Application of Computers in Chemical Analysis: Amino-acid Analysis and Sequence Determination

By B. Sheldrick
ASTBURY DEPARTMENT OF BIOPHYSICS, THE UNIVERSITY OF LEEDS,
LEEDS 2

The widespread introduction of electronic digital computers in recent years has enabled chemists to perform two types of calculation which were previously too complicated, *i.e.*, those which involve complicated mathematical treatment of data and those which involve large numbers of comparisons or sorting movements as in literature searches. The second type of facility is useful when dealing with sequence analysis of chemical compounds, which includes the sequence analysis of proteins or polypeptides, when these are not analysed step-by-step but are investigated by examination of the fragments produced by random hydrolysis, and the establishment of the sequence in a compound by examination of its mass spectrum. Both these types of sequence analysis have much in common as they involve a considerable amount of sorting and comparison but without any complicated mathematical requirements. The advantage of the electronic computer in this field lies in its ability to sort and compare in a thorough and systematic manner the large number of combinations that can occur, and which increases very rapidly with increase of the molecular weight of the compound under examination. For convenience we shall also consider the mass spectrometry of mixtures.

1 Automatic Amino-acid Analysis

The introduction of automation to the analysis of proteins and polypeptides has progressed in a number of discrete steps. First was the introduction of the chromatography column with automatic sampling of the eluent. The sampling was then automated further so that the addition of ninhydrin reagent and the spectrophotometry of the samples resulted in the output of the results as a graphical display on a recorder chart, the chart showing a plot of absorbance against time (or consecutive samples). From this point on, the digital computer entered the picture in a variety of ways and some of the possibilities are shown in Figure 1.

In the left-hand column is shown the method adopted without digital computation; the peaks corresponding to the various amino-acids on the chart are measured, usually by measuring the height of the peak and the number of recorder points marked above a line drawn at the half height of the peak (the accuracy of these measurements is the limiting factor of the accuracy of the final results). The quantities of the amino-acids (c) are then calculated from the relation:

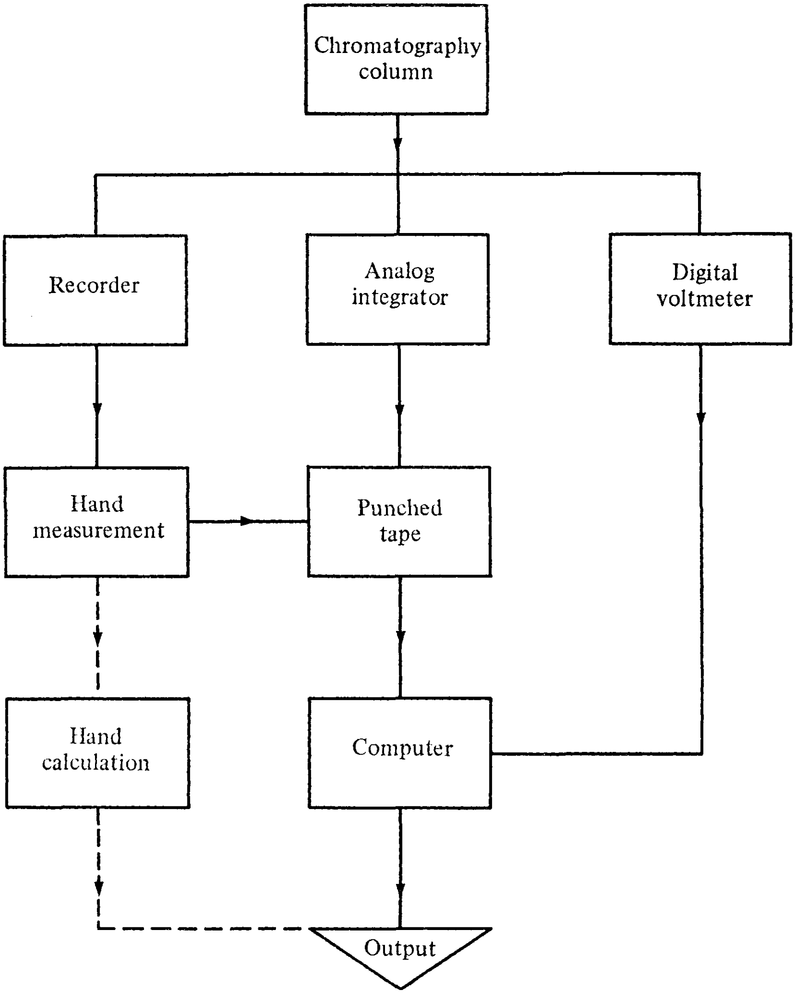

Figure 1

c = no. of dots × peak height × calibration constant, where the calibration constant is given by:

$$\text{const} = \frac{\text{norleucine equivalent} \times \text{norleucine concentration}}{\text{area of norleucine peak}}$$

The value of the calibration constant is found from a run which contains an internal standard (norleucine).

The simplest method of applying a computer to this procedure is merely to carry out the calculations on the data measured by hand. An ALGOL program put forward by Graham and Sheldrick[1] (plus correction) requires the measured data to be typed by hand into a remote terminal,* which is then used by the program, permanently stored in the computer, to produce values, for each amino-acid of μmoles, residues/1000 residues, μg nitrogen, and percentage total nitrogen plus a value for the total nitrogen, which are output at the terminal. Not all the final sets of values are required but, since results are sometimes quoted in the literature as μg nitrogen, sometimes as residues/1000 *etc.*, a small amount of additional computation produces the complete set of data which is suitable for direct comparison with other published results no matter what system is used. This procedure is reasonable as most computers will carry out calculations much faster than data are input, *i.e.*, once the data are read into the computer it is advisable to carry out all the simple calculations which may be required rather than have another program which handles the same set of data.

The program just mentioned suffers the disadvantage that human error is not eliminated from the data handling, as errors can, and do, occur in measuring the peaks and typing the data. Alternative methods have been developed to overcome this which involve direct production of a paper tape (or magnetic tape) by the chromatography equipment. Two methods are available, one in which the peaks produced are integrated, by an analog integrator, to produce a peak area, which is then punched automatically on to paper tape; a second method uses a digital voltmeter to punch each reading on to a paper tape. Obviously, different programs are required to compute the results as the second tape will not have the individual peak areas but will require procedures which recognise a peak and integrate it before carrying out the final calculations.

An example of the first type of program has been published by Starbuck *et al.*[2] This program, written in FORTRAN, was designed to accept information provided either by the hand calculation method or by an integrator, either method giving the required information of the peak area. Additional refinements in this program allow previous standard runs to be averaged in with the run being calculated if required, correction of the results for specified loss or destruction of particular amino-acids, allowance for ammonia production, and calculation of the number of amino-acid residues in one molecule of the protein (the molecular weight of the protein is supplied as part of the input data). An additional program by the same authors calculates the empirical composition of peptides obtained by the proteolytic digestion of proteins.

The completely digitised system may be considered in two parts, the automation of the analyser to produce a digital output and program techniques required to handle the data in this form. Discussions of various forms of digitiser have been given by Yonda *et al*,[3] Porter and Talley,[4] Jones and Spence,[5] Krichevsky,

* Remote terminal: a device for input and output of data connected to a computer by a direct link.

[1] G. N. Graham and B. Sheldrick, *Biochem. J.*, 1965, **96,** 517.

[2] W. C. Starbuck, C. M. Mauritzen, C. McClimans, and H. Busch, *Analyt. Biochem.*, 1967, **20,** 439.

[3] A. Yonda, D. L. Filmer, H. Pate, N. Alonzo, and C. H. W. Hirs, *Analyt. Biochem.*, 1965, **10,** 53.

[4] W. L. Porter and E. A. Talley, *Analyt. Chem.*, 1964, **36,** 1692.

[5] H. J. Jones and D. W. Spence, Infotronics Application Notes, No. 1, 1964.

Schwartz, and Mage,[6] and Cavins and Friedman.[7] These vary according to the type of computer with which they are to be used but basically consist of a coding system which converts the voltage produced by the spectrophotometer into a set of digital values on a paper tape or magnetic tape with suitable arrangements to indicate the termination of a value. The consecutive values are assumed to be produced at equal intervals of time and hence need not be indexed, as a simple count from the first value is sufficient identification. Since every measurement is recorded it is unnecessary to assume, as is done in the height-width calculation, that the peaks approximate to a Gaussian distribution, and a peak area is obtained by summation of all the observed values within it. Usually a recorder trace is produced simultaneously and filed, as this is much more meaningful to observe than a series of numbers if the output is queried. Using a string of digital numbers in place of a string of peak areas (or equivalent data) introduces difficulties which the computer program must be designed to overcome. These are discussed by Jones and Spence[5] for an on-line† integrator and some of the items which have to be dealt with are: (i) recognition or detection of the start and end of a peak; (ii) location of peak maximum; (iii) integration of peak area; (iv) variation of base-line value; (v) resolution of overlapping peaks.

Each item can be dealt with separately by an appropriate sub-routine but in a complete program there must necessarily be considerable overlapping of these sub-routines and care must be taken to ensure that, *e.g.*, two peaks which overlap are not mistaken by the program for two peaks with a peculiar base-line shift.

Consider first of all detection of the three required points of a single peak, the beginning, the maximum, and the end. If we store say, six consecutive numbers of the output string we can readily test these for characteristics which we define for the specific points, *e.g.*, for the beginning of a peak we may specify that

$$x_{t+1} > x_t$$

i.e., that each successive value is greater than the previous value. Alternatively we may specify that

$$x_{t+2} - x_{t+1} > x_{t+1} - x_t$$

i.e., that the differences increase.

Such a test is answered only by yes or no. If yes, then we have the beginning of a peak, if no, then each of the six values in store is shifted one to the left and another value read in from the data string to fill the empty sixth place; and the test repeated. This is equivalent to sliding a window along the string of numbers and checking at each position for the beginning of a peak. Variations can be introduced for the window size and the tests applied, while a similar system with different tests may be used to find the end of the peak. During the search for the end of the peak the intermediate values are stored and, when the process is complete, these are then scanned for the maximum value, *i.e.*, the centre of the peak which can be used for identification of the position of the peak in time.

† On-line: directly connected.

[6] M. I. Krichevsky, J. Schwartz, and M. Mage, *Analyt. Biochem.*, 1965, **12**, 94.

[7] J. F. Cavins and M. Friedman, *Cereal Chem.*, 1968, **45**, 172.

Before the peak area is calculated, however, the two other tests, for base-line drift and for presence of overlapping peaks have to be carried out. A simple method of distinguishing between these two cases is to examine the next few values after the end of the peak; if the values are higher than at the beginning of the peak, then they are tested to see whether or not they fit the criteria used to test for the beginning of a peak. If they do not, then there is a base-line drift which must be allowed for in the calculation of the peak area and, if they do, then another sub-routine must be entered to decide how the areas of the two peaks are to be established. Another simple test is to check if the peak maximum lies about half-way between the beginning and end of the peak; if not, then either the peak is of peculiar shape or overlapping is taking place.

In amino-acid analysis it is unlikely that overlapping peaks will occur as the conditions of the analysis are usually adjusted to ensure that full resolution of all peaks over the range is obtained. The calculation of the buffer pH values in the varigrad* has been described by Burns, Curtis, and Kacser.[8] Exceptions can, however, occur and various methods of resolution are available, depending on how the overlapping occurs.

The simplest form is when two symmetrical peaks of equal height overlap. Areas of the peaks can be calculated by drawing a vertical line from the lowest point between the maxima to the horizontal axis. The area defined by the horizontal axis, the curve from the beginning of the peak to the central minimum, and the vertical, is the area of the first peak, since just as much of the second peak is enclosed in this area as is cut off the first peak by the vertical line. Obviously, if the peaks differ in height or are not symmetrical but skewed, then this method becomes less accurate. A discussion of the resolution of unequal peaks has been given by Fraser and Suzuki[9] where the peaks are analysed into components by using either a Fourier method or by Cauchy Functions.† Though in practice the peaks obtained are not accurately symmetrical, the accuracy lost by assuming that they are will be quite small. An iterative procedure for decomposing the two peaks consists of calculating the theoretical constants for one peak, subtracting the calculated values from the observed values and then calculating the theoretical constants for the second peak. The values obtained for the second peak are then subtracted and the first peak corrected. This process is then terminated when the corrections involved fall below defined limits.

Finally the block diagram for a computer program which carries out the above tests, checks, and calculations, is shown in Figure 2. The only portion not already discussed is the test for the end of the data. This requires either a special number at the end of the data which is recognised by the program or a count of the number of data entered at the beginning of the data.

* Varigrad: a multichamber device used to produce a controlled variation of pH and/or salt concentration.

† Cauchy Function: a function of the type: $y = a/\{1 + [2(x - b)/c]^2\}$.

[8] J. A. Burns, C. F. Curtis, and H. Kacser, *J. Chromatog.*, 1965, **20**, 310.

[9] R. D. B. Fraser and E. Suzuki, *Analyt. Chem.*, 1966, **38**, 1770.

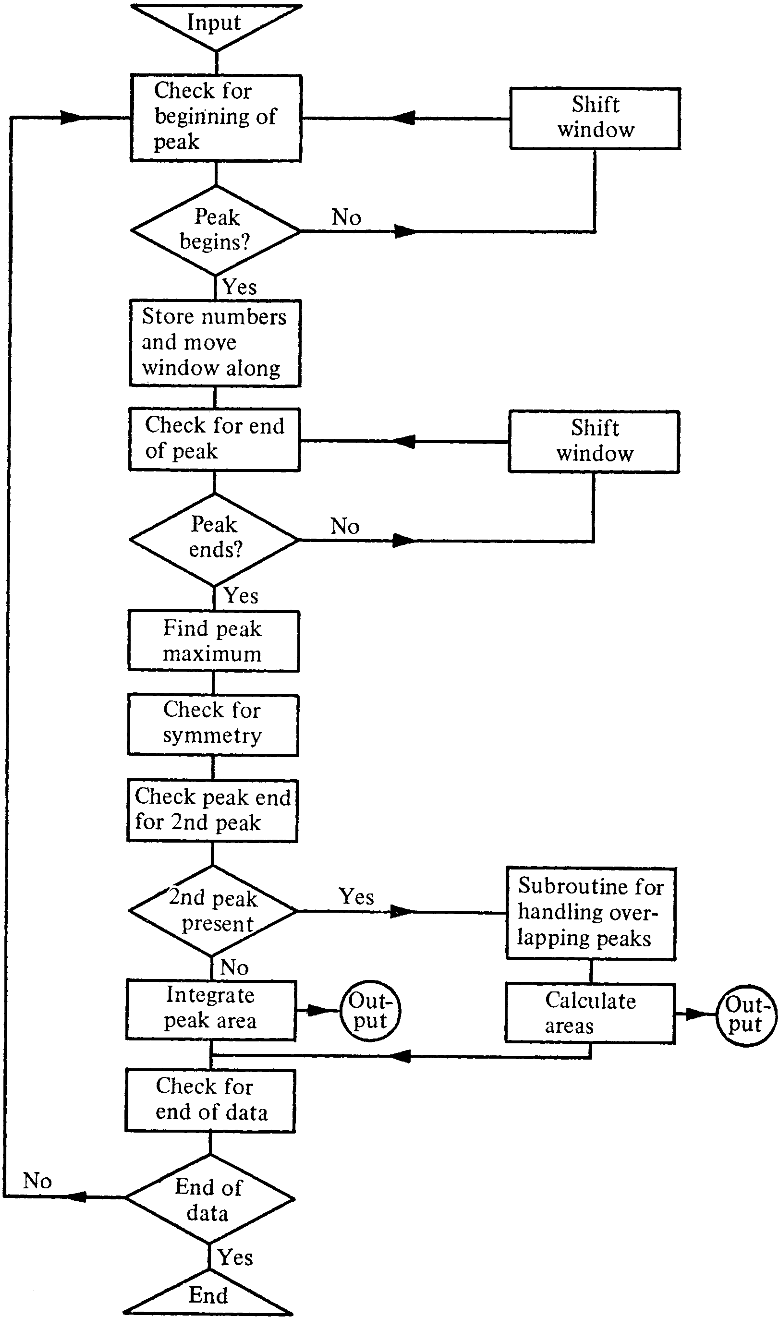
Input
Check for beginning of peak
Shift window
Peak begins?
No
Yes
Store numbers and move window along
Check for end of peak
Shift window
Peak ends?
No
Yes
Find peak maximum
Check for symmetry
Check peak end for 2nd peak
2nd peak present
Yes
Subroutine for handling over-lapping peaks
No
Integrate peak area
Out-put
Calculate areas
Out-put
Check for end of data
No
End of data
Yes
End

Figure 2

The former is more reliable as it will always be the same number and if omitted by an operator will not invalidate the calculations.

In the paper by Krichevsky, Schwartz, and Mage[6] details of a program similar to the above are discussed. An additional refinement introduced is a check and correction for 'noise' (*i.e.*, spurious values caused by random fluctuations in the electronic section of the apparatus), with special precautions taken when testing points near the maximum of a peak. One difficulty arises if the values are produced on paper tape and if this has then to be transferred to card format before the data can be read by the computer. It would perhaps help to avoid difficulties of this nature if all digital information could be recorded on magnetic tape to avoid the necessity of format changes. One factor which will have to be decided if this is ever to become a standard system is whether the data are recorded on the magnetic tape as an analog signal which is later converted to a digital signal by the computer, or whether the data are initially recorded in a digital form on the magnetic tape. The latter has the advantage that an additional translation at the computer is avoided but requires additional equipment in the form of an A–D convertor, to be on-line with the chemical equipment.

Chemical methods of establishing the amino-acid sequence of a protein have been carried out, in many cases on small quantities of material, with considerable success. Due to the difficulties of the chemical processes and the time involved, attention has been directed towards alternative methods. One technique consists of hydrolysing the protein into fragments, analysing the fragments and then calculating the unique chain which could produce the observed fragments. The process is not as fundamental as stepwise analysis and various factors have to be considered.

1. The fragments produced by the hydrolysis must be of reasonable length to allow information about the chain to be established, *e.g.*, total hydrolysis into single peptides would provide no sequential information at all.
2. The points on the chain at which breaks occur should be random and not systematic, *e.g.*, if the chain A.B.C., always breaks into A, B, and C, even if the separate parts undergo random hydrolysis, it would be impossible to distinguish between the chains A.B.C. and A.C.B. or B.A.C.

Bernhard, Bradley, and Duda[10] suggested a program which applies a set of rules in stages to the input data to produce the maximum amount of sequence information which can be established. They assumed that, for each fragment produced by random splitting of the chain, the total composition is known and also the identity of the *N*-terminal amino-acid, *e.g.*, a typical fragment could be represented as —Thr (Ala, Pro, Lys) where the tetrapeptide contains the specified four amino-acids; in this example it is known that threonine is the *N*-terminal group but the sequence of the other three is unknown.

An alternative method of input codes the amino-acids as prime numbers (Sheldrick[11]) with a separation between the items of known sequence, *e.g.*, in

[10] S. A. Bernhard, D. F. Bradley, and W. L. Duda, *IBM Journal*, 1963 246.

[11] B. Sheldrick, *Biochem. J.*, 1966, **100**, 11c.

the following example, if the coding is: Pro, 7; Gly, 2; His, 53; Ala, 3; Val, 11, then the information Pro (Gly, Ala, Val, His) would be represented by 7;3498 where the prime number 7 represents the *N*-terminal amino-acid, proline, and the integer 3498 is the product of the prime numbers, 2 × 3 × 11 × 53. This system has two advantages; all the necessary information is stored in one location limited only by the largest integer which can be stored in the computer, and a test of the data for the presence of a specific amino-acid can be carried out by a single operation, *i.e.*, the product is divided by the appropriate prime number and the answer tested for the presence of any remainder; no remainder indicates that the amino-acid is present. Incomplete sequence data are represented by the two systems as: Pro (Gly, Val, His) Ala and 7; 1166; 3. Similarly, Pro, Val, Gly (His, Ala) = 7; 11; 2; 159.

The process of elucidating the sequence of the amino-acid residues in the original protein from such data has been discussed in various papers. Bernhard, Bradley, and Duda[10] gave some preliminary results using artificially simulated fragments from a known sequence of insulin and showed how the process could be carried out in four stages:

1. The fragments are sorted into groups, each of which has a common known terminal acid.

2. Each group is broken down into sub-groups, the number of such sub-groups being limited to the number of times the *N*-terminal acid occurs in the original chain. Each sub-group contains fragments which have at least one amino-acid in common in addition to the *N*-terminal group and within each sub-group comparisons are used to contrast the information to produce a single sequence which includes all the information available in that sub-group, *e.g.*, two fragments 3;35 and 3;70 would be combined to give 3;35;2.

3. All the contracted fragments are re-listed and replace the original fragments.

4. The sorting process used in 2 is now repeated to merge the contracted fragments into one sequence. The example given in this paper provided the sequence of a chain of twelve units from a collection of eighteen fragments, some of the fragments providing superfluous information. The problem of calculating the minimum number of fragments which is necessary and sufficient to provide an unambiguous sequence for the original chain has not been solved for the general case. It is affected by the following factors:

(i) hydrolysis may not be truly random;

(ii) the number of fragments required may vary according to the arrangement of the amino-acid residues in the original chain—an example given in the above paper deals with the difficulty of determining the sequence 5;7;5;7;5;7;

(iii) variation of the size of the fragments will have an effect—*e.g.*, if all the fragments are only two residues long each occurrence of a duplicate of one of these pairs, in the original chain, will produce an ambiguity;

(iv) errors in the amino-acid analysis or data production may add to the difficulty of producing a sequence.

Further papers dealing with this topic have been published by Dayhoff[12] and Bradley, Merril, and Shapiro.[13] Dayhoff allows for the presence of erroneous data by specifying that removal of a suspected error will remove two or more inconsistencies and not introduce others. Some differences from the previous program of Bernhard, Bradley, and Duda occur because Dayhoff uses data for which the *C*-terminal residues are identified in addition to the *N*-terminal residues. The first process is to collect together groups which have a specified amino-acid residue in common; these are then sub-divided into sub-groups each of which can be condensed to produce a merged sequence. Fragments which cannot be assigned unambiguously to one sub-group are set on one side and this collection of fragments may contain some erroneous pieces in addition to any ambiguous sections. Merging of the sub-group sequences is then used to produce the final sequence (or sequences).

It is pointed out that one amino-acid may be difficult to detect and/or subject to error. Such an acid may be omitted from the sequence determination and added at suitable points, specified by the fragments which contain it, in the final chain. The process of determining the sequence without using a specified amino-acid is also useful to show how much importance can be assigned to that acid, *i.e.*, omission of one acid may have no effect on the final result whereas omission of another acid may produce several alternative possibilities.

The paper by Bradley, Merril, and Shapiro, being the first part of a series, gives a thorough coverage of the problems involved and the rules put forward, ten in number, are designed to give accurate indexing of the various fragments (six rules) with a separate rule to stop the process if internal inconsistencies occur, followed by merging of the fragments and reduction to the final sequence with a final rule to eliminate alternative sequences. The program is iterative, as shown in the block diagram Figure 3, and cycles until either no further overlaps are found or some discrepancy is found in which case a diagnostic message is output. Using this program the authors pointed out a discrepancy between the published composition[14] and the published sequence[15] of the B-chain of insulin and were also able to show that when the presence of three valine residues in the molecules is assumed and not two only, as in the composition data, the discrepancy is removed and the final number of possible sequences produced was six. When fragments produced by acid hydrolysis were added to the input data the sequence, identical to the published sequence, was obtained immediately.

Several textbooks and reviews are extant which deal with the mass spectrometer and its application to the problems arising in organic chemistry (refs.

[12] M. O. Dayhoff, *J. Theor. Biol.*, 1964, **8**, 97.

[13] D. F. Bradley, C. R. Merril, and M. B. Shapiro, *Biopolymers*, 1964, **2**, 415.

[14] F. Sanger and H. Tuppy, *Biochem. J.*, 1951, **49**, 463.

[15] F. Sanger and H. Tuppy, *Biochem. J.*, 1951, **49**, 481.

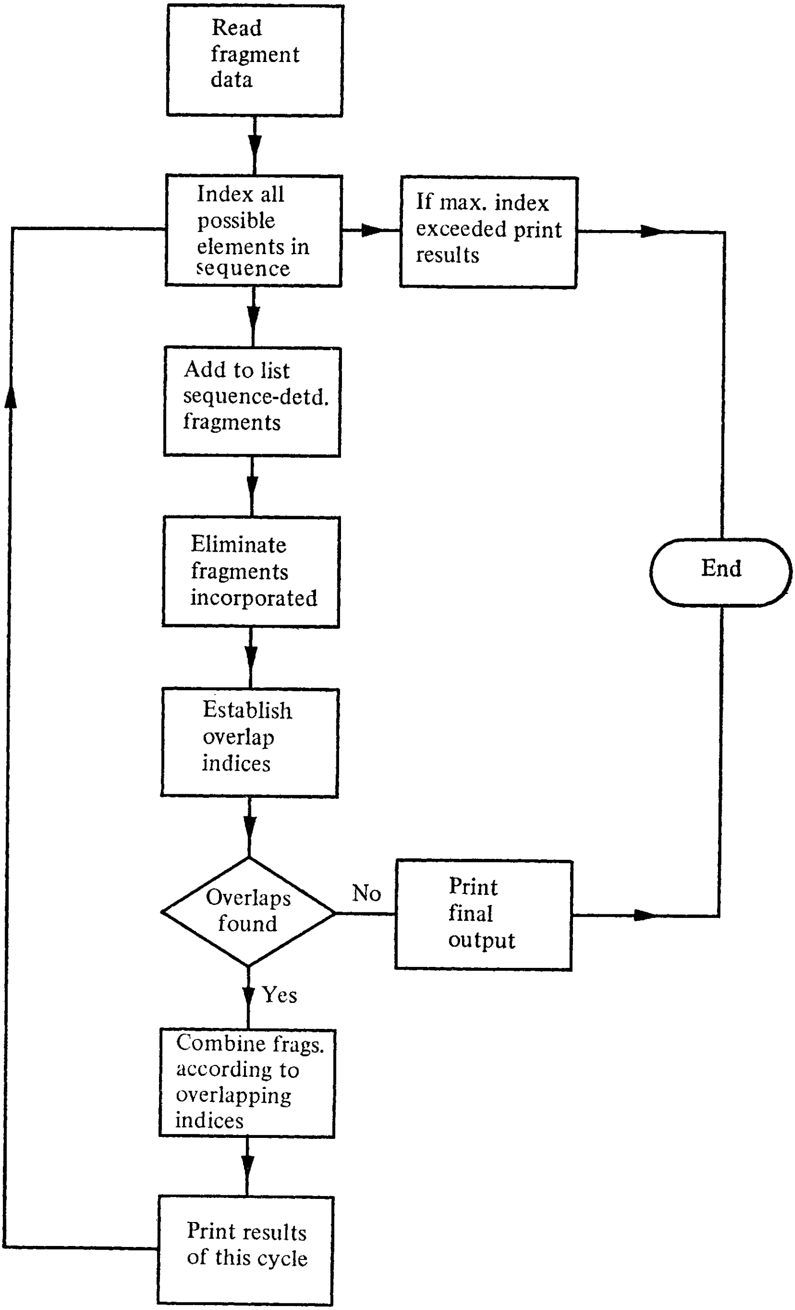
Read fragment data
Index all possible elements in sequence
If max. index exceeded print results
Add to list sequence-detd. fragments
Eliminate fragments incorporated
End
Establish overlap indices
Overlaps found
No
Print final output
Yes
Combine frags. according to overlapping indices
Print results of this cycle

Figure 3 *After Bradley, Merrill, and Shapiro (ref. 13)*

16—25). In a low-resolution system each peak has an integral *m/e* ratio but in a high-resolution system very slight differences can be recorded, *e.g.*, 27.0109 (CHN) and 27.02347 (C_2H_3). With the scanning technique, which can record a spectrum in a few seconds, a popular method is to connect the inlet of the mass spectrometer to the outlet of a gas chromatograph thus identifying the components as they appear even though the quantities involved may be minute. This combination allows the separation and identification of complex mixtures cf volatile organic compounds. The second major use is in structural work where a pure compound is decomposed and the structure elucidated from a study of the resultant spectrum. Both methods may utilise either low- or high-resolution spectrometers but as the cost of a high-resolution spectrometer is higher than that of a low-resolution type there is a tendency to use low resolution with the gas-chromatograph which gives satisfactory results, and a high-resolution type for structural work where the greater precision makes the results unequivocal but requires more calculation.

A. Low Resolution—The sample, usually a mixture of fairly low molecular weight organic compounds, may be analysed directly or may be separated by a gas chromatograph and the components analysed as they appear in succession. The latter method gives a complete spectrum for successive samples and thus some may be of a pure material while others may contain ions from more than one compound. Identification of each spectrum is simpler than that of the total mixture and is facilitated by a library of the mass spectra produced by specific organic compounds, *e.g.*, the A.S.T.M. index of mass spectral data in which the cards are sorted in order of the six most intense lines. This method has also been used with a high-resolution system and will be dealt with in more detail later.

When dealing with a mixture of compounds the mass spectrum is obviously more complicated and various additional factors have to be taken into account. For example, each compound may be ionised to a different extent and may be present in different quantities in the initial mixture thus affecting the height of the peaks produced.

Each peak on the low-resolution spectrum can be assigned an integer i ($= m/e$) and the height of this peak, P_i, produced by the substance j is propor-

[16] H. C. Hill, 'Introduction to Mass Spectrometry', Heydon and Sons Ltd., 1966.

[17] 'Mass Spectrometry of Organic Ions', ed. F. W. McLafferty, Academic Press, New York, 1963.

[18] R. I. Reed, *Quart. Rev.*, 1966, **20**, 527.

[19] J. H. Beynon, 'Mass Spectrometry and its Application to Organic Chemistry', Elsevier, Amsterdam, 1960.

[20] R. I. Reed, 'Ion Production by Electron Impact', Academic Press, London, 1962.

[21] R. I. Reed, 'Application of Mass Spectrometry to Organic Chemistry', Academic Press, London, 1966.

[22] K. Biemann, 'Mass Spectrometry', McGraw-Hill, New York, 1962.

[23] H. Budzikiewicz, C. Djerassi, and D. H. Williams, 'Interpretation of Mass Spectra of Organic Compounds', Holden-day Inc., San Francisco, 1964.

[24] 'Structure Elucidation of Natural Products by Mass Spectrometry', Holden-Day Inc., San Francisco, 1964.

[25] 'Advances in Mass Spectrometry', ed. J. D. Waldron, Pergamon Press, London, 1959.

tional to the amount x_j of that compound. The ratio of proportionality is specified as a_{ij} (see Hopp and Wertzler[26]). Hence, for one peak, from one compound we have

$$P_i = a_{ij} \times x_j$$

It has been shown that the height of such peaks are additive, *i.e.*, if two compounds 1 and 2 give peaks with the same value of i, then

$$P_i = (a_{i1} \times x_1) + (a_{i2} \times x_2)$$

so that if the sample contains n components then

$$P_i\,(i = 1, 2 \ldots m) = \sum_{j=1}^{n} a_{ij} \times x_j$$

The values of the proportionality ratios a_{ij} are known from calibration runs of pure compounds and, for a complete analysis of the observed mass spectrum it is necessary to have some idea of the compounds present, *i.e.*, the solution of the equations gives information about the relative amounts of compounds in the original mixture rather than identification of these, though unsuspected materials may be discovered by a study of the residual spectrum when analysis of the spectrum is thought to be complete in terms of the expected compounds.

Once the instrumental corrections have been made the problem is reduced to solving the m (number of peaks measured) equations for the n values of x_j which requires that $m > n$ and the equations are independent of each other. The equations still, however, contain errors of measurement plus instrumental variations which can introduce varying amounts of error.

Methods of solving the array of equations by inverting the matrix of coefficients to produce equations of the type

$$x_j\,(j = 1, 2 \ldots n) = \sum_{i=1}^{n} b_{ji} \times P_i$$

have been published by Hopp and Wertzler,[26] McAdams,[27] Gillette,[28] and Tunnicliff and Wadsworth.[29] Hopp and Wertzler use a method of triangular inversion (Gauss pivotal) which involves successive transformations of coefficients of the matrix. They include a limitation which replaces any coefficient which is calculated as a negative quantity (an impossible value) by zero and show, by calculation of the variance, that the triangular inverse method produces a more accurate set of coefficients than square inversion.

Gillette uses the Gauss–Seidel iteration to solve the equations and specifies a value, epsilon, to indicate that the required accuracy has been achieved by setting the sum of the squares of the differences between the calculated and observed peak heights to be less than or equal to epsilon. When this condition

[26] H. F. Hopp and R. Wertzler, *Analyt. Chem.*, 1958, **30**, 877.

[27] D. R. McAdams, *Analyt. Chem.*, 1958, **30**, 881.

[28] J. M. Gillette, *Analyt. Chem.*, 1959, **31**, 1518.

[29] D. D. Tunnicliff and P. A. Wadsworth, *Analyt. Chem.*, 1965, **37**, 1082.

is fulfilled the calculated values of the peaks are printed together with the delta peaks, which are the differences between the observed and calculated peaks, so that the user can check the accuracy of the analysis. It is not clear what happens if the initial postulated analysis was in error, by an amount such that the value of epsilon could not be attained by the sum of the squares of the differences, but if the program is taken off by the operator after a reasonable number of iteration cycles the delta values should indicate the nature of the component omitted.

Tunnicliff and Wadsworth include the selection of the spectra to be used in the calculation as part of the program which can select from up to 150 reference spectra, each of which can contain details of 110 mass peaks. The method, based on a stepwise regression programme by Efroymson,[30] chooses a group of spectra from a reference list which, when multiplied by a concentration factor (x_j), gives the best fit (by a least-squares criterion) to the observed spectrum. This has the advantage that the spectra to be used in the calculation are not pre-selected, except in so far that the library is pre-selected, but can, as the authors show, lose accuracy in some hypothetical cases though these may be regarded as unsolvable by hand calculation.

A factor not specifically mentioned above is that the library of spectra should be recorded under the same conditions as the sample, as with variation of temperature various effects may occur, including chemical breakdown or reaction of the components of the mixture, and the relative distribution of the fragment ions may change. The first will result in identification of the products rather than the actual components while the second may produce wrong proportions of compounds or even failure to solve the problem.

The combination of a gas chromatograph, to separate the components of a mixture, followed by mass-spectrographic analysis of the output is a powerful analytical method and is improved by the on-line use of a computer to analyse the details as discussed by Hites and Biemann[31] and Abrahamsson.[32] In the first system the output from the gas chromatograph is sampled at 4-sec. intervals over the 30 min. run and with each spectrum is recorded the sum of the data points which it contains. This sum is proportional to the unresolved ion beam current of the mass spectrometer as measured by the beam monitor and a plot of the value against the serial number of the scan is similar to the gas chromatogram. Use of this plot enables the user to select only those mass spectra of special interest to be processed further and/or retained as permanent records. The mass spectra themselves are identified by comparison with standard spectra as previously discussed.

Abrahamsson gives details of the system used for scanning the library of mass spectra (7500) which are recorded on magnetic tape and points out various methods of simplification to reduce the time required to scan the whole tape.

[30] M. A. Efroymson, 'Mathematical Methods for Digital Computers', eds. A. Ralston and H. S. Wilf, Wiley, New York, 1960.

[31] R. A. Hites and K. Biemann, *Analyt. Chem.*, 1968, **40**, 1217.

[32] S. Abrahamsson, Science Tools, LKB Instrument Journal, 1967, **14**, 29.

As the number of recorded spectra increases some form of simple check test will become important or the computer time required will be wasteful.

B. Metastable Peaks.—These peaks are produced when fragments travelling through the mass spectrograph undergo further fragmentation producing characteristic peaks which have a considerable spread of *m*/*e* when compared with a normal peak. Analysis of the components producing such peaks can be of value in the determination of alternative fragmentation paths. A program by Mandelbaum[33] calculates the possible metastable ion peaks from the equation

$$m^x = (m_2)^2/m_1$$

where m_1 and m_2 are the *m*/*e* values for the normal peaks of the fragments involved. The information required by the program is: (*a*) total number of normal peaks; (*b*) list of *m*/*e* values for these peaks; (*c*) number of metastable peaks; (*d*) list of *m*/*e* values for these peaks.

Values of $(m_2)^2/m_1$ are calculated for all permutations of the normal peaks and, if the calculated value differs from a metastable peak by less than a specified amount ($\pm 0{\cdot}003m^x$ in the example quoted), then this combination of peaks is output together with the value of the metastable peak. The number of combinations which can produce a particular metastable peak may vary and, if the number is large, it may be impossible to assign a specific combination as their origin.

C. High Resolution.—Introduction of the high-resolution mass spectrometer by means of which the *m*/*e* ratio may be quoted to five decimal places has increased the resolution of the mass spectrum so that the elemental composition of a peak, and of the difference between peaks, may be calculated with a high degree of certainty. The amount of arithmetical manipulation has also increased considerably and computer handling of the data processing is almost obligatory to avoid errors and overlooked relationships. The differences which can be detected are perhaps best shown by the following examples where the *m*/*e* ratios are expresed in terms of $^{12}C = 12{\cdot}00000$:

Low resolution	High resolution	Elements present
89	89·02659	$C_3O_3H_5$
	89·06255	$C_4O_2H_9$
88	88·01680	$C_3O_3H_4$
	88·05456	$C_4O_2H_8$

Biemann and McMurray[34] give this example in details of a program which applies certain conditions to the mass spectrum in order to establish the molecular ion peak. This peak is usually the peak corresponding to the highest *m*/*e* value but this may not be the case if (*a*) the compound breaks down so

[33] A. Mandelbaum, *Israel J. Chem.*, 1966, 4, 161.
[34] K. Biemann and W. McMurray, *Tetrahedron Letters*, 1965, 647.

readily that no molecular ion peak appears or (*b*) if an impurity is present which produces a high *m*/*e* peak. Five criteria are listed, all of which can be checked very quickly and success or failure used to steer the search for the molecular ion. The conditions specified for the peak are:

(i) the species may not contain any heavy isotopes;

(ii) the number of hydrogen atoms must be of the same parity as the number of nitrogen atoms (*i.e.*, even if even, odd if odd);

(iii) the mass differences between the peak and peaks of lower *m*/*e* must be equivalent to the loss of a reasonable chemical group;

(iv) if the highest *m*/*e* peak does not fit these criteria then the true molecular ion should be related to these peaks by combinations of atoms which can be lost in simple fragmentation processes;

(v) the elemental composition of the lower *m*/*e* fragments must not require more atoms of one type than the molecular ion processes.

These criteria can be applied to postulated systems and the authors show that the correct molecular ion could still be identified when the two peaks corresponding to M^+ and $(M - 1)^+$ were removed from the mass spectrum of androsterone.

For establishing the chemical fragments which correspond to specific differences between peaks it is useful to have a list of the possible combinations of the elements involved and the *m*/*e* value for each possible fragment. A description of a program which can produce such a list has been published by Tunnicliff, Wadsworth, and Schissler[35] in which they specify the requirements necessary to keep the number of fragments listed within reasonable bounds. These involve limitations, on the number of types of atom allowed or the relative abundances of particular types of atom. The limitations serve to keep the size of the final tables to a manageable size and also restrict the time required for sorting the results into a list of increasing mass.

A similar system to that already described for low-resolution mass spectrometry by Hites and Biemann and by Abrahamsson has been described by Bowen, Chenevix-Trench, Drackley, Faust, and Saunders[36] for a high-resolution system. The postulated system involves an Argus 500 computer directly linked to a high-resolution mass spectrometer to handle both on-line recording and data processing. A 10-sec. scan is sampled and, if every measurement were recorded, would soon overload the data storage. To avoid this, only values which exceed a specified minimum are recorded, thus excluding general background and random noise, only using storage for necessary information. In addition, time is thus provided between peak recording for calculations to be carried out. Calibration involves the presence of a reference compound and the *m*/*e* peaks of the sample are calibrated by interpolation from the known peaks. A perfluoro-compound is recommended for a reference compound if the sample does not contain fluorine, as there is less tendency for overlapping to occur.

[35] D. D. Tunnicliff, P. A. Wadsworth, and D. O. Schissler, *Analyt. Chem.*, 1965, **37**, 543.

[36] H. C. Bowen, T. Chenevix-Trench, S. D. Drackley, R. C. Faust, and R. A. Saunders, *J. Sci. Instr.*, 1967, **44**, 343.

2 Peptide Sequence Determination

Determination of the amino-acid sequence of a peptide or protein by mass spectrometry is an attractive proposition. By an analysis of the *m/e* peaks obtained it should be possible to establish the exact sequence of a very small amount of material. Two main difficulties arise, however, the first being due to the stability of the material, *i.e.*, proteins in general are difficult to volatilise and this stability increases with the size of the molecule. The usual way to overcome this is to attach a known group to the *N*-terminal group to form an ester which will be more volatile than the free peptide. A variety of groups has been used for this and the group may serve a second purpose by acting as a starting point for the analysis of the data. The second difficulty arises when breakdown of side-chains occurs in addition to the breaking of the primary chain. This occurs when residues with sizeable side-chains, *e.g.*, asparagine, proline, serine and may also be affected by interactions with neighbouring side-groups.

Thus, the three main areas of interest are: (*a*) choice of end-group substituent; (*b*) investigation of possible modes of break-down; (*c*) programming of computer to reconstruct the primary chain.

Biemann, Gapp, and Sieble[37] reduced small peptides with $LiAlH_4$ to produce polyamino-alcohols and recorded the low-resolution mass spectra of these products. They found preferential rupture of the primary chain and the bond connecting the side-group to the chain. They point out that a peak corresponding to $M + 1$, produced by an ion–molecule collision, occurs, but this is recognised by its variation of relative intensity with change of pressure and focusing conditions.

Discussions of the terminal substituents and the types of side-chain rupture have been published by Shemyakin *et al.*[38] and by Ovchinnikov *et al.*[39] These papers specify acylation of the free peptide followed by methylation to produce compounds which are sufficiently volatile and break down in a reasonably simple manner. In addition to the breaks at the amide bonds, the resultant fragments can lose the elements of CO producing related peaks and metastable peaks which can be used to confirm the mechanism. Additional factors which may produce partial fragmentation are the nature of the side-groups and the relationships between neighbouring side-groups. These factors are listed below in terms of the nature of the side-group:

> glycine, alanine—little fragmentation; valine, leucine, iso-leucine—some fragmentation of a simple nature; methionine—may lose all the side-chain, or, if some interaction with neighbouring groups occurs, may lose only part of the side-chain; proline—may condense the ring; serine, threonine, and cysteine—may lose the functional substituent; cystine—S—S bond is easily

[37] K. Biemann, F. Gapp, and J. Sieble, *J. Amer. Chem. Soc.*, 1959, **81**, 2274.

[38] M. M. Shemyakin, Yu. A. Ovchinnikov, A. A. Kiryushkin, E. I. Vinogradova, A. I. Miroshnikov, Yu. B. Alakhov, V. M. Lipkin, Yu. B. Shvetsov, N. S. Wulfson, B. V. Rosinov, V. N. Bocharev, and V. M. Burikon, *Nature*, 1966, **211**, 361.

[39] Yu. A. Ovchinnikov, A. A. Kryushkin, E. I. Vinogradova, B. V. Rosinov, and M. M. Shemyakin, *Biokhimiya*, 1967, **32**, 427 (*Biochemistry* U.S.S.R., 1967, 351).

broken; asparagine and aspartic acid—elimination of the β-substituent, by losing first the elements of ammonia (or alcohol) followed by the elements of CO; glutamine and glutamic acid esters—similar steps to asparagine plus some breaks in the C—C bonds of the side-chain; γ-methyl esters of α-glutamic acid—this can lose the elements of water with the formation of a ring to the neighbouring amide group; phenylalanine, tyrosine, histidine, and tryptophan—side-chain elimination as RCH_2 or RCH_2^+ or cleavage of the N—Cα bond.

It is suggested that a method of overcoming the difficulty of vaporising the peptide (for long-chain samples) would be to apply Edman's method to split off several residues followed by mass spectrometry of the remainder.

A computer program to elucidate the peptide sequence of a chain, even before the mass spectrum is considered, has to be organised to start the calculation either from the molecular ion and work downwards to fragments of lower *m/e* or to start with a recognisable end-group and build up the sequence by addition of peptide fragments.

Barber *et al.*[40] say that the second method was not found to be satisfactory and describe a program which starts from the molecular ion. Additional chemical data are provided, if available, and checks are carried out to make all the data compatible; *e.g.*, an example quoted shows that the amino-acid composition data can be changed to fit the molecular formula and molecular weight. During the calculation three types of fragmentation are considered: (*a*) linear peptide break; (*b*) (cyclic peptides)—fragmentation with the loss of an amino-acid residue plus the elements of ammonia; (*c*) as (*b*) but with the loss of an amino-acid residue less the elements of CO.

In addition, the program is designed to allow for cyclodepsipeptides which contain hydroxy-acids in addition to amino-acids residues.

Other programs, by Gavrilov *et al.*,[41] Senn *et al.*,[42] and Biemann *et al.*,[43] start from the other end of the mass spectrum. Gavrilov *et al.* discuss an algorithm* for establishing the chain sequence from low-resolution data of the fragments produced by hydrolysis, *i.e.*, a combination of the first stage of the amino-acid sequence analysis by chromatographic methods and identification of the fragments produced by their mass spectra. The sorting process described is similar to those previously described, and the authors claim success for chains of up to twenty-five residues even in cases where some peaks do not occur. In the latter case an alternative system is proposed in which partial chains are synthesised and finally combined to produce a single chain, though this appears to produce certain ambiguities.

* Algorithm: a mathematically unambiguous set of instructions.

[40] M. Barber, P. Powers, M. J. Wallington, and W. A. Wolstenholme, *Nature*, 1966, **212**, 784.

[41] V. Yu. Gavrilov, A. D. Frank-Kamenetskii, and M. D. Frank-Kamenetskii, *Biokhimiya*, 1966, **81**, 799 (*Biochemistry* U.S.S.R., 1966, 689).

[42] M. Senn, R. Venkataraghavan, and F. W. McLafferty, *J. Amer. Chem. Soc.*, 1966, **88**, 5593.

[43] K. Biemann, C. Cone, B. R. Webster, and G. P. Arsenault, *J. Amer. Chem. Soc.*, 1966, **88**, 5598.

The programs by Senn *et al* and by Biemann *et al.* identify the *N*-terminal substituent group. This process is simplified if the substituent group is markedly different from the normal peptide side-chains, *e.g.*, trifluoroacetyl, and volatility is improved if the terminal carboxy-group is also esterified. Once the *N*-terminal group is found, each possible amino-acid residue is added in turn and a search made of the mass spectrum for the presence of a corresponding peak. This search is obviously shortened if a previous analysis can limit the number of possibilities to be considered. To allow for different fragmentation mechanisms along the chain the search is also repeated for each residue less the elements of CO. A successful result of either search establishes the *N*-terminal residue and the process can be repeated to find the next residue and so on. If an ambiguity arises where two possible residues are found, then the next stage is carried out for both combinations and, if necessary, repeated until one can be rejected. At each stage, in addition to the various amino-acid residues, the ester group OR is tested for the molecular ion. The whole process may also start using the *C*-terminal ester portion. These processes can also be used to check for fragments which are produced by side-chain rupture or re-arrangement if the appropriate values are calculated from chemical considerations.

In all the programs discussed in this article it is essential that the maximum amount of information obtained by the program is output for consideration. This is necessary even if all appears to have fitted perfectly since a critical examination of the results may reveal possible variations which are not considered in the program. Programs can be quite complicated and yet still fail to produce a correct answer merely because of some unforeseen item.