Error-Tolerant Identification of Peptides in Sequence Databases by Peptide Sequence Tags

M. Mann* and M. Wilm

Protein & Peptide Group, European Molecular Biology Laboratory (EMBL), Meyerhofstrasse 1, D-69012 Heidelberg, Germany

We demonstrate a new approach to the identification of mass spectrometrically fragmented peptides. A fragmentation spectrum usually contains a short, easily identifiable series of sequence ions, which yields a partial sequence. This partial sequence divides the peptide into three parts—regions 1, 2, and 3—characterized by the added mass \( m_1 \) of region 1, the partial sequence of region 2, and the added mass \( m_3 \) of region 3. We call the construct, \( m_1 \) partial sequence \( m_3 \), a "peptide sequence tag" and show that it is a highly specific identifier of the peptide. An algorithm developed here that uses the sequence tag to find the peptide in a sequence database is up to 1 million-fold more discriminating than the partial sequence information alone. Peptides can be identified even in the presence of an unknown posttranslational modification or an amino acid substitution between an entry in the sequence database and the measured peptide. These concepts are demonstrated with model and practical examples of electrospray mass spectrometry/mass spectrometry of tryptic peptides. Just two to three amino acid residues derived by fragmentation are enough to identify these peptides. In peptide mapping applications, even less information is necessary.

When sequencing peptides by mass spectrometry/mass spectrometry (MS/MS), one generally strives to obtain a complete sequence of the peptide investigated. The interpretation of a tandem mass spectrum is complex, however, and is done by an expert or a computer program. Practical examples, which often occur in low amounts and may contain residual detergent, usually lead to a list of most likely sequences. For unknowns, it is often difficult to be sure of the correct sequence without employing further chemical derivatization or a few steps of Edman degradation. Along with the instrumental demands and sample amounts needed, this has made MS/MS a tool that has not yet had the impact in practical problem solving in protein chemistry (i.e., as compared to automated Edman degradation) that it might have had.

Many protein chemical problems require more information than the masses of peptides alone but do not require complete sequencing of the peptide. For example, in mass spectrometric peptide mapping of large proteins, the peptide mass alone may not be a sufficient identifier because different peptides may have the same mass and because modified peptides may have unexpected masses. It would be attractive to use MS/MS if fragmented peptides could easily be correlated with the proposed sequence. It would likewise be attractive to use the exponentially growing content of sequence databases in problem solving by MS/MS. With the progress in genome sequencing projects and associated bioinformatics projects, such an approach would become increasingly general. Even today, a large and rapidly increasing percentage of proteins submitted for primary structure characterization can be found in sequence databases.

In our experience, even incomplete, complex, and noisy MS/MS spectra contain "islands" or "runs" of sequence ions that are clearly identifiable. They often serve as a starting point for the interpretation of tandem mass spectra of unknown peptides. Such a run of sequence ions—which we call a "peptide sequence tag"—also identifies with high confidence a peptide in the database.

In this paper, we develop the concept of searching by peptide sequence tags, and we demonstrate the power of the approach with electrospray MS/MS data recorded from low sample amounts of tryptic peptides. Such peptides can be identified with tags as short as two amino acids, and they can be located even in the presence of posttranslational modifications or a sequence difference between the measured peptide and the peptide in the database. Peptide sequence tags can be assigned easily in a tandem mass spectrum, and even an erroneous identification of a tag will not lead to a false positive identification. The current paper is a continuation of work started in Roepstorff's laboratory, which aims at comprehensive correlation of different kinds of mass spectrometric data with sequence databases. Henzel et al. originated the concept of searching a sequence database with a combination of tandem mass spectrometric data and a partial sequence.

(8) The concept of searching sequence databases by MS/MS data was first described in: Mann, M. Identification of Proteins in Sequence Databases using Molecular Weight and Partial Sequence Data. Presented at the 7th Symposium of the Protein Society, San Diego, 1993.
EXPERIMENTAL SECTION

The software program PeptideSearch10,17 was used for the investigations described here. PeptideSearch correlates mass spectrometric and sequence information with sequence databases. It allows searching by the complete molecular weight of a protein, by the masses of a set of peptides derived from sequence-specific cleavage of a protein, and by partial sequence information obtained by either Edman-type reactions or MS/MS data. PeptideSearch runs on the Macintosh series of computers and features an interactive user interface.

Various databases can be searched including SWISS-PROT, the EMBL nucleotide databases, the Protein Identification Resource (PIR) databases, and proprietary amino acid and DNA databases. For SWISS-PROT, completion time for all types of searches is less than 20 s on Macintosh Centris or Quadra computers. On the new Power Macintoshes, search times can be less than 5 s. Searches against a single sequence or a small set of sequences takes only fractions of a second.

The electrospray MS/MS spectra used to demonstrate the new algorithm were obtained on a Sciex API III mass spectrometer equipped with an updated collision cell (Sciex, Toronto, ON, Canada) and a MicroElectrospray source developed by us.20 Briefly, the MicroES source consists of a pulled glass needle with an inner diameter at the tip of 1–3 μm. The outside of the needle is gold-coated. Via a microscope, the tip of the needle is positioned 1–1.5 mm in front of the orifice of the mass spectrometer. The needle can be filled with 0.5–5 μL solution and sprays for more than 30 min when 1 μL is loaded, allowing extensive optimization and different types of MS/MS experiments even with minimal sample quantities. No pumps or valves are needed, and no sheath flow or nebulizer assists the spray. In low-level sequencing, cross contamination can be avoided by using a new needle for every new sample. Searches can be performed in “real time” because the acquisition times of the spectra are long compared to the time it takes to find a sequence tag and to search the database. Furthermore, for intermittent data interpretation the MicroES source can be halted by turning the spraying voltage off. SWISS-PROT, PIR I, PIR II, PIR III, and PatchX were searched with the “B ion” and the “Y-type ions” options.20 Masses are monoisotopic to 1800 Da unless otherwise noted. Mass accuracies were estimated from the signal-to-noise ratio of the spectrum and from the time since the last calibration. Percentage error or absolute error was specified. The mass range of proteins considered was unlimited but could have been narrowed down to improve sequence specificity.

Chicken lysozyme was from Sigma (EC 3.2.1.17; Sigma, St. Louis, MO Lot No. 111 H 7010) and other samples were from the core sequencing facility in our group. In the latter case, peptides were obtained by in-gel digestion of spots that had been excised from gels.22 The digestion mixture was separated by reversed phase HPLC on a 1.6 mm Vydac C8 column (Vydac 218TP, 5 μm, 250 × 16 mm, Serial No. 1213042R from MZ Analysetechnik, Mainz, Germany), and fractions were collected manually. Approximately 2–10% of the sample was used for mass spectrometry and 90–98% could be used for automated Edman degradation if desired. Sample quantities were estimated by UV absorption on the HPLC detector combined with the initial yield of Edman degradation where applicable. In the first example (Figure 2), the sample quantity of the undiluted stock solution of the HPLC fraction was determined by amino acid analysis.

THE ALGORITHM

The sequence tag algorithm will be explained by use of the hypothetical example in Table 1 and the illustration in Figure 1.

Table 1. Sequence Tag Extracted from a Hypothetical Tandem Mass Spectrum

<table>
<thead>
<tr>
<th>enzyme specificity</th>
<th>measured peptide molecular mass&lt;sup&gt;a&lt;/sup&gt;</th>
<th>run of sequence ions</th>
<th>type of ion series</th>
<th>partial sequence</th>
<th>mass of region 1 (m₁)</th>
<th>mass of region 3 (m₃)</th>
<th>search string entered into PeptideSearch</th>
</tr>
</thead>
<tbody>
<tr>
<td>trypsin</td>
<td>2111 ± 0.4</td>
<td>977.4, 1074.5, 1161.5</td>
<td>B series</td>
<td>PS</td>
<td>977.4</td>
<td>949.5</td>
<td>(977.4)PS(1161.5)</td>
</tr>
</tbody>
</table>

<sup>a</sup>Masses are monoisotopic and in daltons.
As indicated in Table 1, a part of a tandem mass spectrum contained a run of sequence ions of masses 977.4, 1074.5, and 1161.5 Da, yielding the partial sequence Pro-Ser. By itself this sequence is too short to be searched in a database.

The tag divides the peptide into three parts: regions 1 and 3 of which the added masses are known (m1 and m3) and region 2 of which the sequence (and the mass m2) is known. Note that the direction of the sequence is not necessarily determined, i.e., the sequence might be PS or SP. The exact number of amino acids in regions 1 and 3 is not known.

We first assume that the ion series is a B series. The added molecular mass m1 of region 1 is defined as the N-terminus of the peptide plus the sum of the amino acid residue masses up to the partial sequence (see Figure 1). The structure of B ions is H - (NH - CHR) - CO\(^{+}\), where \(\text{CHR}\) are the amino acid side chains; thus m1 is equal to the lowest molecular mass in the ion series, 977.4 Da in the example. m2 is the sum of the amino acid molecular masses in region 3 plus the molecular mass of the C-terminus. For B ions m2 is therefore simply the difference between the measured molecular mass of the peptide M and the highest mass in the ion series: m1 = 2111.0 Da - 1161.5 Da = 949.5 Da in the example. The complete sequence tag is thus m1 = 977.4 Da; PS; m2 = 949.5 Da. (Note that according to our definition m1 + m2 + m3 = M.)

The peptide has been derived by tryptic cleavage of a protein. Therefore, in addition to matching the partial sequence, m1 and m3, we require that the amino acid N-terminal to region 1 and the C-terminal amino acid in the peptide be cleavage sites for trypsin (see Figure 1). (Of course, the N- and C-terminus of the protein always match.) Table 2 lists the five criteria to be met by a matching peptide.

A search with the data in Table 1 results in a single match, namely, the peptide ASQSTETQGSSESGLMTVK of Serine/Threonine-Protein Kinase (Accession No. P18961) from which the data for this example had been constructed. The chance that this match is random is less than 3%, according to the equations developed in the Appendix. Since the direction of the sequence is not known, the search is repeated by assuming that the ions are of the Y\(^{+}\) ion type. Because the structure of Y\(^{+}\) ions is H2 - (NH - CHR) - CO\(^{+}\), m1 is the peptide molecular mass M minus the largest mass in the ion series and plus the mass of two hydrogens. In the example, m1 = 2111.0 Da - 1161.5 Da + 2\(m\)H = 951.5 Da. m2 is equal to the lowest mass in the ion series minus the mass of two hydrogens; m3 = 975.4 Da in this case. Furthermore, for Y\(^{+}\) ions the partial sequence has to be reversed before searching. There is no match for the sequence tag m1 = 951.5 Da; SP; m3 = 975.4 Da, demonstrating the discrimination of the algorithm.

All the above calculations, including the calculation of m1 and m3 according to the ion type chosen, are handled automatically by PeptideSearch. The user only enters a "search string" consisting of the lowest mass of the ion series, the partial sequence and the highest mass, and the peptide molecular weight. Searches by other fragment series types, as occur in high-energy collisions, can be performed by adding, for example, the molecular weight of CO (A series) or subtracting the molecular weight of NH (C series) from the masses in the ion series and proceeding analogously for the X and Z series.

Finding a match is only the first step of the procedure. As a second step, all expected fragment ion masses are calculated to verify the found sequence. If a number of possible sequences have been retrieved by the search, inspection of the tandem spectrum quickly reveals the correct match. Predicted fragment ion masses should either be present in the spectrum or their absence should make sense, for example, in the discontinuation of a series because of an internal positive amino acid. Information about the source organism of the matching peptide or other information about the protein can also be used. If the answer to a biological problem has to be based on a single tandem spectrum, the found sequence can be synthesized and fragmented to ascertain its correctness.

The database can be searched on all five or just some of the five criteria listed in Table 2. The first approach yields only exact matches while the second approach is error tolerant. For example, the program can be set to match peptides that have not been cleaved in a sequence-specific manner. More importantly, the algorithm can retrieve all peptides that match just two of the three regions. As the algorithm is implemented now, any mismatch can be tolerated that is localized to one of the three regions.

Consider an amino acid substitution in region 1 of the peptide in the above example. The total molecular weight of the peptide would be changed from the expected value and a search by all five criteria does not retrieve a match. We now specify "match regions 2 and 3" as well as "C-terminal specificity". The program will report peptides that contain the partial sequence "PS", whose amino acid sequence can be elongated from that partial sequence to yield the correct mass m2 and whose C-terminus is a tryptic cleavage site. There are 25 such peptides in SWISS-PROT. The predicted masses of fragment ions in regions 2 and 3 can be.

---

**Table 2. Search Criteria for MS Sequence Tags**

<table>
<thead>
<tr>
<th>match criteria</th>
<th>probability factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 C-terminal cleavage site of preceding peptide</td>
<td>1/10</td>
</tr>
<tr>
<td>(Arg or Lys)</td>
<td></td>
</tr>
<tr>
<td>2 mass of region one (at unit mass accuracy)</td>
<td>1/110</td>
</tr>
<tr>
<td>3 tag sequence (&quot;Pro Ser&quot;)</td>
<td>1/20 x 1/20</td>
</tr>
<tr>
<td>4 mass of region three (at unit mass accuracy)</td>
<td>1/110</td>
</tr>
<tr>
<td>5 C-terminal cleavage (Arg or Lys)</td>
<td>1/10</td>
</tr>
</tbody>
</table>

*See Appendix.*
generated for each of the 25 peptides and compared to the measured one. This procedure yields peptides which are identical to the measured one in regions 2 and 3. As a next step, the predicted masses of the MS/MS fragments corresponding to region 1 are compared to the measured ones to determine where the difference to the found sequence occurs. An amino acid substitution can rapidly be determined in this way. Even unknown or unsuspected posttranslational modifications can in principle be mapped to a single location.

The emphasis of this paper is on extremely short partial sequences; however, every additional amino acid in region 2 will increase the search specificity by a factor of about 20. For example, with a partial sequence length of three amino acids, only two matches would be obtained in the above example instead of 25. With longer partial sequences it becomes more important to consider amino acid mismatches also in region 2. Such mismatches are accommodated by selecting "match regions 1 and 3" and allowing for amino acid errors in the sequence pattern in region 2.

The peptide sequence tag cannot be more specific than the amino acid sequence itself. Thus a peptide sequence tag of a very short peptide with less than six amino acids will often not lead—as is the case with a short sequence of six amino acids—to a unique match in large sequence databases.

RESULTS AND DISCUSSION

Identification of Peptides That Are Identical to a Sequence in the Database. Figure 2 shows a section of the tandem mass spectrum of less than 25 fmol of an HPLC fraction of lysozyme diluted from a stock solution. Four sequence ions can readily be identified at first glance (376.2, 498.4, 617.4, 730.2), yielding the partial sequence (IL)(QK)(IL). Peptides with the same nominal molecular weight could not be distinguished; hence isoleucine/leucine were both allowed in the first and third positions and glutamine/lysine in the second position. The string (376.2)(IL)(QK)(IL) (730.2) was entered into the program together with the Database.

PeptideSearch searched with the sequence tag and a specified mass accuracy of ±0.4 Da; only the tryptic peptide NTDGST-DYGILQINSR from lysozyme was retrieved from the more than 100 000 entries in the sequence databases SWISS-PROT, PIR I, II, and III, and Patch X. The search was then repeated by assuming that the ions were of a B series and—as expected—no matches were found. A random match for this tag would have a probability of less than 1% in spite of the isoleucine/leucine and glutamine/lysine uncertainties (see Appendix). After identifying the candidate sequence, the tandem mass spectrum is compared to the predicted fragment ion masses for that sequence. In Figure 2b, the complete Y" ion series except the Y"16 ion is shown.

If the mass accuracy in the MS/MS measurement had been poor there would be more amino acid possibilities for the partial sequence. For example, if there had been a 1−2 Da uncertainty, isoleucine and leucine would be indistinguishable from asparagine and possibly from aspartic acid (nominal molecular weights of 113, 114, and 115, respectively), just as glutamine and lysine would be indistinguishable from glutamic acid (128 and 129, respectively). Thus, the following search string would have been obtained: (376.2)(LIND)(QKE)(LIND) (730.2). The amino acids in parentheses are all candidates for that position; i.e., the first and third positions of the partial sequence could be either isoleucine, leucine, asparagine, or aspartic acid. Even with this ambiguous partial sequence, only the correct lysozyme peptide is retrieved. Thus it is possible to uniquely match a peptide to the database even when the peptide sequence tag is very short and when the sequence tag contains many ambiguities.

A second example illustrates how a tag can be constructed under less clear-cut conditions than those of Figure 2. A peptide of about 2.7 kDa was isolated after tryptic cleavage of a protein whose identity was to be determined. A 1 μL volume of a solution containing 0.5 pmol of the peptide was injected into the MicroES needle, and several MS/MS experiments with different precursor ions and different collision energies were performed.

Fragmentation spectra of larger peptides are more difficult to interpret completely than those of small peptides. In addition to the problem of generating fragment ions to cover the whole sequence, the charge state of the fragments has to be determined and the mass determination of the multiply charged fragments must be more accurate to distinguish amino acids differing by one mass unit. Due to the overlapping distributions of charge states, more possibilities must be considered when the peptide sequence is being solved.

Figure 3 shows a section of the tandem mass spectrum just above the (M + 3H)3+ precursor ion. The six labeled masses in this section of the spectrum could have charge states of one or two. These charge states can usually be distinguished by their isotope spacing or by the peak shape of the doubly charged ions. However, even if all six ions are equally taken into consideration, there is only one possibility for a stretch of four sequence ions, namely, the doubly charged ions at 965.6, 1030.2, 1073.6, and 1154.8 Da/2.1. Multiplied by two, their mass differences yield the partial sequence ESY and the search string (1930.2)ESY(2308.6). With the measured precursor mass of 2758.3 Da and the assumption that the ions are of Y" type, the sequence tag m1 = 451.7 Da; YSE; m2 = 192.8 Da results. The databases were searched with the sequence tag and a specified mass accuracy of 0.04%. The sequence SSIYSSEEKDPDKKDIEHIGKL was uniquely retrieved, and the protein in question was identified as vaccinia virus protein A4 (Accession No. P20983) because the complete expected sequence of the peptide could be verified in the MS/MS spectra.

If the search criteria were relaxed by either requiring SY as a partial sequence or by not requiring that the peptide is tryptic, the peptide match would still be unique. Furthermore, there are numerous other, and better, possibilities to find sequence tags in the MS/MS spectra of this sample. The above one was chosen to illustrate that it is possible to find tags even if there is no obvious ion series present. If there are several alternative possibilities to construct tags, they can all be searched and the result matched to the MS/MS spectra. Such an iterative procedure is assisted by the speed of PeptideSearch.

Identification of Peptides That Are Not Identical to a Sequence in the Database. The previous examples show the power of the sequence tag method to locate peptides that are already in the database. Many of the most interesting cases, however, involve a change between the peptide as predicted in the database and the actual peptide. Two important instances are peptides with an amino acid difference from the proposed sequence—caused by DNA sequencing errors or by mutation—and posttranslationally modified peptides. The molecular weight of
these peptides does not agree with the predicted one and therefore cannot be used directly for the search.

In such cases, a putative protein sequence or family of sequences is often already available. Then the sequences to compare to are orders of magnitude fewer and the sequence tags become correspondingly more specific (see Appendix). With the increased specificity, peptides can be identified with even shorter partial sequence and with fewer of the match criteria in Table 2.

As a practical example, we deliberately consider a noisy and miscalibrated tandem mass spectrum of a tryptic peptide of
measured molecular mass of 1065.0 Da. The peptide had been isolated from an in-gel digest of a spot on a gel that had been identified by mass spectrometry to belong to the keratin protein family. In spite of the poor quality of the tandem mass spectrum, several pairs of peaks whose mass difference corresponds to an amino acid residue are immediately apparent. For example, the peaks at 573.8 and 702.8 Da yield glutamic acid as a partial sequence. A search of the sequence tag $m_1 = 364.2$ Da, E, $m_3 = 571.8$ Da with tryptic specificity against keratin sequences in SWISS-PROT reveals that such a peptide is not contained in any of them.

In a subsequent search, the same sequence tag is used but it is only required that the regions 1 and 2 match—i.e., that the partial sequence and $m_1$ match—and that the N-terminus of the peptide be preceded by R or K. Several keratins are then identified. Comparison of the retrieved partial sequence (AQYE)
with the tandem mass spectrum shows that regions 1 and 2 do indeed match and that the sequence can be elongated in the C-terminal direction up to the 866.0 Da fragment ion, resulting in the partial sequence AQYEDIA. (Recall that Y ions run in the direction opposite to the normal way of writing sequences and that isoleucine was set equal to leucine for all searches.) Thus, the peptide has been identified and the mass difference between the sequence in the database and the actual sequence has been mapped to the C-terminal two amino acids of the peptide. The remaining mass for that dipeptide is compatible with QK or KK instead of the RK found in the database. Indeed, a low-abundance fragment ion, indicative of the cleavage between QK or KK, can be found in the tandem mass spectrum. Thus the sequence AQYEDIA(QK)K can finally be assigned to the peptide.\(^{23}\)

As could be seen in this example, in a search against a single sequence or a protein family, the requirements on a sequence tag are minimal. For example, usually the cleavage condition does not have to be required even when only one amino acid has been determined by MS/MS. Thus, as long as even minimal MS/MS information can be obtained the peptide can be uniquely located in a sequence even in the presence of a sequence substitution or an unusual cleavage.

A partial sequence of a single amino acid length could easily be assigned erroneously in an MS/MS spectrum, for example, if the difference of two fragment masses results in an amino acid molecular weight even though the fragments are from different ion series. However—because of the specificity of the search and because of the subsequent matching of the retrieved sequence to the actual MS/MS spectrum—no false positive is likely to be obtained in such a case.

As a final example, a cysteine-modified lysozyme peptide is used to illustrate the procedure for identifying modified peptides, this time in the complete database. The peptide had a measured molecular mass of 1372.6 Da, and its doubly charged ion was dissociated. A clear stretch of sequence ions between the doubly and the singly charged ion resulted in the search string (782.3)-NG(IL)(1066.1) and the sequence tag \(m_1 = 308.5\) Da, (IL)GN, \(m_2 = 780.3\) Da. With a mass accuracy to \(\pm0.03\%\), no peptide was found in the databases. This indicates that the peptide was from a novel protein or was modified. Three searches, allowing errors in regions 1, 2, and 3 were now performed. On statistical grounds, random matches are quite possible for this sequence tag; therefore, careful analysis of the found sequence has to be performed. (If the partial sequence length had been four amino acids, no random match would have been expected (see Appendix).)

No match was retrieved when matching regions 1 and 3 by mass, requiring tryptic specificity but allowing any one sequence error in region 2. When \(m_2\), the partial sequence, and the tryptic cleavage site were required, a peptide was matched (IGNAVPVAL-GR). This peptide cannot be reconciled with the measured MS/MS spectrum and must have been a random match. When \(m_1\), the partial sequence, and the tryptic site were required, the peptides GYSLGNWVCAANYESGF..., GYSLGNWVCAAR..., and GYSLGNWVCAAK..., which occur in several lysozymes, were matched.

The agreement with the mass of region 1 and the existence of several predicted Y" ions immediately indicated that the sequence was indeed GYSLGN from the N-terminus. Comparing actual to predicted Y" ions for the next amino acids in the sequence revealed that the sequence was GYSLGNWV as predicted. These eight amino acids are enough to uniquely identify the peptide in the database, and the rest of the variation between the three peptides is likely to be due to differences between species. The next fragment ion corresponding to the cysteine residue was not found, but the C-terminal three amino acids could be verified...
as AAK. Thus, the sequence had to be GYSLGNWWXAAK, where X is a cysteine in the database but has an added molecular mass of 104 Da in the measurement; this is, of course, the mass shift expected for derivatization with vinylypyridine, a procedure which the protein had been subjected to.\textsuperscript{24} While this example is merely intended as an illustration, many posttranslational modifications can in principle be located using the same procedure. (It may, however, only be possible to locate the modification to within several amino acids if the required fragment ions are missing from the tandem mass spectrum.)

\section*{CONCLUSION AND PERSPECTIVE}

We have developed a way to connect mass spectrometry/mass spectrometry data with sequence databases. A short piece of
partial sequence together with the molecular weight of the preceding and trailing region of the peptide, here called a peptide sequence tag, is a unique signature of that peptide. Such a tag can be used to locate the peptide in a given sequence or in a sequence database. We have shown that no more than two to three amino acids are needed to uniquely retrieve tryptic peptides from sequence databases such as SWISS-PROT. A short stretch of amino acids is easily assignable in almost all MS/MS spectra. The likelihood that a match to a given sequence tag is random is easily calculated. In a second step—after finding one or more candidate sequences for the peptide—the predicted sequence ions of a candidate sequence verify the match. Two important cases of mismatches that will confound searches based on the peptide molecular weight alone—amino acid substitutions and posttranslational modifications—do not prevent the identification of the peptide. Using the sequence tag method, these modifications can rapidly be located in a sequence.

Because there are minimal requirements on them, the peptide sequence tags can sometimes be obtained even without dedicated or very sophisticated MS/MS instrumentation—for example by using “in-source fragmentation” in electrospray.

The term peptide sequence tag was inspired by the term expressed sequence tag (EST) coined by Venter et al. and used for the short 200–300 base pairs of sequence randomly collected from each gene in large-scale cDNA sequencing projects.25 Like the ESTs (and the sequence tagged sites (STSS) from which the term EST is in turn derived) a short sequenced region is used to characterize a much longer sequence such as—in the case of the ESTs—the whole gene product. The term furthermore reflects our hope to correlate peptide sequence tags with the ESTs that are now filling the databases at a rate of hundred thousands per year. In this endeavor, the error tolerance of the peptide sequence tag approach will be crucial because the error rate of ESTs is so high that the predicted mass of most peptides is likely to be wrong. ESTs are expected to soon cover most human genes; thus, if our strategy is successful, one could in principle easily identify or at least fingerprint most human gene products.

ACKNOWLEDGMENT

We thank the other members of the Protein & Peptide group at EMBL, especially Tony Houthaeve, Gert Talbo, and Roland Kellner, for their help. Steve Carr of Smith Kline provided helpful criticism on the manuscript.

APPENDIX

Probability of Finding Nonrandom Matches by Use of Mass Spectrometric Sequence Tags. The probability of matching a given amino acid in the tag sequence at any given position in the sequence database is equal to the number of amino acids at a given tag position \(i\) divided by the number of possible amino acids.

Thus, in the case of 1 amino acid out of 20

\[
P_{\text{ith tag position}} = \frac{1}{20}
\]  

The probability of randomly matching a given mass \(m_i\) with unit mass accuracy is equal to the mean molecular weight of the amino acids:

\[
P_{m_i} = \frac{1}{110}
\]

and likewise

\[
P_{m_j} = \frac{1}{110}
\]

The probability of matching a cleavage site at a given position is equal to the number of amino acids fulfilling the cleavage condition divided by the number amino acids

\[
P_{\text{correct cleavage condition}} = \frac{\text{(number of cleavage amino acids)}}{20}
\]

For example, for trypsin we have \(P_{\text{correct cleavage condition}} = \frac{20}{20} = 0.1\).

The chance of a false positive at any one amino acid position is the product of the probabilities of matching the individual requirements (eqs 1–4). If there are \(i\) amino acids in the partial sequence, the following equation holds:

\[
P_{\text{false positive}} = P_{\text{term correct cleavage}}P_{m_i}P_{\text{first tag position}} \times \cdots \times P_{\text{ith tag position}}P_{m_j}P_{\text{term correct cleavage}}
\]

Equation 5 expresses the chance of a random match at any amino acid position in the database. The chance of obtaining no random matches in a database containing \(n\) amino acids is then

\[
P_{\text{nonrandom match}} = (1 - P_{\text{false positive}})^n
\]

Since in general the direction of the partial sequence is not known, the search has to be performed in both directions and there is twice as large a chance to obtain a random match:

\[
P_{\text{nonrandom match}} = (1 - 2P_{\text{false positive}})^n
\]

As an example, at any amino acid position in the sequence database, a tag sequence with two amino acids, tryptic cleavage conditions, and unit mass accuracy result in the following probability of a false positive:

\[
P_{\text{false positive}} = \frac{2}{20} \times \frac{1}{110} \times \frac{1}{20} \approx 2 \times 10^{-9}
\]

A match in SWISS-PROT, which currently contains approximately \(1.5 \times 10^9\) amino acids, will then have a confidence level of greater than 97% of being nonrandom. A sequence database that was 100 times larger (as can be expected from the genome projects) would

References

require a sequence tag of three to four amino acids in length to match with a confidence of greater than 85%.

The above discussion ignores the fact that not all amino acids have the same likelihood of occurring in a given position; e.g., the extreme cases are \( \rho_{Cy} = 0.018 \) and \( \rho_{Leu} = 0.091 \) (in SWISS-PROT). However, all other amino acids are within a factor of about 2 of 0.05. In any case, the exact factors for single amino acids or dipeptides can easily be put into eq 1. Similarly, if a higher or lower than unit mass accuracy has been achieved, eqs 2 and 3 can be adjusted proportionally. Equations 5–7 hold in any case.

Received for review September 7, 1994. Accepted October 20, 1994.*

AC940892J