Application of Pattern Recognition Techniques to Mass Spectrometric Data for Sequencing C-Terminal Peptide Residue Series

Rupika Delgoda and James D. Pulfer*,†

Department of Chemistry, Box 320, University of Papua New Guinea, National Capital District, Papua New Guinea

Received July 28, 1992

The application of pattern recognition to sequence elucidation from fast atom bombardment (FAB), collisionally activated dissociation (CAD), and tandem mass spectrometric data of peptides was investigated. Learning machine techniques for pattern recognition were applied to detect C-terminal series amino acid sequences up to the pentapeptide Try-Gly-Gly-Phe-Leu (YGGFL). The approach conditions the data by building upon known fragmentation pathways of peptides in FAB/CAD-related analysis. The intensities of critical sequence ion peaks are used to describe each pattern in the training set. A well-defined training set is then made use of to classify unknown species. The FORTRAN-77 program is adapted from one first developed by P. C. Jurs. It requires only the input of the critical peak intensities and the length of the unit to be tested. The method has potential in applications to larger peptides provided a database for the training set can be constructed.

INTRODUCTION

Fast atom bombardment (FAB) mass spectrometry[1,2] in conjunction with collisionally activated dissociation (CAD) tandem MS/MS techniques[3-6] has been successfully applied to sequencing peptides.[7-13]

The objective of the current work is to apply pattern recognition[14-17] techniques to the above mass spectrometric methods in order to analyze the structures of unknown peptides when there is little or no accompanying background information. This can be done by merging two broadly based ideas: using known peptide sequences as training sets for classifying unknowns and noting similarities in peptides having similar sequencing in their termini. It has been found that if the intensities of the C-terminal fragmentation pattern sequence is used as the descriptors, correct classification of unknowns (up to the penta residue peptide leucine enkephalin) can be affected.

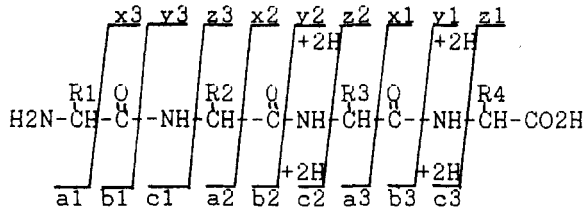

Figure 1. *xyz, abc* scheme for classifying peptide backbone fragments belonging to either the N-terminal or C-terminal sequences.

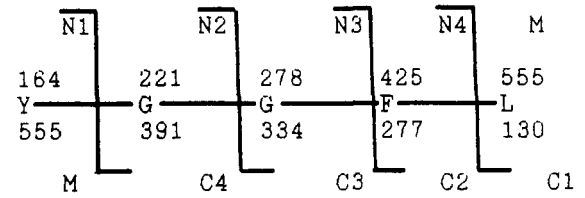

Figure 2. C- and N-terminal fragmentation pattern of leucine enkephalin.

METHODS AND MODELS

Fragmentation by CAD of a mother ion $(M + H)^+$ of a peptide takes place chiefly along the backbone, giving rise to fragmentation ions characteristic of the residues. These ions are summarized in Figure 1, which is a modification by Biemann[18] of the scheme first proposed by Roepstorff and Fohlman.[19] Ions resulting from the three possible cleavage points of the peptide backbone are labeled *a*, *b*, and *c* when the charge is retained by the N-terminal sequence and *x*, *y*, and *z* when retained by C-terminal ions. The most common cleavage takes place at the peptide bond, resulting in *b* and *y* ions. The oxocarbonium ion resulting from the *b* series turns out to be slightly less stable and, therefore, less reproducible than the quaternary nitronium ion. Thus, the C-terminal series proves to be the better set of descriptors. Apart from the backbone ions, some cleavage on side chains does occur[19,20] and can help to differentiate between isomeric leucine and isoleucine.[20] Current work is focused on using pattern recognition techniques to correctly classify the sequence of the biologically important pentapeptide species leucine enkephalin, YGGFL. The fragmentation pattern of this species is given by Katakuse[21] and is shown in Figure 2. Conventionally, the C-terminal ions start from the left, with primary cleavage occurring through the peptide bond, and are accompanied by the addition of two hydrogens so that, in fact, the C-terminal ions occur at mass values of 130, 279, 336, 393, and 555 Da. The unstable oxocarbonium ions of the N-terminal series result in a further loss of carbon monoxide and a resulting variable intensity for the N-terminal series, along with an accompanying proliferation of peaks. It was for these reasons that the C-terminal series was chosen to be the best set of descriptors.

Handling the vast array of peaks observed when larger peptides are analyzed in a difficult task. Several computer-assisted programs have been devised to overcome these problems. To run them usually requires expert analysis of the spectra concerned. For example, the algorithms developed by Siegel et al.,[9] Bartels,[10] and Ziemer et al.[14] start by identifying the significant N-terminal and C-terminal ion series and then use them to help predict primary structure. Others, such as MacProMass[11] and COMPOST,[12] compare known and unknown sequences to affect partial or complete identification. Many of these programs are developed along the

* Author to whom all correspondence should be addressed.

† Paper presented at the 10th International Conference on Computers in Chemical Research and Education, The Hebrew University of Jerusalem, Jerusalem, 91904 Israel, July 12-17, 1992.

0095-2338/93/1633-0332$04.00/0

following lines: a given molecular mother ion mass is used to calculate all possible N- and C-terminal ion sequence peaks using various mass differences derived from the 20 normally occurring amino acid residues. These results are then compared with known sequences and the given mass spectrum.

When an unknown peptide is analyzed with no background information provided, a program which could elucidate sequences using pattern recognition, and thus obviate the need for expert spectral analysis, would be of considerable benefit.

Pattern Recognition. There are occasions when describing a complex phenomena, such as the peaks of the mass spectrum of a large residue peptide, when all the variables must be considered together in order to present a complete picture. One or two parameters considered in isolation may be misleading because of functional interdependence. Also, apparent patterns may be disrupted or missed altogether if only a few descriptors are utilized. The data, then, has to be represented by all the affecting variables. Therefore, if n factors describe the data (n descriptors), then n-dimensional spaces are required to define it. Inherent is the assumption that when the points obtained in n-space are part of a common category, they will cluster together in a limited region.

Pattern recognition consists of a set method for investigating data, represented in the above manner, to assess the degree of clustering and the general structure of the data space. To separate the data into classes, a discriminant is developed which is capable of dividing the regions according to the degree of clustering observed. There are several ways of doing this. In supervised learning,[15] the unknown patterns belonging to a prediction set can be classified into two categories, using rules developed in the training set. In this work, a nonparametric method,[14] the linear learning machine (LM) technique[16] is employed to develop the discriminant from the data itself, in the training set, and to apply it to the prediction set.

Learning Machine Technique. The discriminant, developed as the initial decision surface vector (arbitrarily set), classifies each member of the training set. The descriptor is left unchanged if classification is correct and altered when an incorrect classification occurs. The weight vector $\mathbf{W}$ will change to $\mathbf{W}'$ if correction is needed such that

$$\mathbf{W}' = \mathbf{W} + C\mathbf{X}_i$$

where $\mathbf{X}_i$ is the ith pattern vector incorrectly classified and C is the correction term given by

$$C = -2s/(\mathbf{X}_i\mathbf{X}_i)$$

and s is the scalar $\mathbf{W}\mathbf{X}_i$.

The method is called "learning" because the weight vector, $\mathbf{W}$, improves with experience. The method works best, of course, when the training set size N becomes significantly larger than the number of descriptors in the n-dimensional space, a ratio of 3:1 being optimal.[16] Once the weight vector has been found which correctly classifies the training set, it is used to classify the prediction set. The results are in the form of classifying each member of the prediction set as belonging to either a "yes" or "no" class, e.g., "yes" YGGFL is present in the correct sequence or "no" it is not.

RESULTS

The pentapeptide leucine enkephalin was chosen as the test material. Its N- and C-terminal fragmentation sequences are well-characterized, and it provides enough complexity to serve as a relatively challenging and severe evaluation of the possibilities of pattern recognition techniques applied to sequence elucidation work.

Chart I. Screen printout for unknown peptide NEWDATA5 being checked for the pentapeptide unit L-F-G-G-Y using program LM

```
     ENTER THE UNIT OF YOUR CHOICE(MONO/DI etc.): PENTA

     FILE BEING EXAMINED :PENTA

           THE MAX NO. OF PARAMETERS IN DATA IS:


           THE VALUE OF NDATMAX  =   5

 ENTER THE NEW OUTPUT FILE NAME:
 NEWFILE5

 INITIALIZE RANDOM NUMBER GENERATOR= 345

  NUMBER OF TRAINING SETS, NTRSET:
 13

  CHOSEN NO OF PATTNS FOR PREDICTION SET
 10

  THRESHOLD VALUE IS:
 .1

  NUMBER OF PASSES ALLOWED:


 1000

  THE NUMBER OF DESCRIPTORS:
 5

           THE DESCRIPTORS USED:

           JJ1,JJ2,JJ3 ARE:
 1,5,1

  1  5  1

 THE MASS SPECTRUM OF AN UNKNOWN PEPTIDE CAN BE
              CHECHED FOR THE UNIT YOU ENTERED PREVIOUSLY.

  THE MASS VALUES CONSIDERED ARE:
              86 AND 130 (+/-2 Da) FOR LEU ONLY

              130,279,336,393,555 (+/-2 Da) FOR OTHERS.

 TO CHECK THE PRESENCE OF THESE A.A.ENTER THE PEAK
         INTENSITIES AT THESE M/Z VALUES.

 ENTER THE DATA AS I+01+X1+X2+X3+X4+X5 ID
 1+01+0.05+0.35+0.30+0.15+0.06NEWDATA5
```

After several false starts, it became clear that the ions resulting from the cleavage of the peptide bond would serve as the best descriptors, and, of those, the most reliable were the C-terminal set. A literature review of FAB MS/MS and FAB/CAD MS/MS peptide spectra was carried out to provide as large a set of data of known fragmentation descriptors as possible.

Program LM, developed by Jurs[16] and adapted by Pulfer,[22] was further modified to be applicable to analyze mass spectral data of peptides. Separate files were created in order to consider different sizes of peptide fragments. Those having the unit under consideration were classified as positive and the intensities of the peaks entered at fixed masses characteristic of the sequence. The others were classified negative and their descriptors entered in the same manner. In this way, the training set was considered well-defined. Several peptides were checked to see if their residue sequences fit the training patterns, and, in almost every case, except for the occasional mono and dipeptide sequence data, correct classification was possible.

Program Description. The pattern recognition program is written in FORTRAN-77; see Chart I. It involves the following procedures.

The first consists of data input. The user has the freedom to define the test to suit his needs, i.e., the maximum number of data points to be input, the number of descriptors tested, the number of residues to be analyzed in the peptide unit (whether it is mono-, di-, tri-, tetra-, or pentapeptide), the convergence value before the training routine terminates, the width of the decision surface (as the threshold value), and other machine-directed variables which will allow the user to

manipulate the program in order to obtain a full characterization of the peptide of interest.

When entering the data, the intensities of the C-terminal peak sequence are input in an order corresponding to the masses of the sequence. In the case of leucine enkephalin, YGGFL, the first slot represents the mass of the 130-Da fragment, the second the mass of the 279 fragment, and so on. All relative intensities are entered correct to two decimal places, along with an arbitrary positive or negative value, indicating whether the data is thought to contain the peptide unit or not. Once the data entry is complete, the training routine begins. The data in the training set are obtained from separate data files, depending on the length of the peptide sequence being analyzed: the MONO file if only one residue is being identified, DI if two, and so on. A separate data file (shown in Figure 3) is required for each type of sequence being analyzed. This may, at first sight, appear cumbersome; however, if a negative result occurs for a given data file, then that has the tendency of ruling out large classes of peptides and narrows the search dramatically.

Training of the set continues until a weight vector is found which classifies all the data correctly or until it reaches the user-defined convergence value. Then the weight vector is applied to the prediction set, which includes the unknown data, and the set is classified.

The output resides in a user-named file for convenient future reference. It consists of the training routine, a record of the number of data sets that are incorrectly classified by successive weight vectors, the final weight vector, and the classification of the prediction set. The patterns will either be classified into the "yes" class, having the unit in its C-terminal sequence, or into the "no" class, which does not have the unit in the C-terminal sequence. The label arbitrarily set for the unknown will be checked and the result printed as "correct yes/no class" or "incorrect yes/no class" so that if the unknown pattern belonging to the "yes" class was inadvertently labeled "no", i.e., negative, then it will appear in the class "incorrect no". In summary, data belonging to the "correct yes" and "incorrect no" contain the unit of interest.

The data that fall into the threshold region of the weight vector cannot be correctly classified and, as such, are classed in the "not predicted" set.

Application of Pattern Recognition Program. The C-terminal sequence ions of leucine enkephalin were chosen as the test case. The mass values of the sequence ions were 130, 279, 336, 393, and 555 Da. In the special case of leucine, the mass values used were 86 and 130 Da, as employed by Kulik et al.[23]

As mentioned above, the data for the specific units are shown in Figure 3, where each file is tailored to match the unit under examination. The data identifiers are cross-referenced to their literature source, which is kept on file and will be supplied upon request. The data entries are tagged by number and membership in the two possible classes, followed by the five descriptors and the identifying label for ease of correlation to the literature and/or where the unit is located on the chain.

Typical screen output for the program is shown in Charts I and II. The data, 0.05, 0.35, 0.30, 0.15, and 0.06 named NEWDATA5, was entered with a positive label as the unknown peptide to be tested for the pentapeptide sequence L-F-G-G-Y. The output, as entered in the NEWFILE5 datafile, is illustrated in Chart II. Notice that NEWDATA5 is correctly classified "yes", indicating that this sequence does correlate to leucine enkephalin, YGGFL.

If the data entered were 0.90, 0.45, 0.87, 0.00, and 0.00, then this sequence would correctly predict the tripeptide sequence L-F-G. Mono- and dipeptide trials have also been generated; all of which are on file and accessible upon request.

DISCUSSION

The first example represents a set of data which was checked for the pentapeptide unit and the second for the tripeptide. In both cases, the program predicts that the unit being considered is, in fact, present because of its C-terminal sequence pattern.

Implicit in this analysis is the idea that the unit under consideration be present in the same order as that of the training set data, i.e., the unknown in the pentapeptide analysis could have the sequence X-Y-G-G-F-L, where X represents other amino acid residues, and the tripeptide analysis could have the sequence X-G-F-L.

The pattern recognition program is based on two rules: First, for a pattern to be positively classed, it must contain all sequence ions being searched; any other combination would class it in the "no" category. Second, if the unknown contains all the required peaks, no matter what their intensity, it would be correctly classed in the "yes" category. This is a valid assumption to make because longer peptide chains may give lower intensities than shorter ones, even though both may have the same amino acid residue sequence. These rules are embodied in the positive and negative classes of the training sets.

The pattern recognition program is based on a single assumption: that each amino acid residue has a unique mass. In most cases this is correct, but there are a few exceptions: leucine and isoleucine on one hand and lysine and glutamine on the other. It is possible to identify a sequence containing these isomers provided that there is one or more unique side-chain peaks present to resolve the ambiguity. In the case of leucine and isoleucine, there are two other ions labeled *d* and *w* which can do that.[20]

Within this structure, there are still many peptides that can be resolved by this method, and all can be resolved to within the two pairs stated above. The advantages clearly override the disadvantages of those few ions which require further analysis.

In some spectra, due to instrumental errors or methodology used, a few critical peaks may, from time to time, not be observed. This could hinder the pattern recognition program, especially if the unit being considered is not very long because a larger percentage of the pattern descriptors are absent. Experience has shown this to be the case, i.e., the technique has a higher degree of robustness for longer peptides because, in those instances, even if one or two of the peaks are absent, the remainder provide sufficient correlation to correctly classify the pattern. In the shorter units, only a limited number of sequence ions, and thus descriptors, are available, leading to a higher chance of misclassification.

This phenomena is clearly illustrated in the second example, where the tripeptide is being analyzed. One data set ought to have been classified "correct, no" but was predicted to belong to the "incorrect, no" category which means that it does have the correct tripeptide sequence. Usually in such cases, it is the small number of peaks being considered which causes the error. This trend becomes more pronounced when checking for the presence of mono- and dipeptide residues. In summary, pattern recognition is usually more applicable, i.e., has greater reliability of result in the sequence elucidation of longer peptides. This is a considerable advantage, as analysis of

```
-01+0.00+0.00+0.00+0.00+0.00OTHER-30
-01+0.00+0.00+0.00+0.00+0.00OTHER-30
+01+0.24+0.09+0.00+0.00+0.00DI-C-30
-01+0.00+0.00+0.00+0.00+0.00OTHER-30
+01+0.06+0.02+0.00+0.00+0.00DI-C-30
-01+0.00+0.00+0.00+0.00+0.00OTHER-30
-01+0.20+0.00+0.00+0.00+0.00OTHER-31
+01+1.00+0.00+0.00+0.00+0.00DI-N-30
+01+0.10+0.03+0.00+0.00+0.00DI-C-30
-01+0.29+0.00+0.00+0.00+0.00OTHER-31
-01+0.00+0.05+0.00+0.00+0.00OTHER-30
-01+0.04+0.00+0.00+0.00+0.00ISOLEU-N-31
+01+0.75+0.00+0.00+0.00+0.00DI-N-30
-01+0.31+0.00+0.00+0.00+0.00OTHER-4
+01+0.05+0.05+0.00+0.00+0.00 C-12
+01+0.45+0.00+0.00+0.00+0.00C-32
-01+0.00+0.54+0.00+0.00+0.00OTHER-32
-01+0.00+0.03+0.00+0.00+0.00OTHER-30
-01+0.20+0.00+0.00+0.00+0.00OTHER-33
-01+0.00+0.70+0.00+0.00+0.00OTHER-12
-01+0.00+0.06+0.00+0.00+0.00OTHER-31
```

(a) mono

```
-01+0.00+0.00+0.00+0.00+0.00OTHER-20
-01+0.00+0.00+0.00+0.00+0.00OTHER-34
+01+0.05+0.39+0.00+0.00+0.00PENTA-20
-01+0.00+0.28+0.00+0.00+0.00OTHER-20
+01+0.05+0.37+0.00+0.00+0.00TRI-C-12
-01+0.00+0.00+0.00+0.00+0.00OTHER-18
-01+0.39+0.00+0.00+0.00+0.00OTHER-20
-01+0.00+0.00+0.00+0.00+0.00OTHER-36
+01+0.09+0.22+0.00+0.00+0.00DI-C-30
-01+0.00+0.00+0.00+0.00+0.00OTHER-18
-01+0.06+0.00+0.00+0.00+0.00OTHER-31
-01+0.23+0.00+0.00+0.00+0.00OTHER-31
-01+0.30+0.00+0.00+0.00+0.00OTHER-31
-01+0.00+0.00+0.00+0.00+0.00OTHER-31
-01+0.54+0.00+0.00+0.00+0.00OTHER-32
-01+0.00+0.00+0.00+0.00+0.00OTHER-4
-01+0.00+0.00+0.00+0.00+0.00OTHER-35
-01+0.09+0.00+0.00+0.00+0.00OTHER-33
-01+0.00+0.05+0.00+0.00+0.00OTHER-26
-01+0.00+0.19+0.00+0.00+0.00OTHER-33
-01+0.70+0.00+0.00+0.00+0.00OTHER-12
-01+0.00+0.00+0.00+0.00+0.00OTHER-18
-01+0.00+0.05+0.00+0.00+0.00OTHER-12
```

(b) di

```
-01+0.00+0.00+0.09+0.00+0.00OTHER-20
-01+0.00+0.00+0.00+0.00+0.00OTHER-34
+01+0.05+0.39+0.29+0.00+0.00PENTA-20
-01+0.00+0.28+0.46+0.00+0.00OTHER-20
+01+0.05+0.37+0.39+0.00+0.00TRI-C-12
-01+0.00+0.00+0.00+0.00+0.00OTHER-18
-01+0.39+0.00+0.00+0.00+0.00OTHER-20
-01+0.00+0.00+0.00+0.00+0.00OTHER-36
+01+0.09+0.22+0.00+0.00+0.00DI-C-30
-01+0.00+0.00+0.00+0.00+0.00OTHER-18
-01+0.06+0.00+0.20+0.00+0.00OTHER-31
-01+0.23+0.00+0.00+0.00+0.00OTHER-31
-01+0.30+0.00+0.52+0.00+0.00OTHER-31
-01+0.00+0.00+0.00+0.00+0.00OTHER-31
-01+0.54+0.00+0.00+0.00+0.00OTHER-32
-01+0.00+0.00+0.00+0.00+0.00OTHER-4
-01+0.00+0.00+0.00+0.00+0.00OTHER-35
-01+0.09+0.00+0.00+0.00+0.00OTHER-33
-01+0.00+0.05+0.00+0.00+0.00OTHER-26
-01+0.00+0.19+0.12+0.00+0.00OTHER-33
-01+0.70+0.00+0.00+0.00+0.00OTHER-12
-01+0.00+0.00+0.00+0.00+0.00OTHER-18
-01+0.00+0.05+0.00+0.00+0.00OTHER-12
```

(c) tri

```
-01+0.00+0.00+0.09+0.20+0.00OTHER-20
-01+0.00+0.00+0.00+0.00+0.00OTHER-34
+01+0.05+0.39+0.29+0.15+0.00PENTA-20
-01+0.00+0.28+0.46+0.00+0.00OTHER-20
+01+0.05+0.37+0.39+0.00+0.00TRI-C-12
-01+0.00+0.00+0.00+0.00+0.00OTHER-18
-01+0.39+0.00+0.00+0.00+0.00OTHER-20
-01+0.00+0.00+0.00+0.00+0.00OTHER-36
+01+0.09+0.22+0.00+0.00+0.00DI-C-30
-01+0.00+0.00+0.00+0.00+0.00OTHER-18
-01+0.06+0.00+0.20+0.00+0.00OTHER-31
-01+0.23+0.00+0.00+0.00+0.00OTHER-31
-01+0.30+0.00+0.52+0.00+0.00OTHER-31
-01+0.00+0.00+0.00+0.00+0.00OTHER-31
-01+0.54+0.00+0.00+0.00+0.00OTHER-32
-01+0.00+0.00+0.00+0.00+0.00OTHER-4
-01+0.00+0.00+0.00+0.00+0.00OTHER-35
-01+0.09+0.00+0.00+0.00+0.00OTHER-33
-01+0.00+0.05+0.00+0.00+0.00OTHER-26
-01+0.00+0.19+0.12+0.00+0.00OTHER-33
-01+0.70+0.00+0.00+0.00+0.00OTHER-12
-01+0.00+0.00+0.00+0.21+0.00OTHER-18
-01+0.00+0.05+0.00+0.00+0.00OTHER-12
```

(d) tetra

```
-01+0.00+0.00+0.09+0.20+0.00OTHER-20
-01+0.00+0.00+0.00+0.00+0.00OTHER-34
+01+0.05+0.39+0.29+0.15+0.05PENTA-20
-01+0.00+0.28+0.46+0.00+0.00OTHER-20
+01+0.05+0.37+0.39+0.00+0.00TRI-C-12
-01+0.00+0.00+0.00+0.00+0.00OTHER-18
-01+0.39+0.00+0.00+0.00+0.00OTHER-20
-01+0.00+0.00+0.00+0.00+0.00OTHER-36
+01+0.09+0.22+0.00+0.00+0.00DI-C-30
-01+0.00+0.00+0.00+0.00+0.20OTHER-18
-01+0.06+0.00+0.20+0.00+0.00OTHER-31
-01+0.23+0.00+0.00+0.00+0.00OTHER-31
-01+0.30+0.00+0.52+0.00+0.00OTHER-31
-01+0.00+0.00+0.00+0.00+0.00OTHER-31
-01+0.54+0.00+0.00+0.00+0.00OTHER-32
-01+0.00+0.00+0.00+0.00+0.38OTHER-4
-01+0.00+0.00+0.00+0.00+0.12OTHER-35
-01+0.09+0.00+0.00+0.00+0.00OTHER-33
-01+0.00+0.05+0.00+0.00+0.00OTHER-26
-01+0.00+0.19+0.12+0.00+0.00OTHER-33
-01+0.70+0.00+0.00+0.00+0.00OTHER-12
-01+0.00+0.00+0.00+0.21+0.00OTHER-18
-01+0.00+0.05+0.00+0.00+0.00OTHER-12
```

(e) penta

Figure 3. Data for training and predicting fragment sequences: (a) mono, L-; (b) di, L-F-; (c) tri, L-F-G-; (d) tetra, L-F-G-G-; (e) penta, L-F-G-G-Y.

Chart II. NEWFILES output from program LM for the unknown peptide NEWDATA5 being checked for L-F-G-G-Y

```
   FILE BEING EXAMINED :PENTA

INITIALIZE RANDOM NUMBER GENERATOR=
          345

NUMBER OF TRAINING SETS, NTRSET:

           13

CHOSEN NO OF PATTNS FOR PREDICTION SET

           10

THRESHOLD VALUE IS:

          0.100

THE NUMBER OF DESCRIPTORS:

            5

         THE DESCRIPTORS USED:

         JJ1,JJ2,JJ3 ARE:

1  5  1
   PTTN ID # =   1   BELONGS TO: NEWDATA5
 1  0.05 0.35 0.30 0.15 0.06
   PTTN ID # =   2   BELONGS TO: OTHER-20
 2  0.00 0.00 0.09 0.20 0.00
   PTTN ID # =   3   BELONGS TO: OTHER-34
 3  0.00 0.00 0.00 0.00 0.00
   PTTN ID # =   4   BELONGS TO: PENTA-20
 4  0.05 0.39 0.29 0.15 0.05
   PTTN ID # =   5   BELONGS TO: OTHER-20
 5  0.00 0.28 0.46 0.00 0.00
   PTTN ID # =   6   BELONGS TO: TRI-C-12
 6  0.05 0.37 0.39 0.00 0.00
   PTTN ID # =   7   BELONGS TO: OTHER-18
 7  0.00 0.00 0.00 0.00 0.00
   PTTN ID # =   8   BELONGS TO: OTHER-20
 8  0.39 0.00 0.00 0.00 0.00
   PTTN ID # =   9   BELONGS TO: OTHER-36
 9  0.00 0.00 0.00 0.00 0.00
   PTTN ID # =  10   BELONGS TO: DI-C-30
10  0.09 0.22 0.00 0.00 0.00
   PTTN ID # =  11   BELONGS TO: OTHER-18
11  0.00 0.00 0.00 0.00 0.20
   PTTN ID # =  12   BELONGS TO: OTHER-31
12  0.06 0.00 0.20 0.00 0.00
   PTTN ID # =  13   BELONGS TO: OTHER-31
13  0.23 0.00 0.00 0.00 0.00
   PTTN ID # =  14   BELONGS TO: OTHER-31
14  0.30 0.00 0.52 0.00 0.00
   PTTN ID # =  15   BELONGS TO: OTHER-31
15  0.00 0.00 0.00 0.00 0.00
   PTTN ID # =  16   BELONGS TO: OTHER-32
16  0.54 0.00 0.00 0.00 0.00
   PTTN ID # =  17   BELONGS TO: OTHER-4
17  0.00 0.00 0.00 0.00 0.38
   PTTN ID # =  18   BELONGS TO: OTHER-35
18  0.00 0.00 0.00 0.00 0.12
   PTTN ID # =  19   BELONGS TO: OTHER-33
19  0.09 0.00 0.00 0.00 0.00
   PTTN ID # =  20   BELONGS TO: OTHER-26
20  0.00 0.05 0.00 0.00 0.00
   PTTN ID # =  21   BELONGS TO: OTHER-33
21  0.00 0.19 0.12 0.00 0.00
   PTTN ID # =  22   BELONGS TO: OTHER-12
22  0.70 0.00 0.00 0.00 0.00
   PTTN ID # =  23   BELONGS TO: OTHER-18
23  0.00 0.00 0.00 0.21 0.00
ITIAL WEIGHT VECTOR

1  0.100
2  0.100
3  0.100
4  0.100
5  0.100
6  0.200

CHANGE ANY INITIAL WEIGHT VECTOR VARIABLE

 7  0.000

        TRAINING ROUTINE
 7  4  2  2  2  2  1  0  1  0  0

        WEIGHT VECTOR
          0.032
          1.445
         -0.020
          0.044
          0.124
         -0.185

        FEEDBACKS    21


 EDICTION WITH THRESHOLD =    0.10

      9 NUMBER PREDICTED

      1 NUMBER NOT PREDICTED

      1 NUMBER PREDICTED INCORRECTLY

        OVERALL            NO CLASS            YES CLASS
      1/  9  88.89       1/  8  87.50       0/  1 100.00

       AN IDENTIFYING SUMMARY FOLLOWS:

       THE PATTERNS NOT PREDICTED ARE:

OTHER-33      0.00 0.19 0.12 0.00 0.00
       END FILE

       THE NUMBER OF CORRECT YES CLASSIFICATIONS

NEWDATA5      0.05 0.35 0.30 0.15 0.06
       END FILE

       THE NUMBER OF INCORRECT YES CLASSIFICATIONS

       END FILE

       THE NUMBER OF CORRECT NO CLASSIFICATIONS

OTHER-20      0.00 0.00 0.09 0.20 0.00
OTHER-36      0.00 0.00 0.00 0.00 0.00
OTHER-31      0.23 0.00 0.00 0.00 0.00
OTHER-31      0.30 0.00 0.52 0.00 0.00
OTHER-31      0.00 0.00 0.00 0.00 0.00
OTHER-4       0.00 0.00 0.00 0.00 0.38
OTHER-35      0.00 0.00 0.00 0.00 0.12
       END FILE

       THE NUMBER OF INCORRECT NO CLASSIFICATIONS

 HER-20       0.00 0.28 0.46 0.00 0.00
       END FILE

 THE PATTERN NUMBERS IN NTRSET ARE:

 23  20  19   6   3   4  22   7   8  10
 12  16  11   0   0   0   0   0   0   0


 THE PRSET NUMBERS AND DATA FOLLOW:

NEWDATA5      0.05 0.35 0.30 0.15 0.06
OTHER-20      0.00 0.00 0.09 0.20 0.00
OTHER-20      0.00 0.28 0.46 0.00 0.00
OTHER-36      0.00 0.00 0.00 0.00 0.00
OTHER-31      0.23 0.00 0.00 0.00 0.00
OTHER-31      0.30 0.00 0.52 0.00 0.00
OTHER-31      0.00 0.00 0.00 0.00 0.00
OTHER-4       0.00 0.00 0.00 0.00 0.38
OTHER-35      0.00 0.00 0.00 0.00 0.12
OTHER-33      0.00 0.19 0.12 0.00 0.00
```

long, complicated peptides is an area not treated with ease by other techniques.

Another attractive feature of pattern recognition is that in many instances the entire spectrum need not be analyzed at once. An analysis of the peaks over certain ranges will indicate the presence or otherwise of a particular sequence. Then, by extending the range in a controlled manner, the way the sequence builds up can be carefully developed by the LM technique.

The technique can be applied to any peptide of interest, not just to those which follow certain predefined rules arising out of their biological evolution. What is required is the development of a data file defining the positive and negative classes for the sequence of interest. Thus, unusual or synthetic amino acid residues can be analyzed.

One aspect, which did complicate descriptor selection dealt with the problem of comparing ion intensities from different mass spectra. Intensity is a function of several experimental parameters including the pressure of the collisional gas,[24] dimension of the collisional cell,[21] atomic mass of the colliding particle, and translational energy of the fast atoms and precursor ions.[20] Therefore, the spectra in the training sets were obtained with as much similarity as possible. However, not a great deal of time was spent on this as, has been stated before, the presence of the peak intensity at specific mass positions is of more importance than its absolute magnitude.

The analysis requires the C-terminal sequence to have a free carboxyl radical in each case, whether it be spectra used to generate the classification set or for an unknown. Then, if a data set is classified "yes" then its C-terminal nature is automatically known. If, however, it is classified "no", then clearly the terminal residue must be something else other than a carboxyl radical.

Further work in this type of analysis could extend not only to larger sequences but to progressively increasing the range analyzed and, from that, build up a picture of the correct

sequence. Also, no work has been done on N-terminal sequences. This could serve as a ready backup check, and possibly discount any errors generated by the absence of some peaks in the C-terminal series.

Last, data banks need to be prepared for the most likely sequences found in biologically active molecules. These would be used to confirm their presence by mass spectrometric analysis.

ACKNOWLEDGMENT

R.D. would like to thank her family for enduring encouragement.

REFERENCES AND NOTES

(1) Barber, M.; Bordoli, R. S.; Sedgwick, R. D.; Tyler, A. N. Fast Atom Bombardment of Solids as an Ion Source in Mass Spectrometry. *Nature* **1981**, *293*, 270–275.
(2) Barber, M.; Bordoli, R. S.; Elliot, G. J.; Sedgwick, R. D.; Tyler, A. N. Fast Atom Bombardment Mass Spectrometry. *Anal. Chem.* **1982**, *54*, 645A–657A.
(3) Tomer, K. B. The Development of Fast Atom Bombardment Combined with Tandem Mass Spectrometry for the Determination of Biomolecules. *Mass Spectrom. Rev.* **1989**, *8*, 445–482.
(4) Biemann, K.; Martin, S. A.; Scoble, H. A.; Johnson, R. S.; Papayannopoulos, I. A.; Biller, J. E.; Costello, C. E. In *Mass Spectrometry in the Analysis of Large Molecules*; McNeal, C. J., Ed.; Wiley: New York, 1986; p 131.
(5) Hunt, D. F.; Yates, J. R., III; Shabanowitz, J.; Winston, S.; Hauer, C. R. Protein Sequencing by Tandem Mass Spectrometry. *Proc. Natl. Acad. Sci. U.S.A.* **1986**, *83*, 6233–6237.
(6) Covey, T. R.; Huang, E. C.; Henion, J. D. Structural Characterization of Protein Tryptic Peptides via Liquid Chromatography/Mass Spectrometry and Collision-Induced Dissociation of Their Doubly Charged Molecular Ions. *Anal. Chem.* **1991**, *63*, 1193–1200.
(7) Sakurai, T.; Matsuo, T.; Matsuda, H.; Katakuse, I. PAAS 3: A Computer Program To Determine Probable Sequence of Peptides from Mass Spectrometric Data. *Biomed. Mass Spectrom.* **1984**, *11*, 396–399.
(8) Ishikawa, K.; Niwa, Y. Computer-Aided Peptide Sequencing by Fast Atom Bombardment Mass Spectrometry. *Biomed. Environ. Mass Spectrom.* **1986**, *13*, 373–380.
(9) Siegel, M. M.; Bauman, N. An Efficient Algorithm for Sequencing Peptides Using Fast Atom Bombardment Mass Spectral Data. *Biomed. Environ. Mass Spectrom.* **1988**, *15*, 333–343.
(10) Bartels, C. Fast Algorithm for Peptide Sequencing by Mass Spectroscopy. *Biomed. Environ. Mass Spectrom.* **1990**, *19*, 363–368.
(11) Lee, T. D.; Vemuri, S. MacProMass: A Computer Program to Correlate Mass Spectral Data to Peptide and Protein Structures. *Biomed. Environ. Mass Spectrom.* **1990**, *19*, 639–645.
(12) Papayannopoulos, I. A.; Biemann, K. A Computer Program (COMPOST) for Predicting Mass Spectrometric Information from Known Amino Acid Sequences. *Am. Chem. Soc. Mass Spectrom.* **1991**, *2*, 174–177.
(13) Carson, S. D.; Baggenstoss, B. Identification of Peptides within a Known Protein Sequence Using COMSEQ Analysis of Data Containing Multiple Sequences. *Comput. Methods Programs Biomed.* **1991**, *35*, 35–42.
(14) Ziemer, J. N.; Perone, S. P.; Caprioli, R. M.; Seifert, W. E. Computerized Pattern Recognition Applied to Gas Chromatography/Mass Spectrometry Identification of Pentafluoropropionyl Dipeptide Methyl Esters. *Anal. Chem.* **1979**, *51*, 1732–1738.
(15) Kryger, L. Interpretation of Analytical Chemical Information by Pattern Recognition Methods—A Survey. *Talanta* **1981**, *28*, 871–887.
(16) Jurs, P. C. In *Computer Software Applications in Chemistry*; Wiley: New York, 1986; p 186.
(17) Forina, M.; Leardi, R. PARVUS—MS-DOS: A Package for General Pattern Recognition. *Tr. Anal. Chem.* **1988**, *7*, 53–54.
(18) Biemann, K. In *Some Recent Applications of Mass Spectrometry to Biochemistry*; Oliver, R. W. A., Thompson, J. S., Eds.; Techniques Group Colloquium: Manchester, 1988; Vol. 17, p 237.
(19) Roepstorff, P.; Fohlman, J. Proposal for a Common Nomenclature for Sequence Ions in Mass Spectra of Peptides. *Biomed. Mass Spectrom.* **1984**, *11*, 601.
(20) Martin, S. A.; Johnson, R. S.; Costello, C. E.; Bieman, K. In *The Analysis of Peptides and Proteins by Mass Spectrometry*; McNeil, C. J., Ed.; Wiley: New York, 1988; p 135.
(21) Desiderio, D.; Katakuse, I. Fast Atom Bombardment Linked-Field Scanning Mass Spectrometric Analysis of Peptides. *Mass Spectrom.* **1985**, *33*, 351–370.
(22) Pulfer, J. D. Pattern Recognition Analysis of Lead Isotope Ratios as an Aid to Gold Prospecting. Unpublished, 1988.
(23) Kulik, W.; Heerma, W. The Determination of the Amino Acid Sequence in the Fast Atom Bombardment Mass Spectra of Dipeptides. *Biomed. Environ. Mass Spectrom.* **1988**, *17*, 173–180.
(24) Bradley, C. D.; Derrick, P. J. Collision-Induced Decomposition of Peptides. An Investigation into the Effect of Collision Gas Pressure on Translational Energy Losses. *Org. Mass Spectrom.* **1991**, *26*, 395–401.