wilm (31), a partial sequence of 2-3 amino acids is assigned from the fragment mass differences in the MS/MS spectrum. This partial sequence and its mass distance from each end of the peptide (based on the masses of the fragment and molecular ions) are used for the data base search. Although these approaches have proved valuable in several laboratories (24-33), they require enzymatic digestion and often chromatographic separation prior to MS analysis. Typically, the peptide map used for an identification covers no more than 30-70% of the protein sequence (34) so that unknown modifications and amino acid substitutions cannot be determined in the remaining part of the protein.

In this paper the sequence tag search method is extended to the matching of MS/MS data from entire proteins (8-43 kDa) against a large sequence data base. Enzymatic digestion is not used, and some of the proteins contain sequence regions that do not match the data base sequence.

EXPERIMENTAL PROCEDURES

The MS and MS/MS data used (Table 1) are taken from Cornell studies of the specific proteins (11, 17, 19, 20, 22); the data for chicken cytochrome c has not been published previously. These references describe the protein sources, the six Tesla FTMS, and the experimental conditions. Data base searches were performed on a Power Macintosh computer using PEPTIDESSEARCH (24, 31, 35) for searches based on partial amino acid sequences, protein and fragment molecular weights, and combinations of these using a nonredundant sequence data base that contains more than 180,000 proteins.

Abbreviations: FTMS, Fourier-transform mass spectrometry; ESI, electrospray ionization; B and Y" ions, fragment ions from amide bond cleavage containing the N and C terminus, respectively.
sequences compiled from a variety of sequence data bases. (A nonredundant data base is maintained by C. Sanders at the European Molecular Biology Laboratory and can be downloaded from ebi.ac.uk/pub/databases/Peptidesearch.)

RESULTS AND DISCUSSION

For the molecular ion and selected fragment ions of the proteins, the Table 1 values represent the mass of the most abundant isotopic peak, the value used in these searches (number of 13C atoms in italics); misassignment of the most abundant isotopic peak will cause ±1 Da mass errors. The error probability increases with increasing mass and decreasing signal/noise ratio, with actual mass errors for the data used shown in Table 1. Unless otherwise stated, an error of ±2 Da was allowed in the searches reported here.

For identification, these mass data are matched against those expected from each protein of similar molecular weight in the nonredundant sequence data base sequence data base. Extending the concept of peptide sequence tags (31), partial amino acid sequences are sought by identifying fragment ion pairs whose mass differences correspond to that of an individual amino acid —NHCHRCo—; such contiguous mass pairs define an internal sequence. If these represent terminal B or Y ions (23), the smallest fragment ion mass defines the sequence mass position (m1, Fig. 1) with respect to one protein terminus and the difference of the fragment mass versus the molecular ion mass defines the position (m3) with respect to the other terminus. The tag data thus cover the whole protein; discrepancies between the actual protein and the data base sequence in the m1 part of the molecule will not prevent matching of the partial sequence and m3 data mass.

The MS/MS data of intact bovine ubiquitin (17, 19) show a molecular ion at 8564.66 - 5 Da and a series of three fragment ions (1320.69, 1449.75, and 1577.79) whose mass differences 129.06 Da and 128.04 Da correspond to the amino acids E (129.04) and Q (128.06) or K (128.09). The search string (1320.69) E(Q/K) (1577.79); 8564.66 was entered into the search program (Fig. 2). Because of the similar masses, the program considers this “either Gln or Lys”; Leu/Ile are handled similarly. In each search both B and Y type ion assignments for the data are considered. This search resulted in the unique retrieval of the bovine ubiquitin reference sequence when Y* ions were used, whereas no possible candidates were found for B-fragmentation (Fig. 2). The MS and MS/MS data of ubiquitin (17, 19) included >40 other fragment masses; comparison with the expected data showed masses corresponding to 10 B and 15 Y* ions to verify the complete primary structure. Note that the only remaining possibilities for differences between the primary structures of the data base and known proteins would be isomerizations (compensating mass shifts) between the backbone bonds cleaved in forming these fragment ions.

Many protein sequences in the database represent data automatically translated from cDNA sequences but which could have been modified in the expression of the mature protein. The mass spectra (Table 1) of intact human muscle creatine kinase (22) gave M1 = 42970.7 and a fragment ion series yielding the search string (6459.70) (I/L)Q(K/T) (6802.95); 42970.7. With the search restraint of M1 = 42971 ± 2, no possible candidates were found. However, considering only regions 2 and 3 (Fig. 1) with a 42- to 44-kDa restraints on M1, a match was found for the C-terminal data of the correct human creatine kinase sequence. The N-terminal part of the protein sequence tag (region 1) showed a mass 130.6 Da below that expected from the data base sequence; this mass corresponds to an additional methionine residue (131.04 Da). The lack of a DNA-predicted N-terminal methionine is common (36). All of the other five B ions found (22) showed the same mass shift within 0.3 Da to confirm the missing N-terminal Met; 10 Y* ions also confirmed (±0.2 Da) this primary structure. Such common processing of the leading methionine could also be detected by specifying “remove leading methionine” in the search program. The unusual specificity of these data is shown by searches containing only one residue in region 2. Thus, the string; (6701.87) T (6802.95); 42970.7, considering both B- and Y*-fragmentation, also retrieved only the correct creatine kinase, as did a search with the string (3857.33) V (3956.41); 42970.7.

As a further example, for the FTMS data on bovine carbonic anhydrase B (20) a search by the string (7045.18) TT(I/L) (7359.91); 29024.3 (Table 1) did not result in any possible candidates. However, a search using only regions 2 and 3 versus predicted Y*-fragmentation retrieved only the correct carbonic anhydrase. This was also true for a search using the string...
were found. Search proteins could be used for such sequence desorption forms of reference correct possibility. Although the unit resolution of FTMS is particularly important for detecting adduct ions that can introduce large mass errors (9). The sequence tag approach also appears promising for the identification of posttranslational modifications and even amino acid substitutions. The use of more complex and/or multiple protein sequence tags increases the specificity for localizing such modifications. The method can provide rapid and specific identification of proteins, including protein charting (37), without the need for enzymatic digestion. The much higher accuracy of the FTMS data than that required here should provide high matching specificity even with the highly expanded sequence data base that will result from the completion of the genome projects.

FTMS data used here were also measured by Daniel P. Little, Michael W. Senko, and J. Paul Speir. This study was supported by the Danish Biotechnology Program and the National Institutes of Health (Grant GM16609 and Training Grant 08-T2GM07273 to N.L.K.).