Journal of Chromatography, 297 (1984) 235-244
Elsevier Science Publishers B.V., Amsterdam — Printed in The Netherlands

CHROMSYMP. 335

CLUSTER ANALYSIS OF CHROMATOGRAPHIC PROFILES OF URINE PROTEINS

ROGER J. MARSHALL*, RONALD TURNER, HAI YU and EDWARD H. COOPER
Unit for Cancer Research, University of Leeds, Leeds (U.K.)

SUMMARY

A method to examine and classify the chromatograms of urinary proteins, separated on a fast protein liquid chromatography system, is presented. For the analyses a measure of similarity between chromatogram profiles is defined and this is used to implement a statistical cluster analysis technique for the identification of a system of classification. The resulting classification can be related to a clinical assessment of the proteinuria of the sample urines, which are from patients with various renal disorders.

INTRODUCTION

Recent advances in the high-performance liquid chromatographic (HPLC) separation of proteins have opened up new possibilities for the application of this technique to the analysis of complex protein mixtures in body fluids of patients. Preliminary studies of the chromatography of urine proteins on a novel mono-disperse anion exchanger in the past protein liquid chromatographic (FPLC) system has shown that the profiles can often distinguish between glomerular and tubular proteinuria[1]. This suggested that the automated recognition of patterns of the traces of the HPLC separation of proteins in body fluids was a reasonable study to run in parallel with research on the ways to improve the resolution of the chromatographic separation of these proteins. This paper describes the preliminary results from an analytical system designed to examine and classify the chromatograms of urine proteins, separated by FPLC, through the use of statistical cluster analysis techniques.

MATERIALS AND METHODS

Urine samples

Urine samples were obtained from a series of patients with clinically well defined disorders which either affect the filtration and reabsorption of proteins by the kidney or exudation of proteins in the urine by the lower urinary tract. The classification of these samples into the types of proteinuria, as defined by sodium dodecyl sulphate (SDS)-polyacrylamide gel electrophoresis and by the immunochemical measurement of certain key proteins, is shown in Table I.

0021-9673/84/$03.00

TABLE I

CLASSIFICATION BASED ON SDS-POLYACRYLAMIDE GEL ELECTROPHORESIS AND IMMUNOCHEMICAL MEASUREMENT OF SEVEN KEY PROTEINS

Proteinuria	*Diagnosis*
Glomerular	Nephrotic syndrome
Mixed glomerular and tubular	Renal failure, burns (early), acute pyelonephritis, renal transplants
Tubular	Acute pyelonephritis, burns (early and late), renal transplants, *cis*-platinum nephrotoxicity, chronic liver disease
Post-renal	Acute cystitis
Other	Fever

Chromatography

A 0.5-ml sample of urine was desalted on a Sephadex G-25 column (300 × 16 mm I.D.) and the protein fraction chromatographed on a Mono Q anion exchanger HR 5/5 column in the FPLC system (Pharmacia, Uppsala, Sweden), using a bis Tris propane buffer with a NaCl 0–0.35 *M* and pH 7.5–9.5 gradient to elute the proteins; this procedure has been described previously[1,2].

Data base construction

In anticipation of the accumulation of data as the work advances we have set up a system of data storage and retrieval using the Leeds University AMDAHL 470V/7 computer. In this preliminary study chromatograms were transferred to this computer after manual digitisation on an ICL PERQ digitising tablet giving 200–300 data points per trace. This procedure has been used for 71 urinary protein profiles of the diseases listed in Table I. Eventually, it is planned to replace this manual procedure by introducing a microprocessor to capture and filter the signal as it is generated, and to feed it directly to the AMDAHL. Digitised FPLC chromatograms are stored on disk, from which specific chart can be selected by direct access.

MATHEMATICAL METHODS

Cluster analysis

Cluster analysis[3,4] is a general term to describe various statistical approaches to the classification of objects. Broadly, these methods can be divided in hierarchical and optimal partitioning methods. Usually data is assumed to be of the form of a multivariate vector rather than as an analogue signal, such as a chromatogram. However, once a matrix of similarity measures between all pairs of objects is computed, many of the methods can be implemented in terms only of this similarity matrix, which we shall denote by D. To apply cluster analysis to chromatogram profile classification it is therefore necessary to consider measures of similarity between pairs of profiles. For electrophoretic profiles a measure based on the number of coincident peak positions has been suggested[5]. An alternative approach, which is adopted here, is to measure the separation between chromatograms.

Similarity measures

Suppose that $X_1(t)$ and $X_2(t)$ represent two chromatogram signals, where t

denotes elution time or volume. A measure of the distance or separation of X_1 and X_2 is as follows

$$d(X_1, X_2) = \left[\int_a^b \{X_1[t + h_1(t)] - X_2[t + h_2(t)]\}^2 w(t) \, dt\right]^{1/2} \tag{1}$$

Here the integration limits, a and b, represent the end points of the section of the chromatogram of interest and the functions $h_1(t)$ and $h_2(t)$ are introduced to allow for possible non-alignment of the peaks of X_1 and X_2 corresponding to the same constituent protein. They can be obtained by a method which is outlined below. A weighting function $w(t)$ is also incorporated in eqn. 1 to enable more weight to be given to the separation of X_1 and X_2 at sections of the chromatograms which are of particular interest than to sections where, for instance, peaks are not expected.

Eqn. 1 is termed a dissimilarity measure, since it increases with increasing dissimilarity (distance), and it is a measure of the separation of X_1 and X_2 in absolute terms. It may not, for instance, capture the closeness of profiles which are similar in shape but which differ in order of magnitude. An alternative measure, which will discriminate profile shape, can be obtained by standardising X_1 and X_2 so that the area under each is equal to one, that is, by putting

$$x_1(t) = \frac{X_1(t)}{\int_a^b X_1[u + h_1(u)] \, du}; \quad x_2(t) = \frac{X_2(t)}{\int_a^b X_2[u + h_2(u)] \, du} \tag{2}$$

and substituting $x_1(t)$ and $x_2(t)$ for $X_1(t)$ and $X_2(t)$ in eqn. 1. Henceforth, $X_1(t)$ and $X_2(t)$ are referred to as raw chromatograms and $x_1(t)$, $x_2(t)$ as standardised chromatograms. Analyses based on both are discussed below.

Profile timing adjustment

The timing adjustment functions $h_1(t)$ and $h_2(t)$ can be obtained in a manner which is a generalisation of other reported methods[6,7]. Suppose that $T_1, T_2, ..., T_M$ is a set of reference elution times of M usually well defined and well separated peaks. A profile adjustment function $h(t)$ can be obtained by aligning the corresponding observed peaks with these reference positions as follows: define for $T_1, T_2, ..., T_M$ a set of non-overlapping intervals $I_1, I_2, ..., I_M$. The interval width must be judged, by experience, from the variability in the time of the peak identified with the reference point T_i. If, for an observed chromatogram, a significant peak occurs in I_i then it is assumed to be identified with T_i and is aligned with it. This is done for each $i = 1, ..., M$ and $h(t)$ is determined by piecewise linear transformations of the time scale as is illustrated in Fig. 1 for $M = 5$.

Typically, M will be small; we took $M = 3$ in our analyses to correspond to three well-identified peaks, representing the proteins β_2-microglobulin (β_2-m), acid glycoprotein (AGP) with α_1-microglobulin (α_1-m), and albumin. These peaks are eluted characteristically in the intervals (4.5 ± 2), (14.5 ± 1.5) and (19 ± 3) min, so defining the intervals I_1, I_2 and I_3 with the midpoints defining T_1, T_2 and T_3.

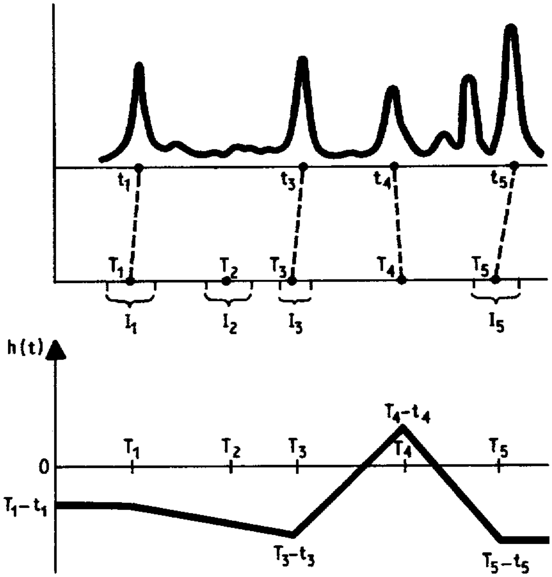

Fig. 1. The profile timing adjustment function for $M = 5$. The appropriate observed peak positions are aligned with the reference set, and h(t) is comprised of a set of linear transformations.

The weighting function

The function w(t) was chosen in this study to give most weight to four regions of peak activity of the chromatogram. These correspond to the three intervals I_1, I_2 and I_3, mentioned above, and a fourth region representing the section of the chromatogram between I_1 and I_2 where retinol binding protein and transferrin, if presented are eluted. The weight function was defined as a series of Gaussian curves centred on I_1, I_2 and, for the fourth region, centred at the 10 min elution mark. The Gaussian curves were each given a standard deviation value of 1.0 except for the I_2 region where a smaller value, 0.75, was used in an attempt to increase the contribution of the β_2-m peak which, although of equal importance is typically narrow and sharper than other peaks. This system of weightings highlights the dissimilarities between chromatograms in terms of peak behaviour, and it is found to be superior to using no weighting at all. It is possible that there is scope for improvement on our choice of w(t).

Computation

The integrals in eqns. 1 and 2 must be computed numerically. For a digitised chromatogram in the form of a set of X(t) values at irregularly spaced time points the integrals in eqn. 2 can be computed using the trapezoidal rule after making the appropriate timing transformation. The evaluation of eqn. 1 presents a little more difficulty since digitised time points $X_1(t)$ and $X_2(t)$ will not normally coincide, at least in manual digitisation at irregular intervals. However, these time points can be interspersed so that at each digitised point of X_1 the corresponding X_2 value, allowing for the profile timing adjustment, can be estimated by linear interpolation and *vice* *versa*. The integration can then be carried out using the trapezoidal rule on the interspersed points.

Using these methods the dissimilarity matrix, D, can be computed by calculating the dissimilarities between each of the N $(N - 1)/2$ profile pairs in a sample of size N. All computations were carried out on the AMDAHL computer using programs written in FORTRAN.

Once D is computed, the CLUSTAN suite of cluster analysis programs[7] can conveniently carry out hierarchical cluster analysis using the CLUSTAN option which allows a user-specified dissimilarity matrix. However, the CLUSTAN procedures for optimal partitioning methods cannot be utilised, given only D, since they require multivariate data as input. Nevertheless, partitioning methods can also be specified in terms of D, and we have made use of a useful algorithm for this purpose described by Spath[4]. For a fixed number of clusters this method allocates objects to clusters in such a way that the total within cluster variability (TWCV) is minimum. If this algorithm is run 1, 2, 3 etc. clusters it may be possible to estimate how many clusters provide the best partition of the data by inspecting the reduction in the TWCV for each run. A significantly large reduction indicates a possible natural set of clusters.

Chromatogram variability

In comparing a number of chromatograms it is useful to be able to plot a region which encapsulates the average chromatogram shape and its variability. One such region can be specified in terms of percentiles. At an elution time, t, the upper and lower pth percentiles of a set of chromatogram values can be estimated by standard statistical methods. If these values are $X_p(t)$ and $X_{1-p}(t)$ the region between $X_p(t)$ and $X_{1-p}(t)$ taken over $a < t < b$ defines a band which outlines the shape and variability of chromatograms, and can be termed a $100 \times (1 - 2p)\%$ band. This concept is illustrated below.

RESULTS AND DISCUSSION

The dissimilarity matrices of 71 urinary chromatogram profiles were calculated, by the methods outlined, for both the standardised and raw chromatograms. We shall use D_1 and D_2, respectively, to distinguish the matrices for hese two cases, and we consider firstly the cluster analyses on D_1.

Fig. 2 shows a dendrogram obtained by the method of single-linkage hierarchical clustering. No particularly well-defined grouping emerge from this plot which is typical of single linkage dendrograms on data in which clusters may exist, but in which the clusters are connected by intermediate objects: an effect known as chaining. Chaining is probably occurring here, since demarcation between chromatogram classifications will probably not be precise. Other hierarchical clustering methods are available to reduce the chaining effect. One of these, Ward's method[8], produces the dendrogram in Fig. 3, when applied to D_1. This diagram suggests a partition of the chromatograms into six clusters, and the 80% bands of the standardised chromatograms for each of these clusters are shown in Fig. 4. Table II shows how the samples are distributed among the six clusters by disease.

The allocation in Table II and the interpretation of patterns in Fig. 4 are

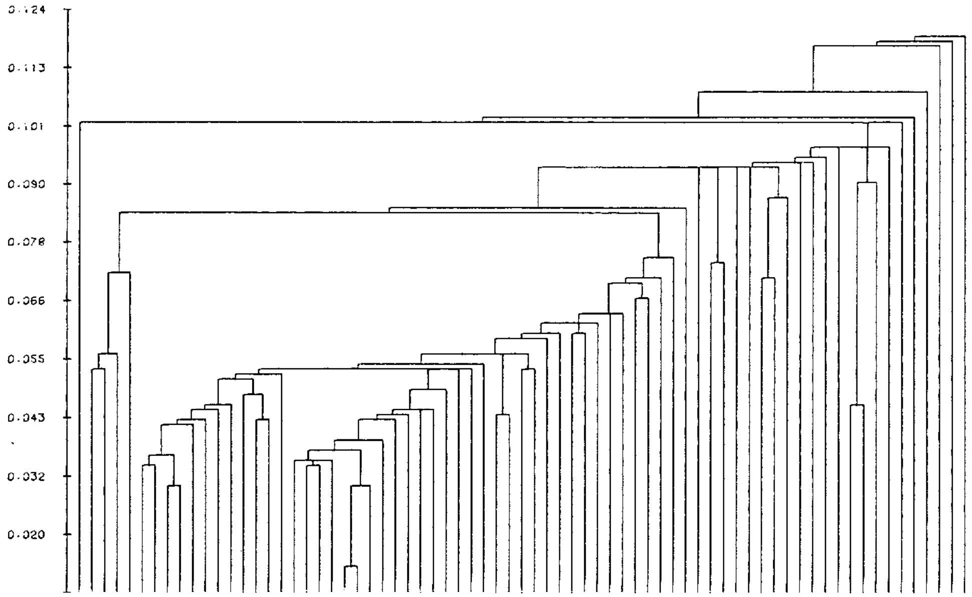

Fig. 2. The dendrogram of standardised chromatogram profiles using single linkage hierarchical classification.

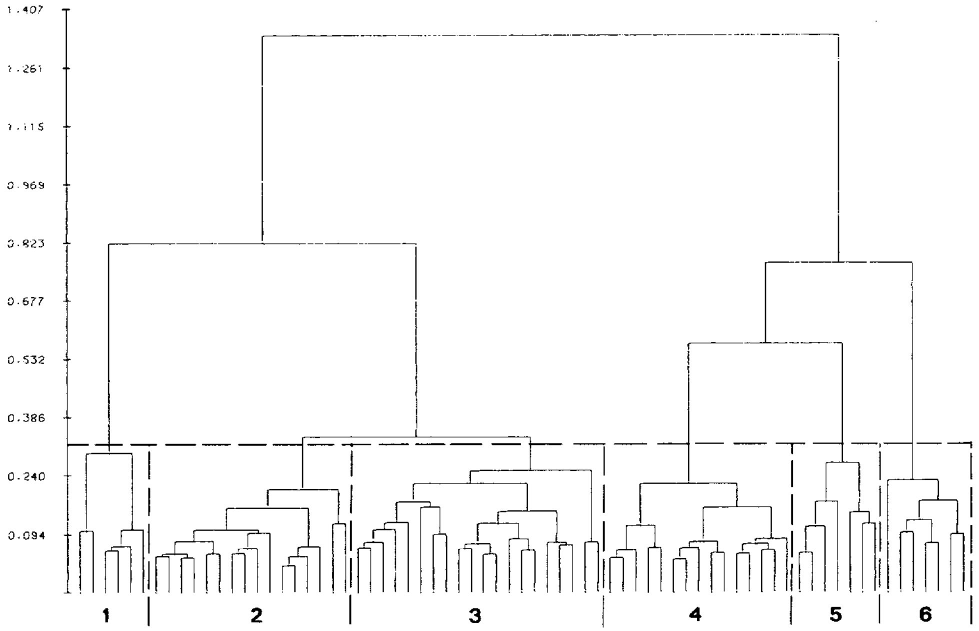

Fig. 3. The dendrogram of standardised chromatogram profiles using Ward's method. Six main clusters can be identified.

TABLE II

THE ALLOCATION BY DISEASE OF CHROMATOGRAMS AMONG THE SIX CLUSTERS IDENTIFIED IN FIG. 3

The disease codes are: FV = fever; PY = acute pyelonephritis; EB = early burns; DB = late burns; CI = *cis*-platinum nephrotoxicity; LI = liver disease; RT = renal graft; RF = chronic renal failure; NP = nephrotic syndrome; AC = acute cystitis.

Cluster	*Disease group*										*Cluster totals*	*Mean total protein (mg/l)*
	FV	*PY*	*EB*	*DB*	*CI*	*LI*	*RT*	*RF*	*NP*	*AC*		
1	4	2	–	–	–	–		–	–	–	6	430
2	–	1	4	5	5	–	1	–	–	–	16	730
3	–	5	2	1	–	2	3	7	–	–	20	1187
4	–	3	3	–	–	2	1	1	–	5	15	2278
5	–	–	1	–	–	–	4	–	–	2	7	1424
6	–	–	–	–	–	–	–	–	7	–	7	14900
Disease totals	4	11	10	6	5	4	9	8	7	7		

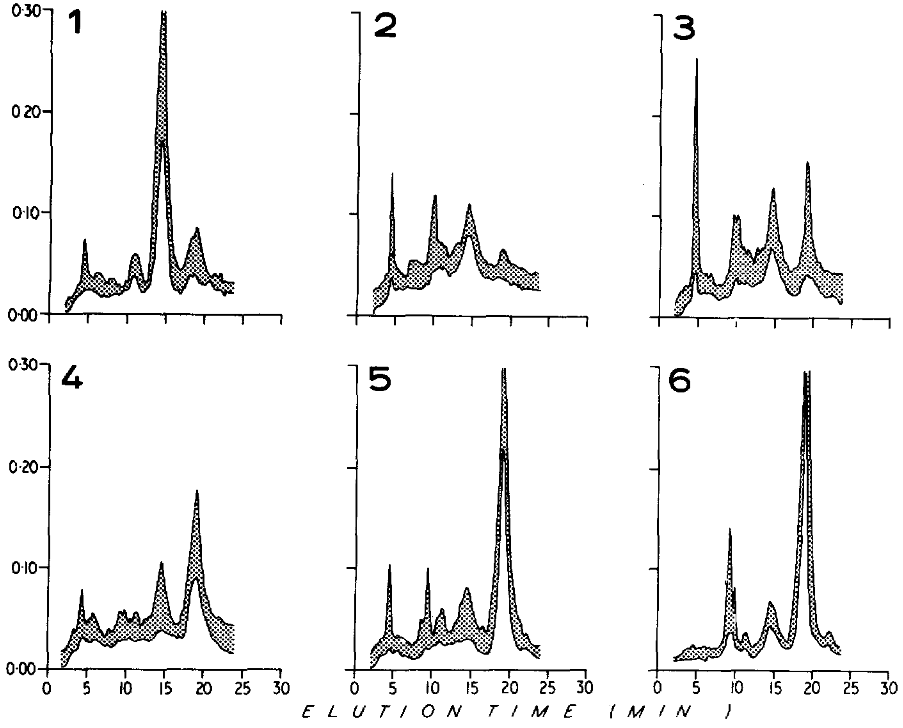

Fig. 4. The characteristic standardised chromatogram patterns of each of the clusters identified in Fig. 3. The 80% bands are shown.

interesting from a clinical point of view, since the classification can be identified, to some extent, with clinical assessment of the proteinuria in the disease categories. Cluster 1, for example, contains chromatograms dominated by a single peak, representing acid glycoprotein (AGP), and its comprises all the fever patients with only two other cases. Excessive AGP proteinuria in fever is due to its high concentration in the blood, causing an overload. The chromatogram pattern in cluster 2 is typical of tubular proteinuria, in which albumin and other high molecular weight proteins do not contribute significantly. The disease cases assigned to this cluster are also, classically, tubular in origin. Cluster 3 can also be identified with tubular type proteinuria but now with a significant albumin content suggesting some degree of glomerular dysfunction. There are close similarities between clusters 4 and 5, except that the albumin peak in cluster 5 is more pronounced. These two clusters can also be identified with mixed glomerular and tubular proteinuria. However, as the albumin content in both clusters is significantly larger than it is in cluster 3, the proteinuria is probably more glomerular in origin. Note that cluster 4 also contains most of the acute cystitis group, with post-renal proteinuria dominated by the exudation of albumin into the urine. Finally, cluster 6 contains only the cases of nephrotic syndrome. Their chromatogram patterns are characterised by a dominant albumin peak and the absence of a β_2-m peak.

An alternative analysis, also based on D_1 but using Spath's[4] optimal partitioning algorithm, has also been carried out to provide a comparison with the above system of classification. When specifying 6 clusters, the allocations in Table III are obtained, and these are in broad agreement with Table II. The algorithm was also run with 1, 2, ... 10 clusters, but at no point was there a significant reduction in the TWCV, rather a gradual decay occurs, supporting the assertion that demarcation between clusters is not clearly defined.

It is also instructive to repeat these analyses using the dissimilarity matrix D_1, which is based on raw chromatogram profiles. The dendrogram for this matrix, using Ward's method, is shown in Fig. 5. Three clusters can be discerned and the distribution of cases in each of these is given in Table IV. Cluster C in Fig. 5 is widely

TABLE III

THE ALLOCATIONS OF CASES AMONG SIX CLUSTERS BY OPTIMAL PARTITIONING

See Table II for disease codes.

Cluster	*Disease group*										*Cluster totals*
	FV	*PY*	*EB*	*DB*	*CI*	*LI*	*RT*	*RF*	*NP*	*AC*	
1	4	1	–	–	–	–	–	–	–	–	5
2	–	4	5	6	5	–	2	5	–	–	27
3	–	3	3	–	–	2	3	2	1	–	14
4	–	3	1	–	–	2	–	1	–	5	12
5	–	–	1	–	–	–	4	–	–	2	7
6	–	–	–	–	–	–	–	–	6	–	6
Disease totals	4	11	10	6	5	4	9	8	7	7	

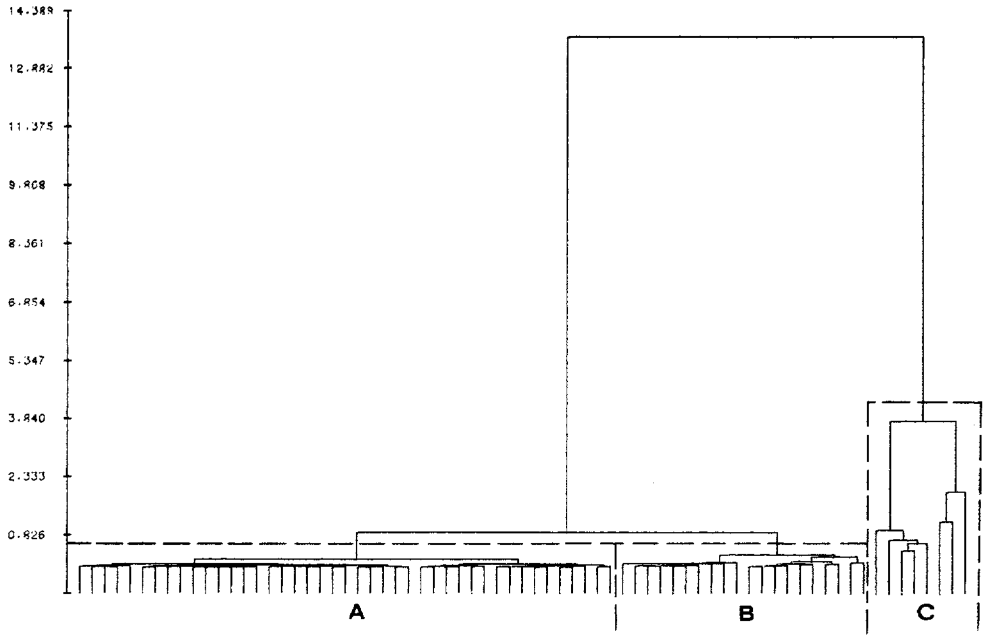

Fig. 5. The dendrogram of raw chromatograms using Ward's method. The clusters A, B and C can be identified.

separated from the other two and it comprises all the nephrotic syndrome patients and a case of renal failure. This separation is not unexpected, since the total protein of these cases is about ten fold that of all others. That the nephrotic cases are clustered both in Fig. 5 and in Fig. 3 reflects the homogeneity of chromatograms for this group, both in shape and in intensity. The isolated renal failure case in cluster C is, however, different in character and is assigned to cluster 3 in Fig. 4. The separation between clusters A and B in Fig. 5 is less marked. Broadly, however, A comprises the pure tubular and mixed proteinuria cases with moderate amounts of total protein, whilst B contains mixed proteinuria cases with larger quantities of albumin and

TABLE IV

THE ALLOCATION OF CASES BY DISEASE, TO THE THREE CLUSTERS A, B AND C IDENTIFIED IN FIG. 5

See Table II for disease codes.

Cluster	*Disease group*										*Cluster total*	*Mean total protein (mg/l)*
	FV	*PY*	*EB*	*DB*	*CI*	*LI*	*RT*	*RF*	*NP*	*AC*		
A	3	8	5	5	5	4	4	5	–	4	43	685
B	1	3	5	1	–	–	5	2	–	3	20	1472
C	–	–	–	–	–	–	–	1	7	–	8	13738
Disease totals	4	11	10	6	5	4	9	8	7	7		

greater total protein content. The mean total protein in each of these clusters is shown in Table IV.

It is interesting to note that when clusters 1, 2 and 3 of Fig. 3 are combined, the cases correspond quite closely to those of cluster A in Fig. 5. Similarly clusters 4 and 5, when combined, correspond to cluster B. This can be checked by combining the appropriate rows of Table II and comparing with Table IV. Thus the classifications that are derived using the matrix D_1 appear to account not only by shape but also for total protein content; that is, there is evidently an inherent association between chromatogram shape and chromatogram intensity. This assertion is supported by a comparison of the mean total protein content, in each of the clusters in Fig. 3. These values are shown in Table II.

In summary, it has been shown how it is possible to sort urinary protein chromatograms from an anion exchanger into groups which are clinically meaningful. That this can be done, provides a basis for the use of the patterns of protein or peptide chromatograms to aid clinical assessment in certain diseases. This work naturally extends to the development of a system to automatically assign individual chromatograms to pattern classes, that is, to implement a pattern recognition system. The distance measure, eqn. 1, may be suitable for this purpose by assigning an individual to the reference pattern to which it is closest.

ACKNOWLEDGEMENTS

We are grateful to Professor M. Wells for his advice on computational aspects and to Dr. J. T. Brocklebank for permission to investigate his patients. Roger Marshall is supported by the Yorkshire Cancer Research Campaign.

REFERENCES

1 E. H. Cooper, R. Turner, E. A. Johns, H. Lindblom and V. J. Britton, *Clin. Chem.*, 29 (1983) 1635.

2 H. Lindblom, U.-B. Axio-Fredriksson, E. H. Cooper and R. Turner, *J. Chromatogr.*, 273 (1983) 107.

3 R. M. Cormack, *J.R. Statist. Soc. Series A*, 134 (1971) 321.

4 H. Späth, *Cluster Analysis Algorithms for Data Reduction and Classification of Objects*, Ellis Horwood, Chichester, 1980, p. 117.

5 P. J. H. Jackman, *Microbios Lett.*, 23 (1983) 119.

6 D. D. Chilcote, *Clin. Chem.*, 19 (1973) 826.

7 D. D. Chilcote and C. D. Scott, *Anal. Chem.*, 45 (1973) 721.

8 J. H. Ward, *J. Am. Stat. Assoc.*, 58 (1963) 236.