A quantitative representation of molecular surface shape. II: Protein classification using Fourier shape descriptors and classical scaling

Steve Leicester [a], John Finney [b] and Robert Bywater [c,1]

[a] *Department of Crystallography, Birkbeck College, London WC1E 7HX, UK*

[b] *Department of Physics and Astronomy, University College London, Gower Street, London WC1 E6BT, UK*

[c] *Biostructure Department, Novo Nordisk A/S, Bagsværd, DK-2880 Denmark*

Received 20 December 1993

A quantitative means of describing molecular shape has been described and developed in the first paper of this series. In this second paper these techniques are further developed to enable the classification of protein "surface" shape. Essentially the molecular surface shape descriptors are considered as components of a vector in a multidimensional space. The closer two vectors are to each other, the more similar are the shapes from which the vectors were derived.

1. Introduction

A general method of obtaining global shape descriptors of the "surface" of a molecule from a three dimensional coordinate structure of the molecule has been developed [1,2]. The process is able to deal with reentrant regions by generating several single valued functions, called subsurfaces, from what would otherwise be a single multivalued function. In this paper, work carried out on a number of protein molecules is described. The actual proteins used are shown in table 1 along with a reference key and their protein data bank code [3]. In section 2 details of obtaining the shape descriptors for protein molecules is described. The shape descriptors may be viewed as a shape spectral representation of the molecule and this is also seen in this section.

The actual technique of shape comparison is examined in more detail in section 3. In particular it is important to see how the results may depend on the number of shape descriptors and also the number of subsurfaces. This will provide a

[1] To whom requests for reprints should be sent.

© J.C. Baltzer AG, Science Publishers

Table 1
Key to protein molecules used in this study.

Key	Data bank code	Molecule
A	2ADK	adenylate kinase
B	2SSI	subtilisin inhibitor
C	4ATC	aspartate carbamoyltransferase
D	4CPA	carboxypeptidase A
E	1CTS	citrate synthase
F	1EST	elastase
G	4CHA	α-chymotrypsin
H	2GCH	γ-chymotrypsin
I	2DFR	dihydrofolate reductase
J	2PTC	trypsin
K	2SBT	subtilisin
L	1LYZ	lysozyme
M	2SNS	staphylococcal nuclease
N	2SGA	proteinase A
O	2PKA	kallikrein A
P	4LDH	lactate dehydrogenase
Q	2APE	acid proteinase
R	1ALP	alpha-lytic protease
S	2PAD	papain
T	2CAB	carbonic anhydrase B
U	3PGK	phosphoglycerate kinase
V	1BP2	phospholipase 2
W	2ACT	actinidin
X	3RP2	proteinase II (structure A)
Y	3RP2	proteinase II (structure B)
Z	1GCR	gamma crystallin (both domains)
a	1GCR	gamma crystallin (domain A)
b	1GCR	gamma crystallin (domain B)
c	5CPA	carboxypeptidase A
d	PTC2	trypsin inhibitor

better understanding of any limitations of the method and thus help interpretation of later results. In section 4 the method is then applied to a large number of protein molecules. This should then enable the identification of groups of proteins that are related by surface shape. Section 5 is devoted to some concluding remarks.

2. Significance of the descriptors

2.1. OBTAINING SHAPE DESCRIPTORS FOR PROTEINS

All protein structures were taken from the protein data bank [3]. Before surface points were generated the protein data bank (PDB) files were edited to remove, in particular, water molecules as well as other structures that were not part of the actual amino acid sequence or sequences, the intention being to deal in the first instance with a straightforward protein structure. Necessary symmetry operations were also applied in the case of oligomeric macromolecules. From these modified PDB files a Connolly dot surface was generated [4]. For this, the probe radius was set at 1.4 Å, thus resulting in points lying on the solvent accessible surface being generated. An alternative choice would be a probe radius of 0 Å, giving the Van der Waals surface. Either choice will provide points which represent the molecular shape and is therefore unlikely to have much effect on the results. However, because of the importance of solvent in molecular interactions, the solvent accessible surface is usually considered a more appropriate surface model. In addition, the solvent accessible surface is smoother than the Van der Waals surface and therefore more easily represented by spherical harmonics.

The surface point density was set at 5/ Å^2. This latter value must be considered as a compromise since as high point densities as possible are desirable. However fig. 1 shows the effect of increasing the point density on the zeroth order descriptor of chymotrypsin (for the first subsurface) and as can be seen there is little change above the value of 5/ Å^2. Note that to obtain Connolly surfaces for very high densities required much memory space which was generally not available. Hence the value of 5/ Å^2 was chosen as a practical working value for the protein molecules.

The Connolly dot surface was used as input to the mapping program. For the protein molecules the mapping program produced from the dot surface six piecewise continuous single valued surface functions or subsurfaces ($r^k(\theta, \phi)$, $k = 1$, ... 6). Actually, in some cases six functions were insufficient as will be seen later but this did not alter any results obtained. To resolve reentrants the separation or step (see [2] was set at 1.4 Å and the maximum search radius was set at ten degrees (i.e. with one degree grid boxes this corresponds to ten further levels of grid boxes) before the search for surface points was abandoned. These conditions were maintained throughout the work to ensure consistency. Figure 2 shows the comparison

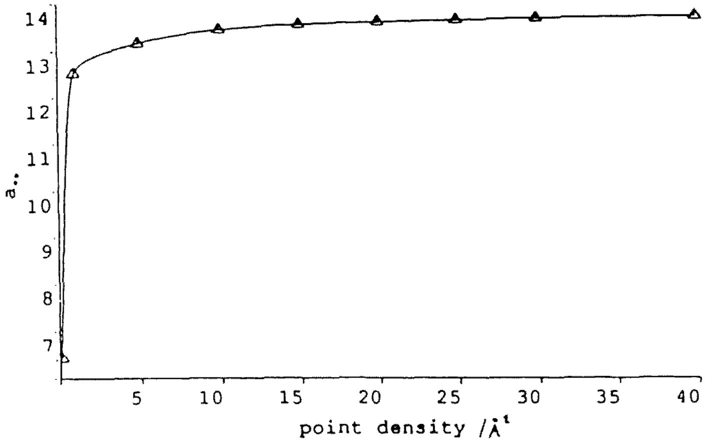


Fig. 1. The effect of point density on computed zeroth order shape descriptor for chymotrypsin.

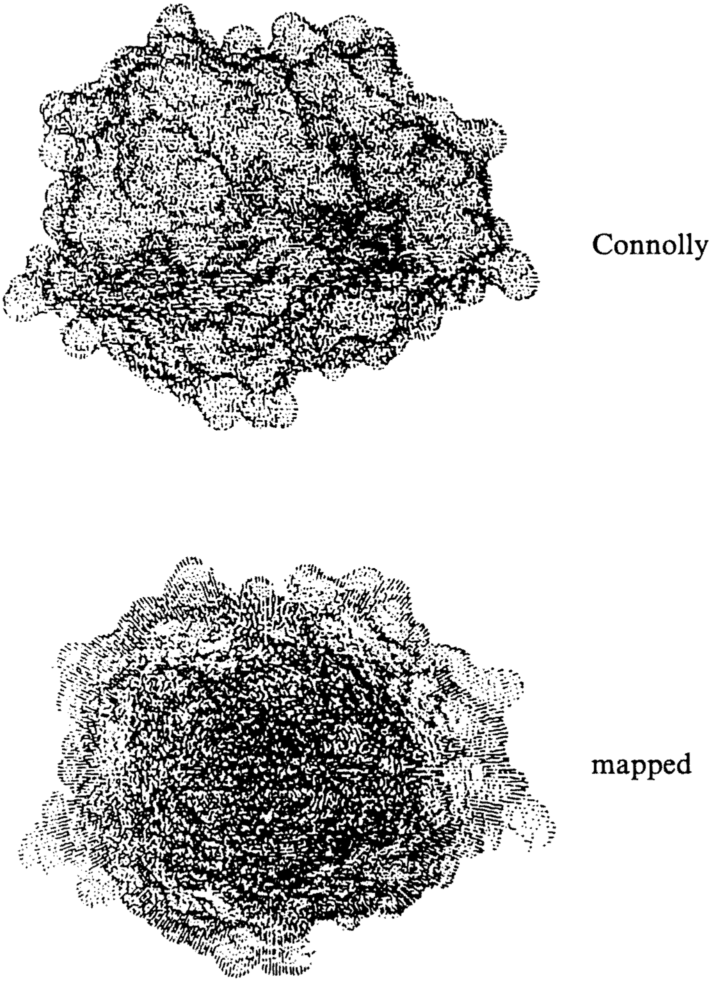


Fig. 2. Comparison of a Connolly dot surface with a mapped polar coordinate version. The molecule is chymotrypsin.

between a dot surface and a mapped surface for chymotrypsin and it can be seen that a rather good representation has been obtained (subsurfaces superimposed).

From the mapped surfaces, harmonic expansion coefficients (the shape descriptors) were calculated. Values were evaluated up to $l = 50$ ($m \leqslant l$) and up to $k = 6$, i.e. for each separate surface function.

2.2. DISPLAYING SHAPE DESCRIPTORS

A number of examples of shape descriptors are shown in fig. 3, these corresponding to the first order subsurface only.

These may be viewed as "shape spectra" the spectra being in some sense a

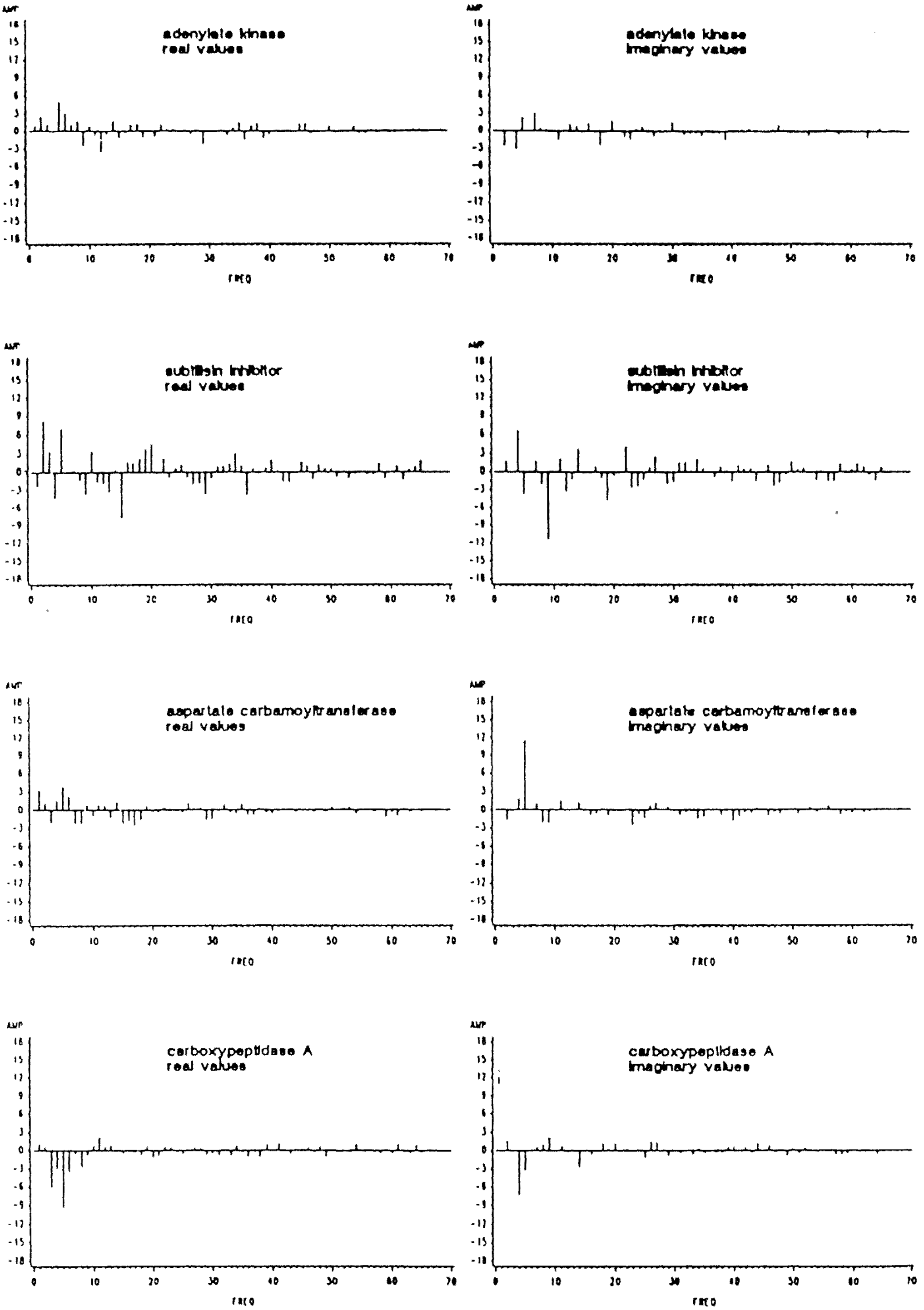


Fig. 3. Shape spectra for a number of protein examples.

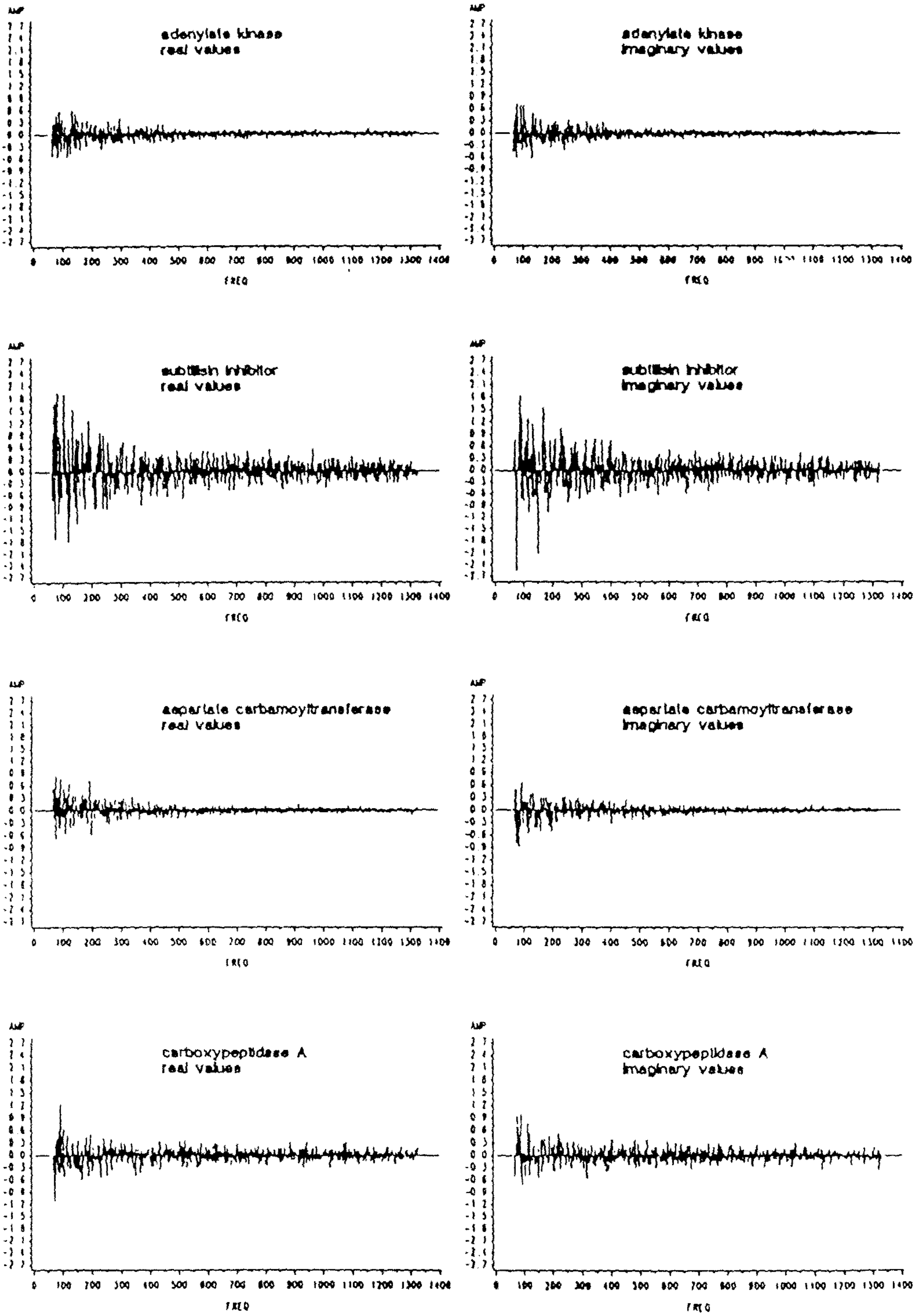


Fig. 3 (continued). Note the change of scale between high and low order shape spectra.

frequency representation of the molecular shape. The term frequency must be qualified since each shape descriptor has two subscripts which have been plotted along a single axis. For each value of l, values for the entire range of m are plotted, before turning to the next value of l. Moving along the x-axis therefore shows the contribution higher order harmonics make to the overall shape and in this sense corresponds to higher frequency terms. It is possible to represent the spectra with l and m along the x- and y-axis respectively say, and with the amplitude along the z-axis but the form given is convenient for a visual inspection of the shape spectra. Note that the descriptors are complex values so that real and an imaginary spectra are obtained in each case.

The spectra as shown are obtained directly from the expansion coefficients and no normalisation has been carried out. In other words, they are raw shape spectra. However, an important observation can be made concerning the descriptors themselves. This is that there is a rapid fall off in the amplitude of the descriptors (there is a change of scale in fig. 3) with much of the shape information apparently being contained within the first few descriptors. This may seem somewhat surprising in view of the complexity of the solvent accessible surface of a protein molecule but actually this rapid early convergence is somewhat curtailed as more descriptors are included. Hence the low order shape descriptors rapidly build up the overall shape (and size) but a great many more are required to represent any fine detail. This is a consequence of the fact that it is a global shape description technique being applied to generally piecewise continuous functions. The zeroth order descriptor is not shown in fig. 3 since this was very much larger than the remaining shape descriptors. This descriptor of course is the one corresponding to size and contains no shape information (i.e. a sphere).

Normalisation was not applied to these descriptors but instead rotationally invariant descriptors were obtained. In fig. 4 are shown plots of invariant descriptors for a number of proteins. In order to have as large a scale as possible the largest value, i.e. the zeroth order descriptor, is not included on the plots since again this was much greater than the remaining descriptors.

In addition, these rotationally invariant descriptors have been size scaled (or size-normalised). This is carried out by dividing all the descriptors of a particular molecule by the zero order descriptor (of the first subsurface). This value corresponds to the size (of the first subsurface) so in effect the descriptor for the first subsurface of each molecule takes the value one. Rotational invariants that result from these size scale descriptors will also be such that in all cases the value $A_0 = 1$ (for the first subsurface). These modified spectra can be used for making a direct visual comparison between the different molecules and again as in the case of the raw descriptors, low order shape descriptors appear to dominate the spectra with a rapid initial tail off in their values. However, it can be seen that beyond values of $l = 20$ there appears to be little variation in the values of the shape descriptors. A second observation that can be made concerns a comparison between molecules. This will be discussed in more detail in subsequent sections but it can be seen that

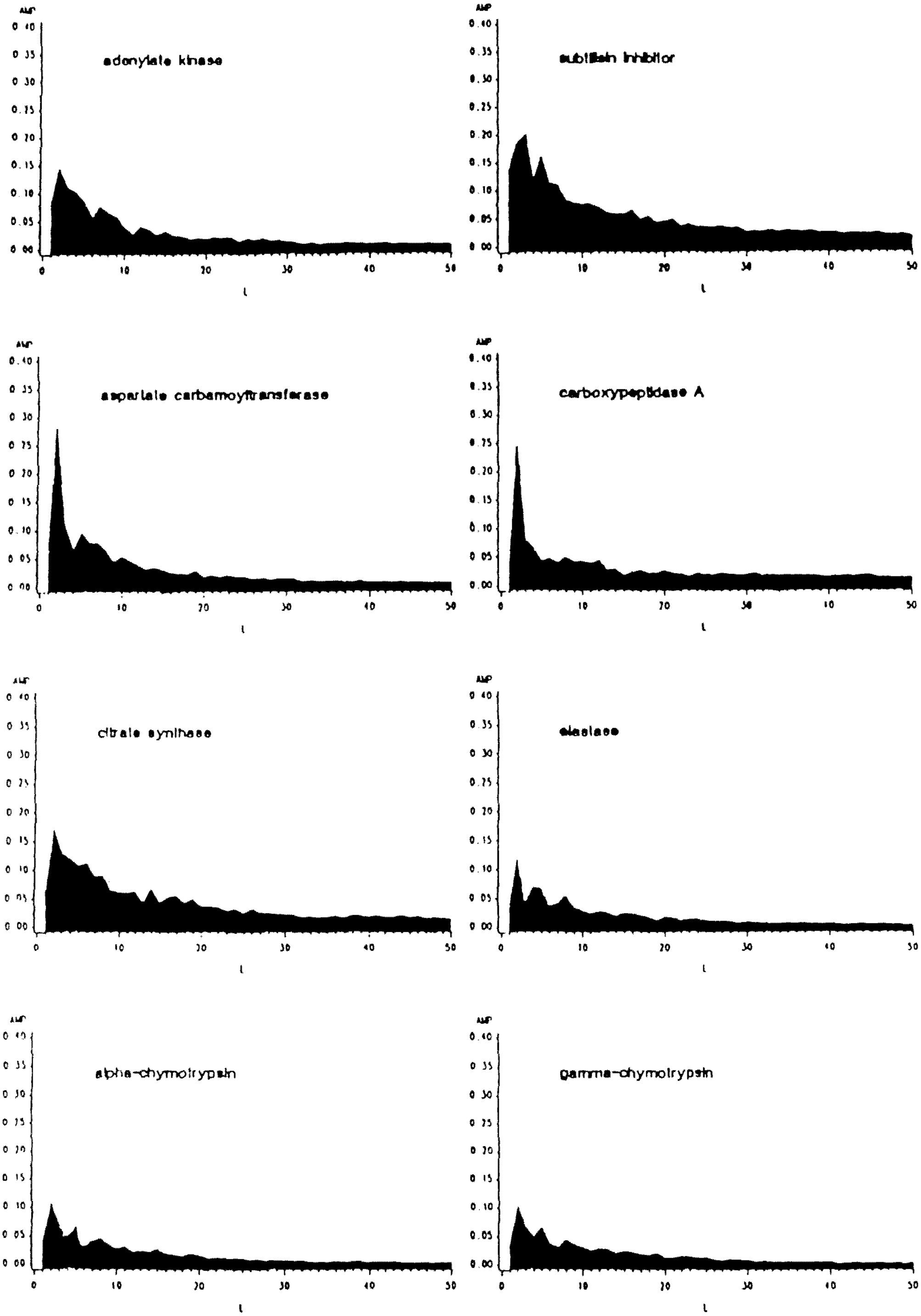


Fig. 4. Size-scaled rotationally invariant shape descriptors of several proteins. These can be used for a direct visual comparison of shape similarities.

molecules that are known to be similar in structure and therefore in shape give a similar shape spectrum. Hence, similarities can be identified between elastase, α-chymotrypsin and γ-chymotrypsin whereas the spectrum of carboxypeptidase A is rather different. Of course, elements of the spectra are also considered as components of Euclidean vectors and the shape difference is quantified through relation

$$D = \left\{ \sum_{k=1}^{K} \sum_{l=0}^{L} (A_l^k - b_l^k)^2 \right\}^{1/2} . \tag{1}$$

The behaviour of this definition will be examined in more detail in the following section.

3. Investigating shape difference properties of the descriptors

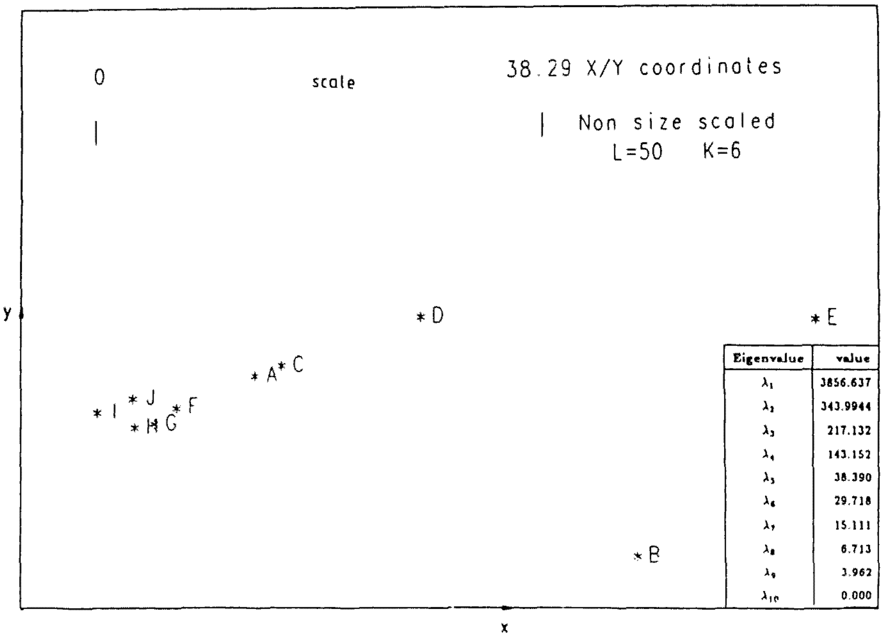
3.1. USING NON-SIZE-SCALED DESCRIPTORS

Initial work was carried out using 10 proteins and shape difference matrices evaluated using eq. (1). Using the method of classical scaling as described in appendix A, shape difference plots have been made and these are shown in fig. 5. It shoúld be emphasised that the motivation for obtaining these plots is to obtain a means of visualising the shape difference measure when applied to a number of molecules.

In these diagrams non-size-scaled rotationally invariant descriptors have been used. In order to interpret these diagrams it is important to observe the eigenvalues obtained in inverting the B matrix. It is seen that in fig. 5 the first eigenvalue is about 10 times greater than the second with the second, third and fourth being of the same order of magnitude. The fifth is then a further order of magnitude smaller. This would indicate that four dimensions are needed to produce a plot with little or no distortion of the Euclidean distance relations as obtained from the shape difference measure and, for example, the cluster in the first plot may be only apparent. However, the second part of fig. 5 shows the spread in values due to the third coordinate. The plots may be therefore viewed as an xy and a zy plot. Thus in the second plot it can be seen that there is little spread above the xy plane and the cluster indicated is real. The values of the eigenvalues indicate that including further coordinates will also have minimal effect on the clustering.

In fact it is possible to quantify the contribution each coordinate makes to the distribution of molecules on the plot (i.e. the shape difference relations). To consider the contribution each coordinate makes or, in particular, the first m coordinates, the following expression can be considered [5]:

$$\sum_{j=1}^{m} \lambda_j / \sum_{j=1}^{p} \lambda_j , \tag{2}$$



Eigenvalue	value
λ_1	3856.637
λ_2	343.9944
λ_3	217.132
λ_4	143.152
λ_5	38.390
λ_6	29.718
λ_7	15.111
λ_8	6.713
λ_9	3.962
λ_{10}	0.000

Fig. 5. Classical scaling plots for 10 proteins (see key). These are obtained from the non-size-scaled descriptors. The *xy* plane is shown.

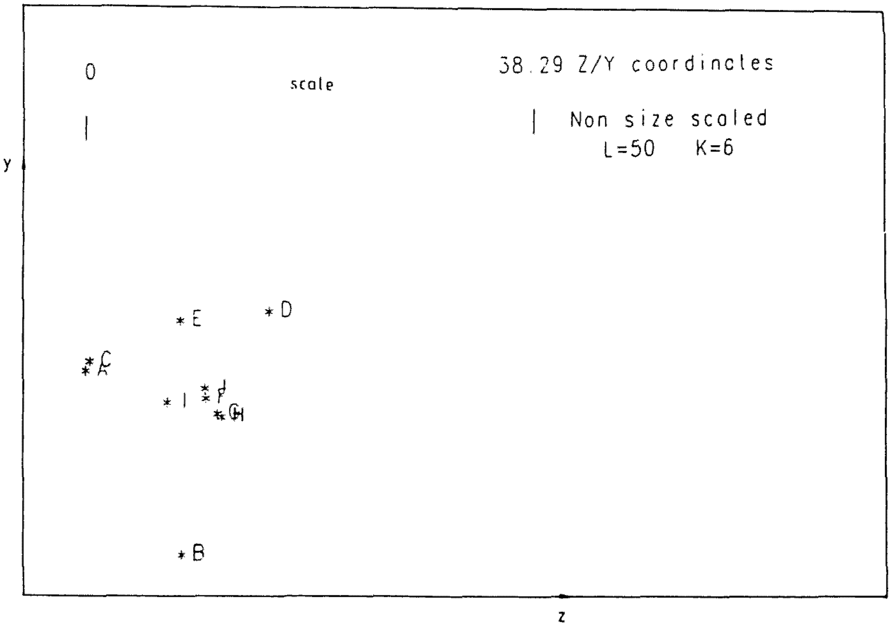
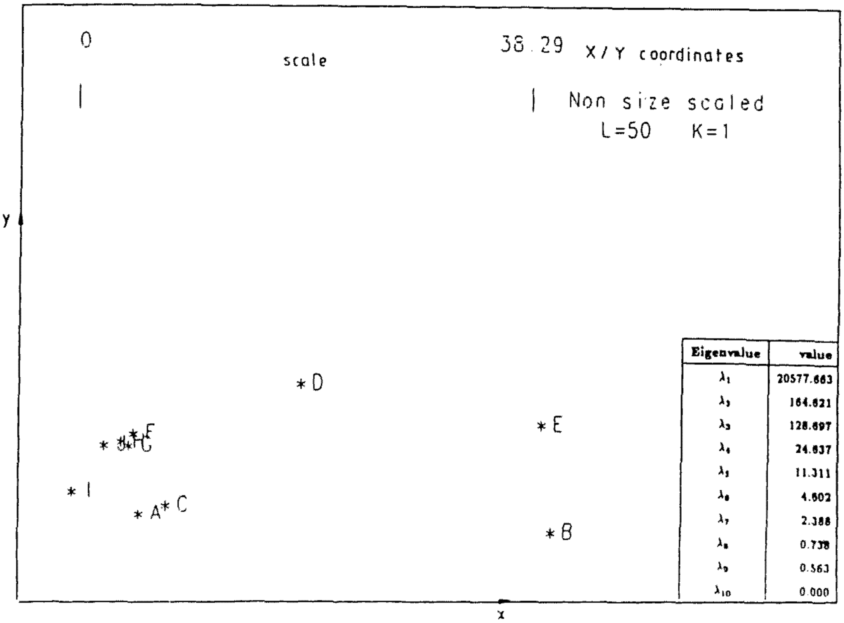


Fig. 5 (continued). This shows the *zy* plane.



Eigenvalue	value
λ_1	20577.663
λ_2	164.621
λ_3	128.697
λ_4	24.637
λ_5	11.311
λ_6	4.602
λ_7	2.388
λ_8	0.738
λ_9	0.563
λ_{10}	0.000

Fig. 6. Scaling plots with the number of subsurfaces reduced to 1. The *xy* plane.

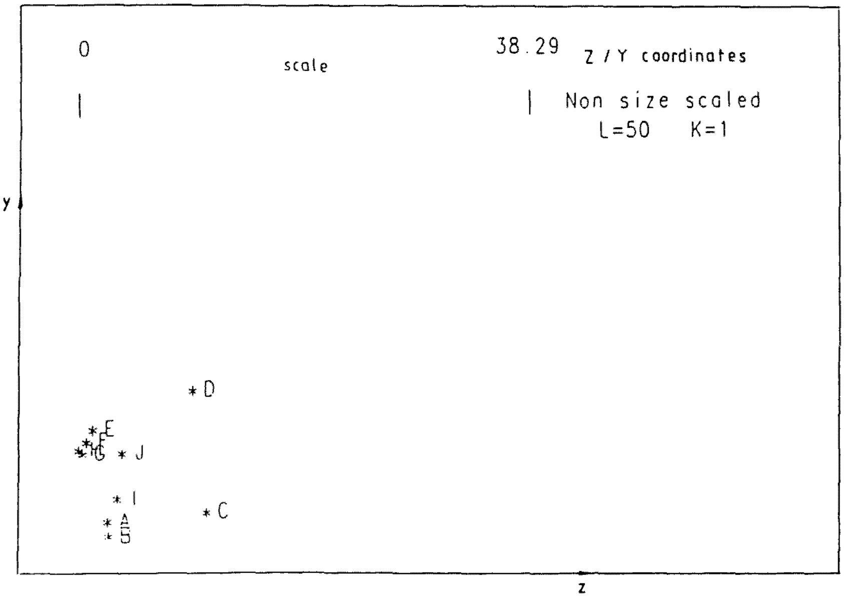


Fig. 6 (continued). The *yz* plane.

where p is the total number of coordinates (i.e. molecules and hence eigenvalues). This gives the fractional contribution that the coordinates make and for the examples of fig. 5 the first two coordinates (x and y) account for 90% of the variation. Including the third component (z), 95% of the variation is accounted for. On this basis it can be seen that in combination the xy and yz plots give an excellent indication of the shape interrelations between the various molecules considered. These plots cannot be used for accurate numerical work but it serves as a convenient way to explore shape interrelationships and to identify possible clustering.

In fig. 6 are shown similar plots but in these cases the number of subsurfaces has been reduced to one – the first subsurface (i.e., the upper limit of k is 1 in eq. (1)).

The eigenvalues also indicate that a three dimensional representation is sufficient to display the shape difference interrelationships with 98% of the variation being accounted for in this case. In the latter situation, reentrants on the surface of the molecules are simply left out of the calculation and it may appear surprising that the overall pattern has not been altered to any marked degree. However, this is simply a reflection of the fact that most of the solvent accessible surface of each protein can be "seen" from the centroid of the protein. It is therefore the shape descriptors corresponding to this first order subsurface (recall that the first order subsurface is that surface which can be seen from the centroid of the relevant molecule) which play the major role in the shape difference values.

This point is further emphasised by the figures shown in table 2. Here the zero order shape descriptors are shown for subsurfaces of the proteins considered. Since the values for the first subsurface are much greater than those for higher order subsurfaces the indication is that the first order surface is of much greater area than the higher order subsurfaces. Therefore it is the former which are important in evaluating the shape difference measure. It can also be seen from this table that a number of molecules may require more than six subsurfaces (i.e. those whose sixth zero order descriptor is not zero) but, from what has just been seen and the fact that these sixth order descriptors are small, including further subsurfaces will not change any of the results.

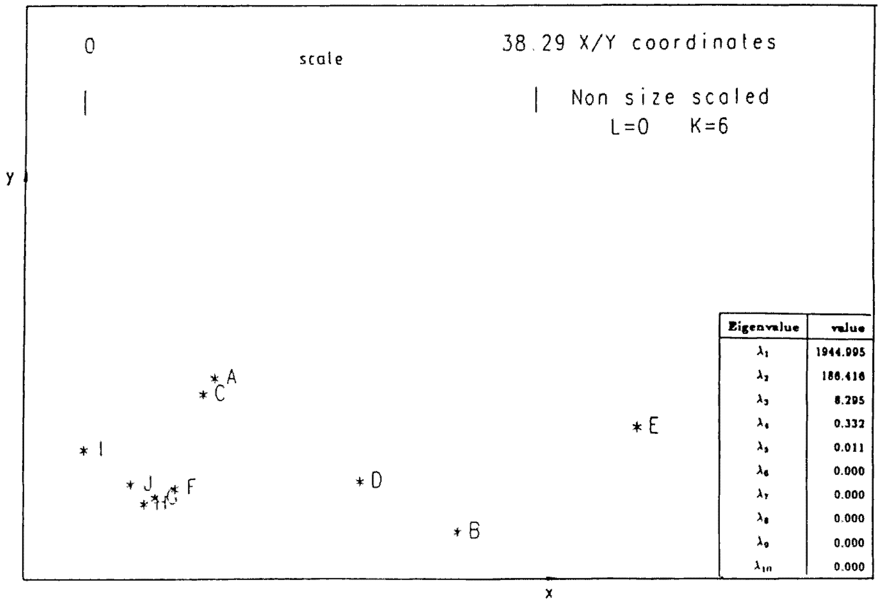
It has been seen that it is the first order subsurface which makes the main contribution to the shape difference relationships. Also to be considered is the effect that a change in the number of shape descriptors (the upper limit L in eq. (1) has on the shape difference measure. It was noted previously that low order descriptors were dominant in the shape spectra and it can be seen from fig. 7 that the major influence in determining shape interrelations between a number of molecules is the zeroth order term. Since there is much similarity between figs. 5 and 7 it must be concluded that it is the size of the molecules which are determining the shape relationships. This is because the zero order descriptor corresponds to the size of the molecule. Note that though the pattern is similar in both cases, the figures will not be to the same scale. However, it is the relative values (distances on the plot, i.e., shape differences) that are of interest.

Table 2
Zero order descriptors for protein molecules. Descriptors have been size normalized relative to the first subsurface.

Protein	Zero order shape descriptor (size-scaled)					
	Subsurface first	second	third	fourth	fifth	sixth
A	1.00000	0.33136	0.08986	0.01252	0.00148	0.00009
B	1.00000	0.28952	0.05803	0.01055	0.00166	0.00030
C	1.00000	0.29918	0.08970	0.01342	0.00235	0.00027
D	1.00000	0.27808	0.09139	0.02241	0.00403	0.00056
E	1.00000	0.43709	0.15005	0.04716	0.01259	0.00305
F	1.00000	0.16922	0.04672	0.00350	0.00030	0.00000
G	1.00000	0.14674	0.03584	0.00306	0.00026	0.00000
H	1.00000	0.13327	0.02919	0.00251	0.00038	0.00000
I	1.00000	0.15306	0.02881	0.00262	0.00013	0.00001
J	1.00000	0.14414	0.03531	0.00281	0.00018	0.00000
K	1.00000	0.17202	0.05173	0.00313	0.00042	0.00000
L	1.00000	0.08391	0.01467	0.00191	0.00052	0.00004
M	1.00000	0.14860	0.03381	0.00197	0.00012	0.00002
N	1.00000	0.06456	0.01218	0.00092	0.00009	0.00000
O	1.00000	0.15593	0.03855	0.00348	0.00033	0.00002
P	1.00000	0.28053	0.10312	0.02193	0.00453	0.00061
Q	1.00000	0.25118	0.07845	0.01398	0.00351	0.00070
R	1.00000	0.12472	0.03128	0.00120	0.00004	0.00000
S	1.00000	0.13592	0.02972	0.00199	0.00003	0.00000
T	1.00000	0.13078	0.02651	0.00241	0.00037	0.00006
U	1.00000	0.42653	0.18444	0.05922	0.01618	0.00339
V	1.00000	0.13267	0.12197	0.00309	0.00059	0.00012
W	1.00000	0.09390	0.01372	0.00053	0.00002	0.00000
X	1.00000	0.15840	0.04752	0.00509	0.00055	0.00002
Y	1.00000	0.15750	0.04465	0.00459	0.00040	0.00002
Z	1.00000	0.14470	0.03305	0.00279	0.00011	0.00000
a	1.00000	0.05138	0.01027	0.00008	0.00000	0.00000
b	1.00000	0.07670	0.01842	0.00108	0.00004	0.00000
c	1.00000	0.14335	0.03904	0.00362	0.00033	0.00000
d	1.00000	0.06035	0.00763	0.00012	0.00000	0.00000

The above points can be seen in much more detail by observing the plots in fig. 8.

In these plots two specific examples are considered, these being the shape difference between trypsin and chymotrypsin and on the same plot the shape difference between trypsin and carboxypeptidase A. It can be seen that already after the $L = 0$ descriptor the main contribution to shape difference has been made. This is also the case if reentrants (i.e. higher order subsurfaces) are included in the calculation.



Eigenvalue	value
λ_1	1944.995
λ_2	188.416
λ_3	8.295
λ_4	0.332
λ_5	0.011
λ_6	0.000
λ_7	0.000
λ_8	0.000
λ_9	0.000
λ_{10}	0.000

Fig. 7. Scaling plots with the number of descriptors reduced to include just the zero order term.

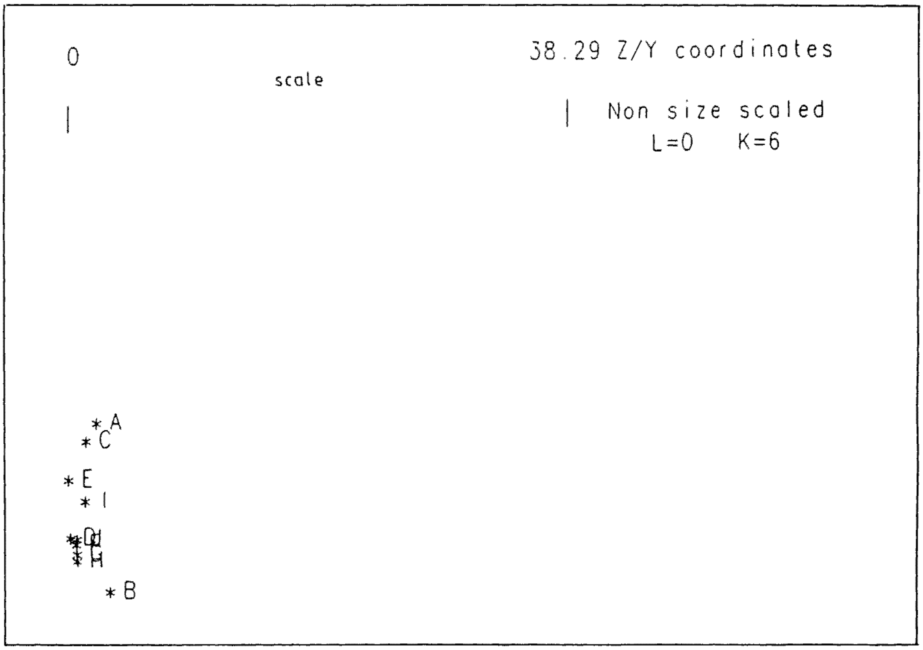


Fig. 7 (continued). The *yz* plane.

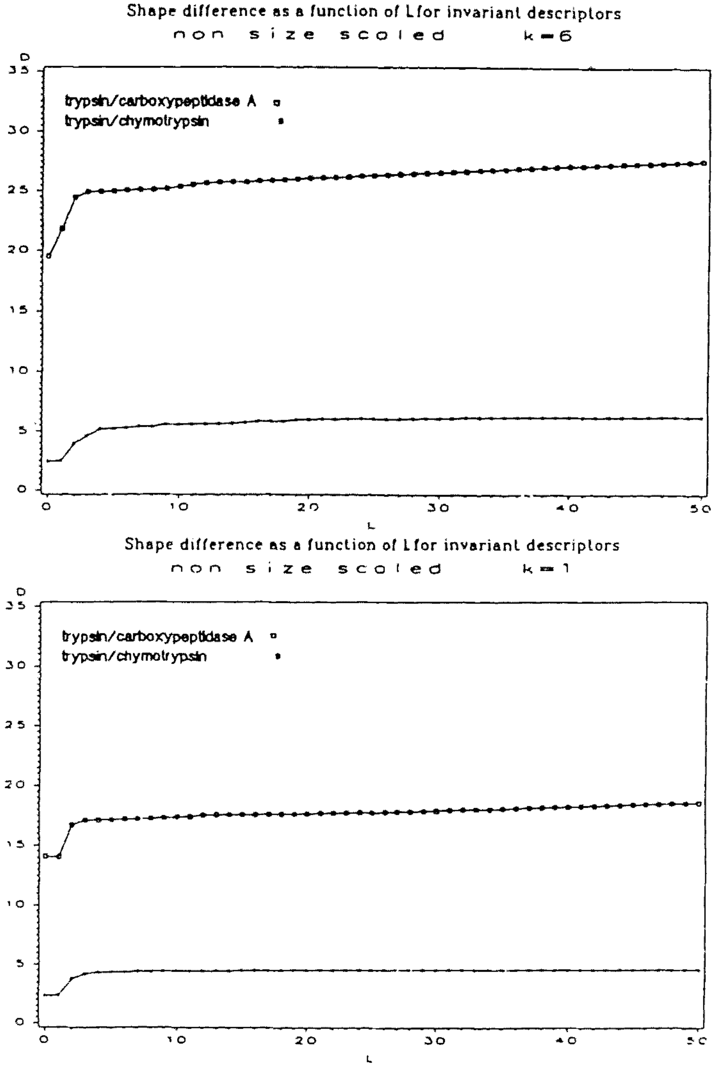
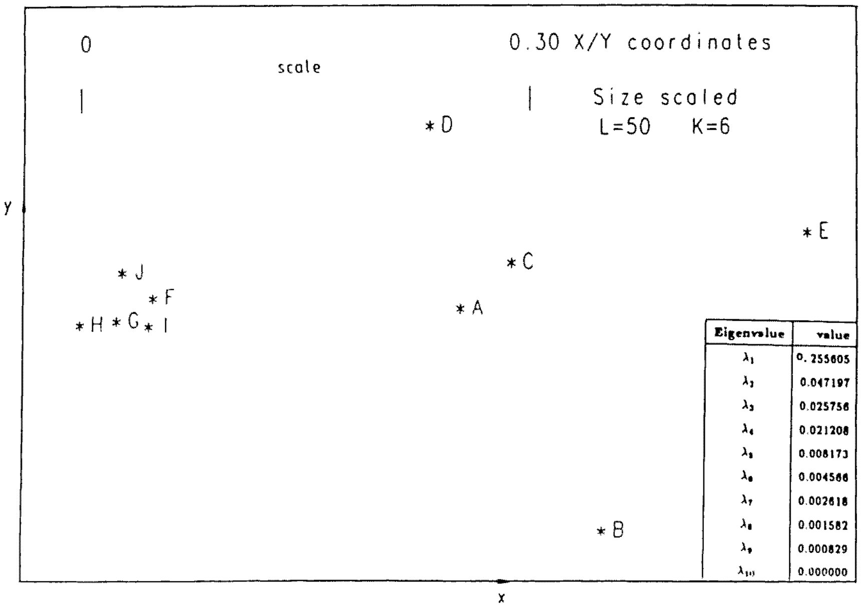


Fig. 8. Shape difference as a function of the number of descriptors.

Similar plots can be obtained for the other proteins and so it is not surprising that the shape difference relations are rather stable under changes in the number of descriptors used and also the number of subsurfaces.

The above results demonstrate that effectively the size of each molecule is the dominant factor and is masking the shape information that is required. Fortunately it is possible to "filter" out the size information by using the size normalised shape descriptors. The results using these are described in the following subsection.



Eigenvalue	value
λ_1	0.255605
λ_2	0.047197
λ_3	0.025756
λ_4	0.021208
λ_5	0.008173
λ_6	0.004566
λ_7	0.002618
λ_8	0.001582
λ_9	0.000829
λ_{10}	0.000000

Fig. 9. Scaling plots obtained using size normalised shape descriptors. The *xy* plane.

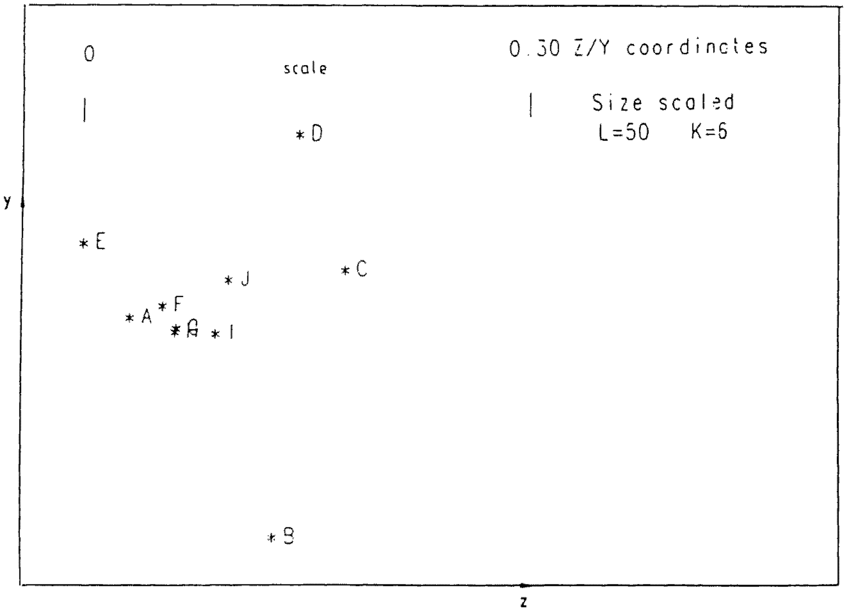


Fig. 9 (continued). The *yz* plane.

3.2. USING SIZE-SCALED DESCRIPTORS

Scaling plots have been obtained using size-scaled descriptors and some of these are shown in fig. 9.

These can be compared with those plots shown in fig. 5 but in the case now being considered size information is excluded. The scale of the plots is considerably reduced since size normalised descriptors are much lower in value than their unnormalised counterparts. Again a similar pattern is observed in the shape relations between the molecules and therefore the relative values of shape difference are not influenced a great deal. It is worth observing that the proteases elastase, α-chymotrypsin, γ-chymotrypsin and trypsin are clustered together. These are all related serine proteases.

Although the dominating influence of molecular size is no longer included it was in fact again found that low order terms were the main contribution to shape interrelations and the effect of including more descriptors and decreasing the number of surfaces used for two examples is shown in fig. 10. An appreciation of the differences in scale between the non-size-scaled plots and the size-scaled plots can be obtained by comparing the shape difference scales of figs. 8 and 10.

Here as in fig. 8 it can also be seen that it is the low order terms which make the dominant contribution to the shape difference measure even though here the effect of molecular size has been "filtered" out.

On the basis of these results it appears that there may be little purpose in evaluating shape descriptors above $l = 10$ if the aim is in exploring shape relationships between molecules.

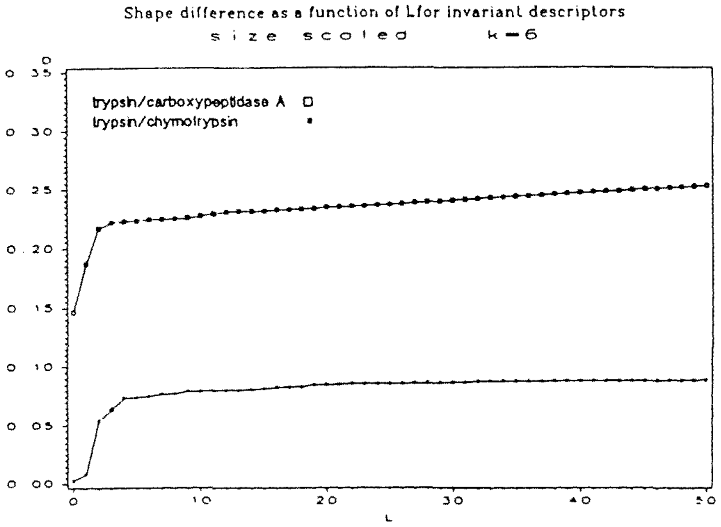


Fig. 10. Shape difference as a function of shape descriptor terms using the size normalised shape descriptors. Again low order terms dominate even though the size contribution is effectively "filtered" out.

4. Shape and groups of related proteins

4.1. GLOBAL AND LOCAL SHAPE COMPARISON

One of the most basic of problems in protein chemistry is that of the relation between amino acid sequence, structure and function. It is only when a proteins amino acid sequence folds into a full three dimensional structure that, for example, an enzyme becomes active. It is not yet possible to predict structure from such a sequence and crystallographic and increasingly NMR techniques are needed to obtain information about a protein structure. Once structures are available, it is then possible to consider the problem of the relation between structure and function of biomolecules. One approach to this problem is to study different proteins which have a similar function with the aim of identifying common structural motifs.

With the work described here the aim is somewhat more modest since the data being considered is that of the solvent accessible surface of a molecule and the shape descriptors do not contain any information as to what lies "behind" the surface. Ultimately though, it is this surface which interacts with the molecule's environment and it is therefore of interest to explore shape relations of the molecular surface between a greater number of protein molecules and to see in what way any shape similarities are reflected in functional similarities.

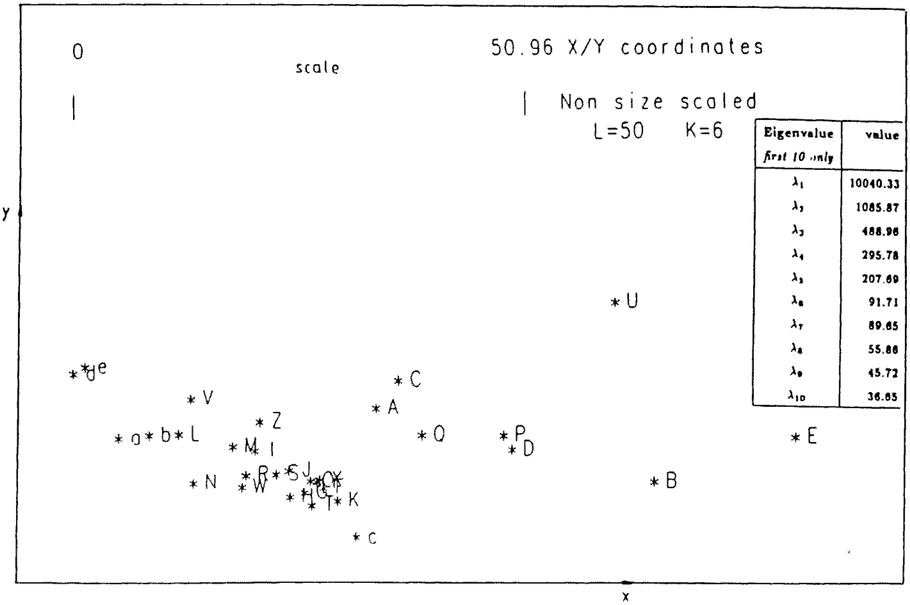
In considering this problem it is important to recall that, with the techniques described here, entire molecules are being compared. In addition, the method is a global shape description technique so no local shape comparisons can be made. However, the active site or sites on a protein are local regions formed by a few key residues, so strictly it would be desirable to compare shapes of these local regions in order to identify possible shape similarities and functional similarities. The local and global aspects of the problem therefore impose a limitation. Proteins retaining common structural characteristics during a process of divergent evolution would still be expected to show similarities in overall shape though, and on this basis global shape comparisons can be made.

4.2. NON-SIZE-SCALED AND SIZE-SCALED PROTEIN SHAPE COMPARISONS

Bearing in mind the limitations outlined in the previous subsection, scaling plots for a large number of proteins are shown in fig. 11 and fig. 12.

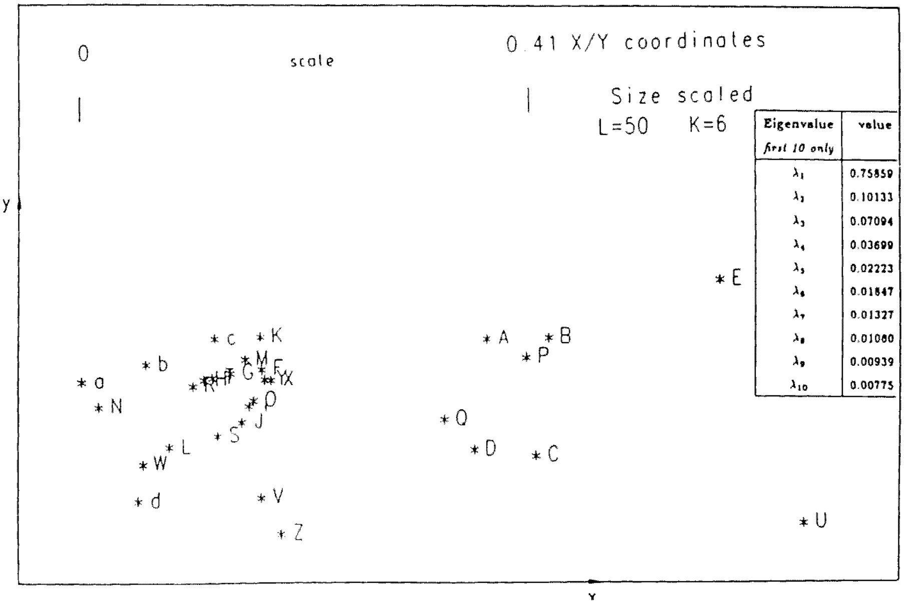
The former shows a plot obtained from the non-size-scaled descriptors and from what has been seen previously it may be supposed that these results are a reflection of relative sizes of the molecules.

However, all the serine proteases are grouped in the main cluster and this appears also, though not to the same extent, in fig. 12 (these proteases have codes F, G, H, J, K, N, O, R, X, Y). This has been obtained using the size-scaled descriptors and therefore is a pure shape comparison. Clustering of structurally related



Eigenvalue *first 10 only*	value
λ_1	10040.33
λ_2	1085.87
λ_3	488.96
λ_4	295.78
λ_5	207.69
λ_6	91.71
λ_7	89.65
λ_8	55.86
λ_9	45.72
λ_{10}	36.85

Fig. 11. Non-size-scaled shape comparison between a large number of proteins.



Eigenvalue *first 10 only*	value
λ_1	0.75859
λ_2	0.10133
λ_3	0.07094
λ_4	0.03699
λ_5	0.02223
λ_6	0.01847
λ_7	0.01327
λ_8	0.01080
λ_9	0.00939
λ_{10}	0.00775

Fig. 12. Size-scaled shape comparison between a large number of proteins.

serine proteases is certainly expected – the enzymes of trypsin, chymotrypsin and elastase have both similar amino acid sequences and similar crystallographic structures. In the case of the size-scaled plots, the eigenvalues indicate that the first two coordinates contain 80% of the shape difference variation with the value being 88% in the non-size-scaled case. Hence, the grouping on these plots is significant (i.e. not a consequence of representing the data in too few dimensions) and it is therefore somewhat surprising that as well as the structurally related serine proteases just mentioned a number of other molecules appear in the same group. This applies more to the size normalised plot where, in particular, there are a number of other serine proteases included as well as the thiol protease papain, carbonic anhydrase B, staphylococcal nuclease and dihydrofolate reductase.

Also note that here there is carboxypeptidase A (labelled C) and a second carboxypeptidase A (labelled D) at a relatively large distance from the former. The key in table 1 can be referred to for details of other proteins. The difference between these two is that the former is the potato inhibitor structure whilst the latter is the unbound bovine structure. The inhibitor was removed from the PDB file before a Connolly surface was obtained. Although it is known that carboxypeptidase A undergoes large structural changes on binding [6], it would be expected that there would be much similarity in shape, an expectation which is at variance with the indications of the scaling plots.

It is not clear why a number of non-related enzymes should appear to be closely related in shape though it seems likely that this is a symptom of using the rotationally invariant descriptors which do not contain any phase information. More exactly, each descriptor of order l is a measure of the total contribution made by the set of spherical harmonics of order l in building the overall shape of the molecule in hand. This will have the consequence that similar shapes will result in similar shape descriptors but different shapes may produce similarities in the rotationally invariant descriptors. In other words, there may be a degeneracy problem in the rotational invariants.

It is expected that using such shape descriptors as described may involve additional complications so that the technique may better be considered as a first step in exploring shape relations between numbers of molecules. From this point of view largely dissimilar molecules may be eliminated from any further consideration, whilst those displaying certain similarities may be examined more closely by using the full set of original raw descriptors. This implies a large number of expensive normalisations but nevertheless a considerable reduction in the number that might otherwise be needed.

To see this, consider a set of thirty molecules which are formed into two clusters of fifteen molecules using the above techniques. Further analysis of the two groups would require $2 \times (15 \times 14)/2 = 210$ normalisations. However, the number of normalisations required in a full analysis of the original ungrouped set of thirty molecules would be $30 \times 29/2 = 435$. Hence, in such an idealised situation the number of normalisations would be reduced by around a factor of two.

5. Conclusion

It has been seen how shape descriptors are obtained for protein molecules and how they may be used in exploring shape relations between various molecules. One of the problems appears to be the lack of sensitivity so that, in particular, low order shape descriptors dominate the Euclidean shape difference measure and it is these which represent gross shape features of the molecules. Hence, it is comparisons of a low resolution view of the molecules that are being made using the classical scaling.

It has also been pointed out that it is the rotational invariants that have been used to define a shape difference measure so again information concerning the full detailed shape of the molecule is missing from the analysis. It seems likely that a combination of these two considerations could account for some of the peculiarities displayed in the scaling plots. In order to assess this, calculations would have to be carried out using the original "raw" descriptors, requiring a large number of normalisations. However, the technique could be used as a first step by identifying structures which display large differences in shape and then excluding these from any normalisation calculations.

Appendix A

CLASSICAL SCALING

It is an easy task to calculate Euclidean distances between pairs of points either directly from the data matrix X (coordinates of points) in which case the distance matrix is given by

$$d_{rs}^2 = \sum_{j=1}^{p}(x_{rj} - x_{sj})^2 \tag{3}$$

or, where n is the number of coordinates each point has (i.e., the dimensionality), using the $n \times n$ matrix $B = X^{\mathrm{T}}X$ which is the matrix of scalar products between each pair of points. In the above, the matrix X has general elements x_{ij}. Hence, each term of B is given by

$$b_{ij} = \sum_{k=1}^{p} x_{ik}x_{jk} \,. \tag{4}$$

The distance between points r and s is therefore given by

$$d_{rs}^2 = \sum_{k=1}^{p} x_{rk}x_{rk} + x_{sk}x_{sk} - 2x_{rk}x_{sk} \tag{5}$$

$$= b_{rr} + b_{ss} - 2b_{rs} \,. \tag{6}$$

The problem to be solved however is the inverse of above. The matrix of distance values is given and it is required to determine the coordinates. The problem can be approached in two stages. First the matrix B is obtained and then it is factorised in the form $B = X^{\mathrm{T}}X$. To obtain B, eq. (6) needs to be inverted, which has no unique solution. However, with the constraint that the data is mean centered (i.e. the "centre of gravity" of x lies at the origin, or mathematically $\bar{x} = 0$, where x refers to either column or row vectors of X) eq. (6) can be summed over r, over s and over r and s. This gives the following:

$$\sum_r d_{rs}^2 = T + nb_{ss}\,, \tag{7}$$

$$\sum_s d_{rs}^2 = nb_{rr} + T\,, \tag{8}$$

$$\sum_{r,s} d_{rs}^2 = 2nT \tag{9}$$

with the trace

$$T = \sum_{r=1}^{n} b_{rr}\,. \tag{10}$$

Equations (7) to (10) can be rearranged to give

$$b_{rr} = \frac{1}{n}\sum_s d_{rs}^2 - \frac{T}{n}\,, \tag{11}$$

$$b_{ss} = \frac{1}{n}\sum_r d_{rs}^2 - \frac{T}{n}\,, \tag{12}$$

$$2\frac{T}{n} = \frac{1}{n^2}\sum_{r,s} d_{rs}^2\,. \tag{13}$$

Indicating the averages expressed in the above equations by using a dot the solution for B is given by

$$b_{rs} = -\tfrac{1}{2}[d_{rs}^2 - d_{r\cdot}^2 - d_{\cdot s}^2 + d_{\cdot\cdot}^2]\,. \tag{14}$$

Having obtained the matrix B a possible coordinate matrix can be obtained by carrying out an eigenvector analysis of B. If there are k eigenvalues arranged such that $\lambda_1 \leqslant \lambda_2 \leqslant \ldots \leqslant \lambda_k$ and k eigenvectors $\boldsymbol{e}_1, \boldsymbol{e}_2 \ldots \boldsymbol{e}_k$ of unit length then the coordinate matrix is given by [5]

$$X = (\boldsymbol{f}_1\boldsymbol{f}_2 \cdots \boldsymbol{f}_k)\,, \tag{15}$$

where $\boldsymbol{f}_i = (\sqrt{\lambda_i})\boldsymbol{e}_i$.

A program has been written to carry out the above analysis with the eigenvalues and eigenvectors being calculated by the method of Jacobi [7]. It may in fact require more than two or three dimensions to represent the data, this being indicated by the number of non zero eigenvalues. If it is the case that the two highest valued eigenvalues are much greater than the remaining ones then using the corresponding two eigenvectors to obtain a two dimensional coordinate will provide a good representation of the data. Similarly for three or more dimensions, but clearly if there were four large eigenvalues then it would not be possible to obtain an accurate visual representation. Results shown in the following chapters however clearly show that for the distance matrices obtained for molecules, a two or three dimensional representation can generally be used to display shape difference interrelationships as contained in the distance matrix.

References

[1] S.E. Leicester, J.L. Finney and R.P. Bywater, J. Mol. Graph. 6 (1988) 104.

[2] S.E. Leicester, J.L. Finney and R.P. Bywater, J. Math. Chem. 16 (1994) 315.

[3] F.C. Bernstein, T.F. Koetzle, G.J.B. Williams, E.F. Meyer, M.D. Bryce, J.R. Rodgers, O. Kennard, T. Shikanouchi and M. Tasumi, J. Mol. Biol. 112 (1977) 535.

[4] M.L. Connolly, Molecular Surface Program, QCPE Bull. 1 (1981) 74 (QCPE No. 429).

[5] C. Chatfield and A.J. Collins, *Introduction to Multivariate Analysis* (Chapman & Hall, London, 1980).

[6] L. Stryer, *Biochemistry* (Freeman, New York, 1988).

[7] W.H. Press, B.P. Flannery, S.A. Teukolsky and W.T. Vetterling, *Numerical Recipes – The Art of Scientific Programming* (Cambridge University Press, 1987).