A quantitative representation of molecular surface shape. I: Theory and development of the method

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, DK-2880 Bagsværd, Denmark*

Received 20 December 1993

The importance of molecular shape in many areas of biochemistry and biomolecular interactions is well recognised. In spite of this a rigorous and widely applicable means of defining and quantifying molecular shape has not been available. This paper, the first of a series of papers, presents a new method of quantifying molecular "surface" shape. The development of the technique, based on Fourier shape descriptors is discussed in some depth including the computer programs written to implement and test the method. A subsequent paper will present results obtained from the application of the new quantitative molecular shape descriptors.

1. Introduction

Considerations of molecular shape play a very important role in many areas of biochemistry and medicinal chemistry, especially when one is concerned with questions of molecular similarity or recognition of one molecule by another.

Attempts to represent molecular shape in a quantitative way are rendered difficult by the complexity and irregularity of molecular surfaces and the lack of symmetry in all but the simplest molecules [1]. Most authors have been content to use parameters like molecular surface area or volume, that express size but not shape, or else some approximate method such as STERIMOL [2] in which much of the shape information is lost.

In order to make an accurate representation of the molecule, and comparisons with other molecules it is essential to encode as much of the shape information as possible. A number of interesting mathematical methods have been developed for tackling this problem, some based on the idea of homology groups [1,3] of algebraic

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

topology. We have chosen another approach, based on the idea of expressing shape in terms of Fourier descriptors [4,5]. A similar approach was adopted by Max and Getzoff [6].

The technique is novel in the field of molecular sciences, but in other fields a number of applications have already been described in the literature [7–9]. Although the majority of methods have been restricted to the two dimensional domain, techniques have been described for dealing with three dimensions but with the limitation of being unable to deal with reentrant surfaces [10]. This severely limits the usefulness of these methods.

In this paper, the first of a series of papers, the Fourier descriptor method is further developed to enable it to be applicable to molecular surface shape including those with reentrants.

A brief resume of the Fourier shape descriptor method will be given first, followed by a description of how to extend the method to deal with three dimensional surface shape.

2. The method of Fourier descriptors

The method of representing shape information using Fourier descriptors was first suggested by Cosgriff [11] and the basic principle is straightforward. In its simplest form, a closed contour, i.e. the shape to be described, is represented as a parametric function as shown in fig. 1.

As the parameter ϕ increases, the shape is generated and, since the contour is closed, continually increasing the parameter implies repeated tracing-out of the contour. Hence there results in effect a periodic function of the parameter ϕ which can therefore be represented in terms of a Fourier series as follows:

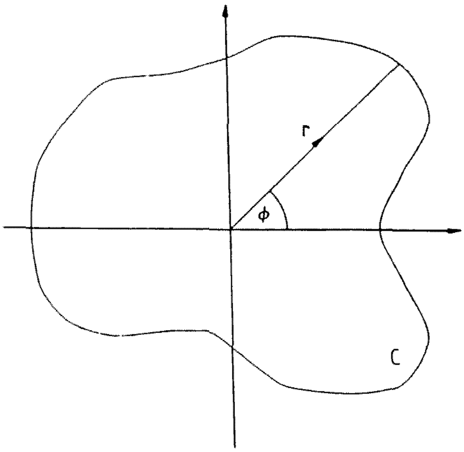

Fig. 1. The shape of a contour C, is represented as a parametric function $r(\phi)$.

$$r(\phi) = \sum_{n=-\infty}^{+\infty} a_n e^{i\frac{2\pi}{T}n\phi}\,, \tag{1}$$

where T is the periodicity of the function, in this case 2π. The expansion coefficients of the series, given by

$$a_n = \int_0^T r(\phi) e^{-i\frac{2\pi}{T}n\phi}\,d\phi\,, \tag{2}$$

are known as Fourier descriptors and it is these that contain the shape information of the contour.

From this basic idea the technique has been developed by a number of researchers in order to tackle such problems as aircraft recognition [7], automatic machine part recognition [8] and for character recognition [9]. In the form described above there is, however, a serious difficulty which occurs if a contour has a reentrant region. This may result (depending upon the severity of the reentrant) in the parametric function being multivalued over some range of the parameter (see fig. 2).

It is of course possible to represent only a single valued function as a Fourier series.

In order to overcome this problem other methods of parameterising the contour have been used. For example, the parameter may be considered to be elapsed time as the contour is traced out at uniform speed [7] or, alternatively, a function can be defined which measures the angular direction of the curve as this varies with arc length [12].

Within these approaches it then becomes possible to describe very general shapes using the Fourier descriptor technique. Unfortunately, as will be seen later, the problem of reentrants will recur in the development of Fourier descriptors for molecular surfaces and this will be somewhat less straightforward to overcome.

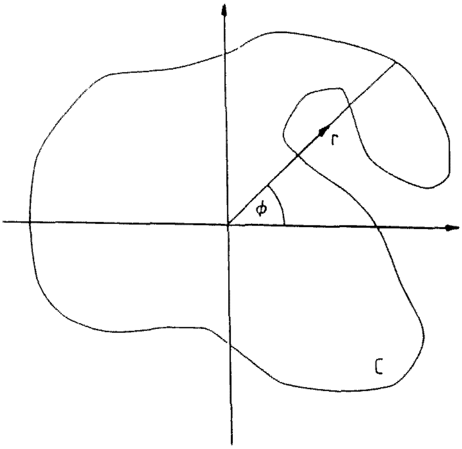

Fig. 2. A contour described by a function of the polar angle ϕ may result in the function being multivalued over one or more ranges of this parameter.

In order to compare quantitatively two shapes, Fourier descriptors for each shape are calculated which are then considered as a pair of N dimensional vectors $\vec{\boldsymbol{A}} = (a_1\ a_2\ \ldots\ a_N)$, $\vec{\boldsymbol{B}} = (b_1\ b_2\ \ldots\ b_N)$, where N equals the number of descriptors obtained for each shape. From this perspective an obvious quantitative measure of shape difference is the Euclidean distance between end points of the two vectors given by

$$D = |\vec{\boldsymbol{A}} - \vec{\boldsymbol{B}}| = \left\{ \sum_{n=N}^{+N} (a_n - b_n)(a_n - b_n)^* \right\}^{1/2} . \tag{3}$$

The measure of course depends on the number of descriptors used (N in the above equation) and so should be consistent throughout any set of calculations. The number of descriptors required will in general depend on the complexity of the shape or the accuracy of representation desired (resolution) and this is one point for investigation concerning the molecular descriptors developed and described in this series of papers.

Apart from the number of descriptors used, i.e., the dimensionality of the shape vector, of more fundamental importance is the dependence of the shape difference measure on how the coordinate system is originally set up. This is due to the fact that, for example, a rotation of the contour (or coordinate system) will affect the values obtained for the Fourier descriptors. To overcome this, a procedure referred to as normalisation must be carried out on the descriptors before a comparison can be made and there are two basic approaches to this. The first is to consider positioning the shapes in some standard orientation before a calculation of shape difference is made [7] and the second is to minimise the distance measure with respect to coordinate operations. It is possible to consider the minimisation process in real space since it follows from Parseval's identity [13] that the distance measure defined above has a representation (strictly in the limit of infinite N) in real space given by

$$D = \frac{1}{\sqrt{2\pi}} \left\{ \int_0^{2\pi} |r_A(\phi) - r_B(\phi)|^2 \, \mathrm{d}\phi \right\}^{1/2} . \tag{4}$$

From the above, normalisation can be seen to be a process of rotating and translating one object with respect to the second such that a best fit according to the minimum of eq. (4) is obtained. One advantage of the Fourier descriptors is that these coordinate operations have a straightforward representation in Fourier space or, in other words, a simple effect on the descriptors. Also note that normalisation which involves orienting the shapes to some standard position is faster than minimising the distance measure but is also suboptimum [14]. In this case only the two largest descriptors (and a third to resolve ambiguities) are used in the normalisation procedure. Since it is often the case that most of the "energy" of Fourier descriptors is contained within as few as two descriptors [15], it can be considered as effectively a normalisation with respect to gross shape features, i.e., those features which contribute the most to the shape difference measure. Actually it is also possible to use the Fourier descriptors to construct a second set of shape descriptors which are invariant under rotations and this approach has also been used in shape recognition [16]. Normalisation and particularly the construction of rotationally invariant descriptors will have important bearing on the development of molecular shape descriptors.

2.1. FOURIER DESCRIPTORS IN THREE DIMENSIONS

The method of Fourier descriptors for quantitative shape comparison is well developed and has been very successful. It is also based on the powerful foundation of Fourier analysis and seems to offer a promising approach to developing molecular shape descriptors. One point to note is that the use of descriptors has in general been confined to describing two dimensional contours and this is where most of the work has been carried out. The reason for this is that the method was developed in the field of image processing and has therefore concentrated on flat, i.e. two dimensional, images.

To obtain three dimensional Fourier descriptors, the shape, now a surface, expressed as a function of two parameters, is required. Considering the two parameters u and v, this function may then be expanded as a product of two Fourier series as follows:

$$f(u, v) = \sum_{m} \sum_{n} a_m b_n \mathrm{e}^{inu} \mathrm{e}^{imv} . \tag{5}$$

Some work has in fact been done in using Fourier descriptors to characterise shape in three dimensional [10,17] but in these methods a Cartesian coordinate system is used and one of the parameters used for describing the surface is z, the height above the xy plane. Hence they are based on dividing an object into sections parallel to the xy plane and then applying to each section of the object the straightforward two dimensional technique described previously.

There are two main difficulties in extending the two dimensional method to characterising the shape of three dimensional surfaces. The first problem is that of obtaining a parametric description of the surface which results in a single valued function. With a cross sectional method as just described this can be overcome in the same way as in the two dimensional situation. Hence one parameter is z and the second for example is the time parameter when travelling around each sectional contour at uniform speed. In this way, a single valued function of two parameters may be obtained which can then be expanded as the series given by eq. (5). The second difficulty concerns normalisation and the development of rotationally invariant shape descriptors. To normalise the distance measure the effect of rigid body coordinate operations on the Fourier descriptors must be considered and descriptors are therefore required on which these operations have a straightforward effect. In particular this concerns the effect of three dimensional rotations on the shape descriptors, and in the work dealing with three dimensional shape description just mentioned the shape descriptors have no straightforward dependence on general rotations. This is clear because of the sectioning approach taken; only rotations about the z-axis will result in a simple modification of the Fourier descriptors – a reflection of the fact that they are, in the final analysis, inherently two dimensional methods.

The problem is therefore one of parameterising a surface such that a single valued function is obtained and finding a suitable set of functions in which to expand the parametric function such that rigid body coordinate operations have a straightforward representation. In the latter case spherical harmonics would seem to be most appropriate, though in fact rotations still have a somewhat complicated effect on the expansion coefficients. However, with this choice of basis functions it becomes a straightforward matter to obtain rotationally invariant descriptors from the original expansion coefficients and this will be discussed in section 8. The idea of exploiting the rotational properties of spherical harmonics is not new, and has been used in developing the fast rotation function used in making a rotational search for examining Patterson maps derived from diffraction data [18].

Choosing spherical harmonics as basis functions implies using polar coordinates as parameters of the surface and therefore similar problems are expected to arise as in the two dimensional case when reentrant regions are encountered. In spite of this, it was decided to proceed within this approach and seek some other method of overcoming the problems of reentrants and indeed a way of avoiding this difficulty will be described later. In the following section, an account of spherical harmonics will be given and in particular those properties which are relevant to the shape descriptor method.

3. Spherical harmonics

The traditional way of considering spherical harmonics is as solutions to the Laplace equation

$$\nabla^2 f(x,y,z) = 0\,; \tag{6}$$

this approach will be described here. A more detailed discussion can be found in Hobson [19]. To obtain solutions to the Laplace equation the Laplacian operator is written in polar coordinates and the solution $f(r,\theta,\phi)$ (a different function to that expressed in equation (6)) is supposed separable such that $f(r, \theta, \phi) = R(r)\Theta(\theta)\Phi(\phi)$. This results in the following equation:

$$\left\{r^2\left(\frac{\partial^2}{\partial r^2}+\frac{2}{r}\frac{\partial}{\partial r}\right)+\left[\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right)+\frac{1}{\sin^2\theta}\frac{\partial^2}{\partial\phi^2}\right]\right\}R(r)\Theta(\theta)\Phi(\phi)=0\,. \tag{7}$$

Since the first term depends on r only and the second two terms depend on θ and ϕ

only, the only way they can be equal is if they are constant. For convenience this constant is written in the form $l(l+1)$ to give the following two equations:

$$r^2\left(\frac{\mathrm{d}^2}{\mathrm{d}r^2}+\frac{2}{r}\frac{\mathrm{d}}{\mathrm{d}r}\right)R(r)=l(l+1)R(r)\,, \tag{8}$$

$$\left[\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right)+\frac{1}{\sin^2\theta}\frac{\partial^2}{\partial\phi^2}+l(l+1)\right]\Theta(\theta)\Phi(\phi)=0\,. \tag{9}$$

Considering the case of the angular dependent equation this can again be separated into ϕ and θ dependent terms and equating these terms to a constant $-m^2$ there result two differential equations:

$$\frac{\mathrm{d}^2}{\mathrm{d}\phi^2}\Phi(\phi)=-m^2\Phi(\phi)\,, \tag{10}$$

$$\left[\sin\theta\frac{\mathrm{d}}{\mathrm{d}\theta}\left(\sin\theta\frac{\mathrm{d}}{\mathrm{d}\theta}\right)+l(l+1)\sin^2\theta-m^2\right]\Theta(\theta)=0\,. \tag{11}$$

The solution to the former is straightforward and is given by

$$\Phi(\phi)=\mathrm{e}^{im\phi}\,, \tag{12}$$

where the integration constants have been set to 0 (the phase) and 1 (the amplitude). In the above, the restriction on m to integer values ensures continuity. The second of the above equations, (11), is Legendre's equation, the general solutions of which are the associated Legendre polynomials $P_l^m(\theta)$ [19]. With the restrictions $l\geqslant 0$ and $|m|\leqslant l$ the solutions to the angular part of the Laplacian are the spherical harmonics which in complex form are given by

$$Y_l^m(\theta,\phi)=P_l^m(\theta)\mathrm{e}^{im\phi}\,. \tag{13}$$

The spherical harmonics are orthogonal over a unit sphere so that with a suitable normalisation

$$\int_0^{2\pi}\mathrm{d}\phi\int_0^{\pi}\mathrm{d}\theta\sin\theta\, Y_l^m(\theta,\phi)(Y_{l'}^{m'})^*(\theta,\phi)=\delta_{mm'}\delta_{ll'}\,. \tag{14}$$

Using the orthogonality property a function defined over a sphere $r(\theta,\phi)$ may be formally expanded in terms of spherical harmonics as follows:

$$r(\theta,\phi)=\sum_{l=0}^{\infty}\sum_{m=-l}^{+l}a_{lm}Y_l^m(\theta,\phi)\,. \tag{15}$$

Multiplying through by $(Y_{l'}^{m'})^*(\theta,\phi)$ gives

$$r(\theta,\phi)(Y_{l'}^{m'})*(\theta,\phi)=\sum_{l=0}^{\infty}\sum_{m=-l}^{+l}a_{lm}Y_l^m(\theta,\phi)(Y_{l'}^{m'})^*(\theta,\phi) \tag{16}$$

and integrating over a sphere

$$\int_0^{2\pi}\int_0^{\pi} \mathrm{d}\Omega\, r(\theta,\phi)(Y_{l'}^{m'})^*(\theta,\phi)$$
$$= \sum_{l=0}^{\infty}\sum_{m=-l}^{+l} a_{lm}\int_0^{2\pi}\int_0^{\pi} \mathrm{d}\Omega\, Y_l^m(\theta,\phi)(Y_{l'}^{m'})^*(\theta,\phi)\,. \tag{17}$$

Hence from the orthogonality of the spherical harmonics

$$\int_0^{2\pi}\int_0^{\pi} \mathrm{d}\Omega\, r(\theta,\phi)(Y_{l'}^{m'})^*(\theta,\phi) = \sum_{l=0}^{\infty}\sum_{m=-l}^{+l} a_{lm}\delta_{ll'}\delta_{mm'} \tag{18}$$

and the expansion coefficients (ignoring the dummy primes) are given by

$$a_{lm} = \int_0^{2\pi}\int_0^{\pi} \mathrm{d}\Omega\, r(\theta,\phi)(Y_l^m)^*(\theta,\phi)\,, \tag{19}$$

where $\mathrm{d}\Omega \equiv \mathrm{d}\phi\,\mathrm{d}\theta\sin\theta$.

Note that changing the order of integration and summation requires the series in eq. (11) to be uniformly convergent and this is generally the case for piecewise continuous functions. At discontinuities, convergence is in the mean, i.e., the series converges to the average value of the function from the two sides of the discontinuity [19].

The final properties that are required from the spherical harmonics are the effect of rotating the coordinate system. Under such an operation rotated spherical harmonics are given in terms of the unrotated ones by the expression [20]

$$Y_l^m(\theta',\phi') = \sum_{q=-l}^{+l} D_{qm}^{(l)}(\alpha,\beta,\gamma)\,Y_l^q(\theta,\phi)\,, \tag{20}$$

where the matrices $D_{mq}^{(l)}(\alpha,\beta,\gamma)$ form irreducible representations of the rotation group $\mathcal{R}(3)$. These matrices are unitary which is expressed by the following [21]:

$$\sum_{q=-l}^{+l} D_{mq}^{(l)}(\alpha,\beta,\gamma)(D_{m'q}^{(l)})^*(\alpha,\beta,\gamma) = \delta_{mm'}\,; \tag{21}$$

this will be exploited in the development of shape descriptors discussed in the following section.

4. Using spherical harmonics to obtain three dimensional Fourier shape descriptors

Some of the properties of spherical harmonics have been described in the previous section and these can now be exploited in developing three dimensional

Fourier descriptors. Proceeding as for the two dimensional case, a surface is represented as a function of polar coordinates $r(\theta, \phi)$ which is then expanded in terms of the spherical harmonics, eq. (11). The expansion coefficients are the required Fourier shape descriptors given by eq. (19). As in the two dimensional case, these descriptors contain the shape information and can be used to define a quantitative measure of shape difference as follows:

$$D = \left\{ \sum_{l=0}^{L} \sum_{m=-l}^{+l} |a_{lm} - b_{lm}|^2 \right\}^{1/2} . \tag{22}$$

The upper limit, L, is set according to the number of descriptors used.

With the new descriptors normalisation also needs to be considered. In this work, the development of a suboptimum strategy has not been considered. Hence, to obtain the true measure of shape difference eq. (22) would have to be minimised with respect to coordinate operations. With respect to translations this is straightforward and occurs when the centroids of the two shapes coincide. In the case of rotations, the effect on the expansion coefficients can be derived as follows. Expanding a rotated function using rotated expansion coefficients, a'_{lm}, gives

$$r(\theta', \phi') = \sum_m a'_{lm} Y_l^m(\theta, \phi) . \tag{23}$$

However, rotated spherical harmonics and unrotated expansion coefficients may be used to form an expansion. Thus

$$r(\theta', \phi') = \sum_m a_{lm} Y_l^m(\theta', \phi') , \tag{24}$$

and using eq. (20) gives

$$r(\theta', \phi') = \sum_m a_{lm} \sum_q D_{qm}^{(l)} Y_l^q(\theta, \phi) \tag{25}$$

$$= \sum_m \sum_q a_{lm} D_{qm}^{(l)} Y_l^q(\theta, \phi) . \tag{26}$$

Renaming dummy indices m and q gives

$$r(\theta', \phi') = \sum_q \sum_m a_{lq} D_{mq}^{(l)} Y_l^m(\theta, \phi) \tag{27}$$

and swapping the order of summation,

$$r(\theta', \phi') = \sum_m \left(\sum_q a_{lq} D_{mq}^{(l)} \right) Y_l^m(\theta, \phi) . \tag{28}$$

Comparing the above with eq. (15), the relation between rotated and unrotated expansion coefficients or shape descriptors is therefore given by

$$a'_{lm} = \sum_{q=-l}^{+l} D^{(l)}_{mq} a_{lq} \,, \tag{29}$$

with the shape difference measure now taking the form

$$D = \left\{ \sum_{l=0}^{L} \sum_{m=-l}^{+l} \left| a_{lm} - \sum_{q=-l}^{+l} D^{(l)}_{mq}(\alpha, \beta, \gamma) b_{lq} \right|^2 \right\}^{1/2} . \tag{30}$$

It is this expression which should be minimised with respect to α, β and γ.

It can be seen that this expression is still somewhat complicated, largely because of the need to evaluate the rotation matrix $d^{(l)}_{qm}(\beta)$ (where $D^{(l)}_{mq}(\alpha, \beta, \gamma) = \mathrm{e}^{im\gamma} d^{(l)}_{mq}(\beta) \mathrm{e}^{iq\alpha}$). Strategies for tackling the above problem will be discussed in the second of this series of papers though in fact this later work will be dominated by use of invariant descriptors to be described presently.

It has been mentioned that rotationally invariant descriptors have been used in the two dimensional case and again because of the choice of spherical harmonics as basis functions it becomes possible to calculate descriptors which are invariant under three dimensional rotations. This again exploits the unitary properties of the rotation matrices. Begin by taking the modulus squared of eq. (29) and summing over m.

$$\sum_{m=-l}^{+l} |a'_{lm}|^2 = \sum_{m=-l}^{+l} \sum_{q=-l}^{+l} |D^{(l)}_{mq} a_{lq}|^2 \tag{31}$$

$$= \sum_{m=-l}^{+l} \sum_{q=-l}^{+l} D^{(l)}_{mq} a_{lq} (D^{(l)}_{qm})^* a^*_{lq} \tag{32}$$

$$= \sum_{q=-l}^{+l} a_{lq} a^*_{lq} \sum_{m=-l}^{+l} D^{(l)}_{qm} (D^{(l)}_{mq})^* . \tag{33}$$

Thus, using eq. (21)

$$\sum_{m=-l}^{+l} |a'_{lm}|^2 = \sum_{q=-l}^{+l} a_{lq} a^*_{lq} \delta_{qq} \tag{34}$$

$$= \sum_{q=-l}^{+l} |a_{lq}|^2 \tag{35}$$

$$= \sum_{m=-l}^{+l} |a_{lm}|^2 . \tag{36}$$

Therefore define rotationally invariant descriptors A_l by

$$A_l = \left\{ \sum_{m=-l}^{+l} |a_{lm}|^2 \right\}^{1/2} . \tag{37}$$

Using two sets of such descriptors a shape difference measure is defined as follows:

$$D = \left\{ \sum_{l=0}^{L} (A_l - B_l)^2 \right\}^{1/2} . \tag{38}$$

Hence there are now two approaches to quantifying shape, defined by (22) and (38). Although both are referred to by D it will always be clear from the context whether the original descriptors or the rotationally invariant descriptors are being considered.

5. The problem of reentrants

The three dimensional Fourier descriptors as described will only be successful if the function $r(\theta, \phi)$ is single valued. Hence it is not immediately applicable to surfaces with large reentrant regions since identical problems will occur as are illustrated in fig. 2. A completely general method therefore requires the problem of reentrants to be overcome whilst remaining within the framework of polar coordinates and spherical harmonics. Some ideas of overcoming the difficulty are presented in the following subsections.

5.1. USE OF A FOUR DIMENSIONAL HYPERSPHERE

The first approach considered in order to overcome the reentrant difficulties was to use a third parameter which would enable multivalued regions to be distinguished. To be more precise, the multivalued function $r(\theta, \phi)$ is a function defined over a sphere in three dimensions and by projecting this function onto a 4-sphere a single valued function of three polar angles would result. It would still be possible to use spherical harmonics as basis functions, though of course a product of spherical harmonics with some other orthogonal basis set would be required to allow inclusion of the third polar angle. Such a product would be satisfactory since although a four dimensional sphere is being considered only rotations in three dimensions need be considered when normalising a suitably defined distance measure or obtaining rotationally invariant descriptors.

Actually it is also possible to consider an analogous situation in the two dimensional case and this is illustrated in fig. 3.

Here a multivalued contour is mapped to a sphere to become a single valued function of θ and ϕ. Note the function is a step function and has the values 0 or 1. However, being piecewise continuous, Fourier methods are still applicable. This function may then be expanded in terms of spherical harmonics and thus a set of shape descriptors obtained in this way. Again, it is not necessary to consider full

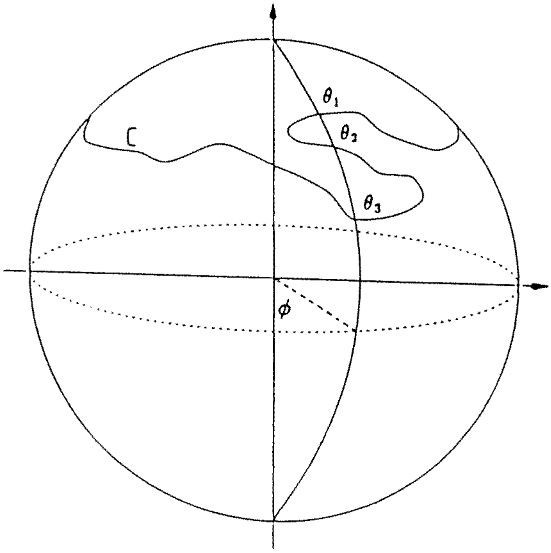

Fig. 3. A contour containing a reentrant is mapped onto a sphere. In this way it becomes a single valued function of two polar angles. Note that the function takes the value 1 inside the contour and 0 outside. Hence a piecewise continuous function defined over a sphere is obtained.

three dimensional rotations when carrying out normalisation but only rotations of the sphere about the z-axis. Therefore a product of ordinary Fourier series would be a suitable method of obtaining shape descriptors.

The ideas just discussed seemed to offer a way around the problem of reentrants even though there would be a considerable increase in computational expense. Some preliminary experiments were therefore carried out, restricted to the case of two dimensional to three dimensional mapping. This work high-lighted a major difficulty of the method and concerned the relation between a series expansion of a function and the derivative of that series expansion and the original function.

It is recalled that the shape of a contour is represented by a function on a sphere taking values zero or one so that the actual shape of the contour is defined by where on the sphere this function changes value. Hence, effectively the shape of the two dimensional contour is represented by the derivative of a step function on a sphere. Now, it is the case that a series expansion represents the function on the sphere and not the derivative of this function and although many terms in the series expansion may produce a good representation of the surface function this may not be the case in terms of derivatives. Hence the original surface function has zero derivative everywhere except at the contour boundary where it becomes infinite. In contrast, the function represented by a finite series expansion will have a derivative which changes over the entire sphere. This therefore calls into question the reliability of shape information contained in the coefficients of such a series expansion, so this approach was not pursued any further. In the following subsection, a further attempt at overcoming the reentrant problem is discussed.

5.2. USING EXTENDED PERIODICITY TO OVERCOME REENTRANTS

A second method considered for overcoming the reentrant problem involved making use of the periodicity of a function defined over a sphere. Figure 4 shows the principle in the case of two dimensions with a contour containing a single reentrant.

The contour is considered as a function in the range 0 to 2π and as illustrated over the subrange a to b it is multivalued. It is possible to consider this function as a new single valued function in the range 0 to 6π and of periodicity 6π so that regions on the contour that have multiple values are shifted by integer multiples of 2π as illustrated. In this way the original multivalued function becomes a piecewise continuous single valued function which may then be expanded as a Fourier series.

This idea may also be carried over into three dimensions by obtaining from a multivalued function in the range $\theta = 0$ to π, $\phi = 0$ to 2π, a new piecewise continuous single valued function in the range $\theta = 0$ to $\pi, \phi = 0$ to T. The new period T would depend on the form of the reentrants, i.e. the degree of multiplicity of the function. Spherical harmonics are now modified so that shape descriptors are now given by

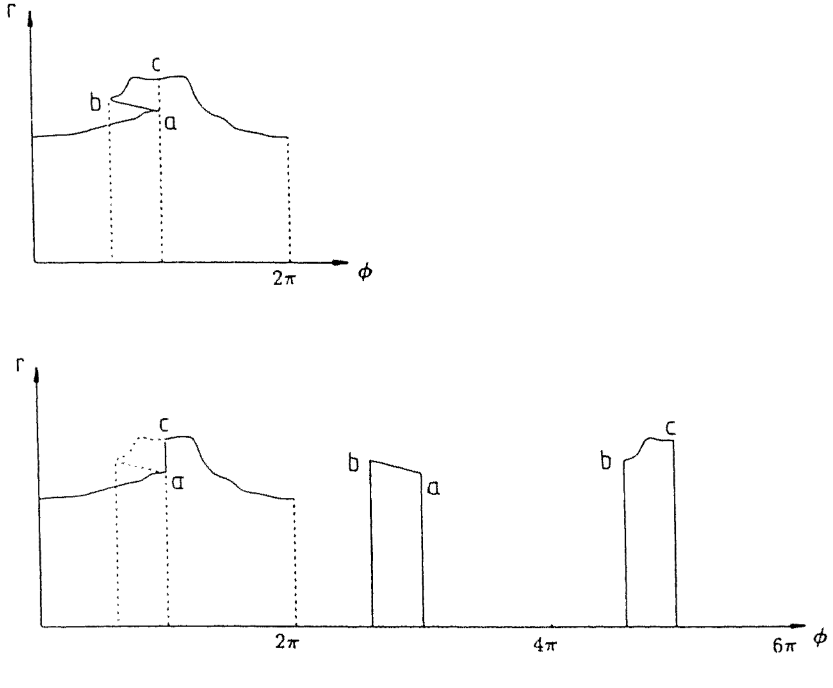

Fig. 4. It is possible to obtain a single valued function of the polar angle ϕ if the contour contains a reentrant. This is done by increasing the periodicity of the function and translating multivalued regions by integer multiples of 2π.

$$a_{lm} = \int_0^T \mathrm{d}\phi \int_0^\pi \mathrm{d}\theta \sin\theta\, \bar{r}(\theta,\phi) P_l^m(\theta) \mathrm{e}^{i\frac{2\pi}{T}m\phi}\,, \tag{39}$$

where $\bar{r}(\theta,\phi)$ is obtained from the original multivalued function $r(\theta,\phi)$.

This in principle allows shape descriptors to be obtained for any molecule but unfortunately there is a further problem when the normalisation procedure is considered. The effect of rotation on the new shape descriptors may be treated by defining a new coordinate, $\phi' = 2\pi\phi/T$, such that eq. (39) becomes

$$a_{lm} = \int_0^{2\pi} \mathrm{d}\phi' \int_0^\pi \mathrm{d}\theta \sin\theta\, \bar{\bar{r}}(\theta,\phi') P_l^m(\theta) \mathrm{e}^{im\phi'}\,. \tag{40}$$

From this point of view it seems that a function representing the shape of the molecular surface $\bar{\bar{r}}(\theta,\phi)$ is expanded in terms of ordinary spherical harmonics as previously described and therefore the effect on the expansion coefficients of rotating this function can be considered in the normal way. To obtain the new function $\bar{\bar{r}}(\theta,\phi)$, the original multivalued function $r(\theta,\phi)$ is transformed into a single valued function of increased periodicity in ϕ. This new surface representation, $\bar{r}(\theta,\phi)$, is then rescaled to produce a single valued function, $\bar{\bar{r}}(\theta,\phi)$, now defined over a sphere.

Unfortunately because of the way reentrants are resolved it is not true that rotating the original molecule by $R(\alpha,\beta,\gamma)$ will produce a similarly rotated reentrant resolved function. Hence, if

$$r(\theta,\phi) \rightarrow \bar{\bar{r}}(\theta,\phi)\,, \tag{41}$$

then a corresponding relation for rotated versions does not hold. That is

$$R(\alpha,\beta,\gamma) r(\theta,\phi) \rightarrow R(\alpha,\beta,\gamma) \bar{\bar{r}}(\theta,\phi)\,. \tag{42}$$

For this reason this method of dealing with reentrants was not considered further and the method to be described in the following subsection was finally adopted.

5.3. USE OF MULTIPLE SURFACES – THE SUBSURFACE SOLUTION

This method is somewhat similar to that described in the previous subsection. There it was seen that reentrant regions were translated by integer multiples of 2π and then the periodicity of the function adjusted to obtain a single valued function. With the subsurface approach reentrant regions are again obtained but these are then considered as separate subsurfaces. Again, this can be seen by referring to fig. 4, which shows the situation for a contour. In the approach now adopted, the contours in the range 0 to 2π, 2π to 4π and 4π to 6π as shown in fig. 4 are considered as separate subcontours each with a periodicity of 2π and in the range 0 to 2π. The contours are ordered so that referring to fig. 4 that contour in the range 0 to 2π is the first order contour (subcontour) and so on to higher order contours.

By analogy this procedure is carried out for surface functions resulting in a number of separate subsurfaces which are again ordered according to the procedure just described. Effectively this means that the first order surface is that surface which can be "seen" from the centroid, the second order surface is that surface which is obscured by the first order surface only and so on. Clearly, if the surface function is single valued then there is only one subsurface and the higher order surfaces are taken as zero. In this way there results a number of separate surface functions each one of which is piecewise continuous and single valued. It can therefore be represented in terms of spherical harmonics. There then results a separate set of shape descriptors which can be incorporated into an overall measure of shape difference. Hence, summing over subsurfaces, shape difference is defined as follows:

$$D = \left\{ \sum_{k=1}^{K} \sum_{l=0}^{L} \sum_{m=-l}^{+l} |a_{lm}^{k} - b_{lm}^{k}|^{2} \right\}^{1/2} . \tag{43}$$

In the above, the sum over k corresponds to summing over each subsurface and the upper limit K is set to the desired number of subsurfaces to be included. Again, invariant shape descriptors can be derived from each set and these used to define a rotationally invariant shape difference measure given by

$$D = \left\{ \sum_{k=1}^{K} \sum_{l=0}^{L} (A_{l}^{k} - b_{l}^{k})^{2} \right\}^{1/2} . \tag{44}$$

Having defined this new measure of shape comparison, some points concerning the significance of the descriptors should be made. Firstly, the zero order descriptor a_{00} corresponds to the size of the surface or subsurface. More exactly, if $\bar{r}$ is the average value of the surface function (its average radius) then

$$\bar{r} = \frac{\int \int_{\Omega} r(\theta, \phi) \, \mathrm{d}\Omega}{\int \int_{\Omega} \mathrm{d}\Omega} . \tag{45}$$

Now $a_{00} = \int \int_{\Omega} r(\theta, \phi)/2\sqrt{\pi} \, \mathrm{d}\Omega$, where spherical harmonic $Y_0^0(\theta, \phi) = 1/2\sqrt{\pi}$. Therefore, the average radius (a measure of the size of the object encapsulated by the surface) is given by $a_{00}/2\sqrt{\pi}$. Similarly it can be shown that the $l = 1$ shape descriptors contain the coordinates of the centroid of the surface ($\bar{x}, \bar{y}$ and $\bar{z}$). Now, it was mentioned earlier that the shape difference measure is a minimum when the centroids of the two surfaces being considered coincide (subject to orientation also). This can be achieved by placing both surfaces such that their centroids lie at the origin of the coordinate system. In this situation the $l = 1$ descriptors will all be zero and can therefore be excluded from any analysis. However, this is only the case if the centroid of each separate subsurface is placed at the origin. If the centroid of the entire multivalued surface is placed at the origin then the centroids of the subsurfaces will not necessarily also coincide with this point. Hence the $l = 1$ descriptors will not be zero and should therefore be considered along with all other terms.

It now seems that there is a general method for obtaining three dimensional Fourier shape descriptors which is able to deal with reentrant surfaces whilst remaining within the realm of spherical harmonics. In particular it is still possible to deal with normalisation in the usual way. In the following sections the practice of obtaining and using the shape descriptors of a given molecule is described.

It is worth noting here that in the work of Max and Getzoff [6], though the existence of reentrants was remarked upon, no satisfactory solution to the problem was presented.

6. Evaluating harmonic descriptors

6.1. CHOICE OF INTEGRATION METHOD

It is supposed that a surface representation of a molecule is obtained and it is then required to evaluate the double integral, eq. (19) given earlier, i.e., the harmonic shape descriptors. There are three factors to consider in choosing how to evaluate the integral; these being the speed of integration, the accuracy of the result and the ease of implementation. As far as integration methods are concerned, there are a number to choose from [22], but for simplicity and ease of implementation it was decided to use Simpson's rule.

One of the factors influencing speed is the time taken to evaluate the integrand and the dominating factor here is evaluating the spherical harmonics or more specifically the associated Legendre polynomials. There are in fact a number of formulas for evaluating the associated Legendre polynomials [23] and the one adopted here is a recursion technique given by Press [24]. Unfortunately, for high order harmonics a great many recursions are required and this becomes expensive in computer time. It is possible to save considerable time if the Legendre polynomials are evaluated once only at equal angular intervals and the results stored in an array at the beginning of the integration program (this also allows code vectorisation if such a facility is available). Hence, during the integration no more time is needed to obtain values for a high order Legendre polynomial than that required for a low order one, in both cases this being a simple matter of array access. This is then suited to using an integration method where the integrand is evaluated at equal predefined intervals, in this case Simpson's rule. It is of course a double integral which is to be evaluated so that the inner integral is evaluated using Simpson's rule which then forms the integrand of the outer integral this again being evaluated by Simpson's rule. Also included in the inner integral is the following correction factor [22]:

$$-\frac{3}{90}\sum_{n_{\text{odd}}} f_{n-2} - 4f_{n-1} + 6f_n - 4f_{n+1} + f_n \,. \tag{46}$$

This is added to the following Simpson rule for integrand f:

$$\frac{h}{3}\left[f_0+f_N+2\sum_{n_{\text{odd}}} f_n+4\sum_{n_{\text{even}}} f_n\right], \tag{47}$$

where the integrand is divided into N (even N) equal intervals of width h and f_0 and f_N are the integrand values at the lower and upper limits of integration, respectively.

6.2. SOME RESULTS ON THE ACCURACY OF THE SHAPE DESCRIPTORS

In order to assess how accurate the resulting descriptors are when evaluated using the above method, the procedure was tested using the square of an ellipsoidal function

$$r(\theta,\phi)=(a\sin\theta\cos\phi)^2+(b\sin\theta\sin\phi)^2+(c\cos\theta)^2\,. \tag{48}$$

For such a shape it is possible to calculate descriptors analytically up to $l=2$ and all higher order descriptors are zero. The shape descriptors are therefore given by

$$a_{lm}=\int_0^{2\pi}\mathrm{d}\phi\int_0^{\pi}\mathrm{d}\theta\sin\theta\{(a\sin\theta\cos\phi)^2+(b\sin\theta\sin\phi)^2 \\ +(c\cos\theta)^2\}(Y_l^m)^*(\theta,\phi)\,. \tag{49}$$

Evaluating the above analytically and by computer program therefore gives an indication of the accuracy of the chosen technique and some figures are shown in table 1 for different values of the parameters a, b, c.

In general it was found that up to $l=50$ the results were accurate to four significant figures. A further way of assessing the accuracy of the calculated descriptors is to reproduce the original shape from the descriptors, i.e., using eq. (15) and the method has again been seen to produce excellent results [5].

7. The molecular shape

7.1. THE MOLECULAR SURFACE MODEL

A method for evaluating shape descriptors is now available and it is therefore necessary to obtain a suitable representation of the surface of a molecule which allows the integration to be carried out. This surface is the shape for which hitherto no general method of shape analysis had been developed.

The simplest case that may be considered is to take an entire molecule and obtain a set of descriptors for this molecule. This will be the case when interest is centered on the entire molecule and its shape properties rather than on some specific region such as a receptor site on the molecule. The latter situation will be dealt with presently but in either case the three dimensional atomic coordinate structure

Table 1
Comparisons of computed and actual values for a number of shape descriptors.

$a = 1.0, b = 2.0, c = 3.0$

Descriptor	Actual value	Computed value
a_{00}(real)	16.542902839	16.542902978
a_{00}(imag)	0.000000000	0.000000000
a_{10}(real)	0.000000000	0.000000855
a_{10}(imag)	0.000000000	0.000000000
a_{11}(real)	0.000000000	−0.000000285
a_{11}(imag)	0.000000000	0.000000002
a_{20}(real)	6.869767140	6.869766705
a_{20}(imag)	0.000000000	0.000000000
a_{21}(real)	0.000000000	0.000000000
a_{21}(imag)	0.000000000	0.000000000
a_{22}(real)	−1.941625940	−1.941625700
a_{22}(imag)	0.000000000	−0.000000003

$a = 2.6, b = 51.5, c = 0.5$

Descriptor	Actual value	Computed value
a_{00}(real)	3142.277128760	3142.277221764
a_{00}(imag)	0.000000000	0.000000000
a_{10}(real)	0.000000000	0.000078771
a_{10}(imag)	0.000000000	0.000000000
a_{11}(real)	0.000000000	−0.000118233
a_{11}(imag)	0.000000000	0.000000003
a_{20}(real)	−1404.872771656	−1404.872796894
a_{20}(imag)	0.000000000	0.000000000
a_{21}(real)	0.000000000	0.000000000
a_{21}(imag)	0.000000000	0.000000000
a_{22}(real)	−1712.184002292	−1712.183992708
a_{22}(imag)	0.000000000	0.000280836

of the molecule is needed. In the case of protein molecules structures can be those obtained using crystallographic techniques and stored in the Brookhaven protein data bank [25]. For small molecules, structures can be calculated by, for example, using the AMPAC package [26]. Once a structure is available it is then necessary to obtain some representation of the surface "envelope" of the molecule and this is done using Connolly's program [27]. This generates coordinates of points which lie on the so called solvent accessible surface of the molecule [28]. This is not the only model of the surface that may be taken so for example energy surfaces could also be considered.

When interest is centered around some local region of a molecule there are two approaches that may be taken in order to obtain a "molecular envelope". The first is to generate surface points for the entire molecule and then use some criterion for extracting those points that lie on the region of interest. Such a criterion will in general be based on a cutoff distance so that surface points lying within a given distance from some chosen centre are obtained.

In practice it is convenient to consider structures of bound complexes and obtain points on each surface which lay within a certain distance from the other surface. This is illustrated in fig. 5 and a program, the interface program, has been written to carry out this process.

The second possibility is to decide, perhaps with the aid of interactive computer graphics, which atoms form a region of interest; for example, those atoms which form a binding or active site of a molecule. This subset of atom coordinates can then be used by the Connolly program for the generation of surface points over the desired region of interest.

The essential difference in the two approaches is in the way in which an exposed region that was otherwise embedded inside the molecule is "closed" off. In the former this region remains open with no surface points whereas in the latter atoms become fully exposed and surface points are generated (compare fig. 5 and fig. 6).

The latter case may be considered somewhat artificial since the embedded

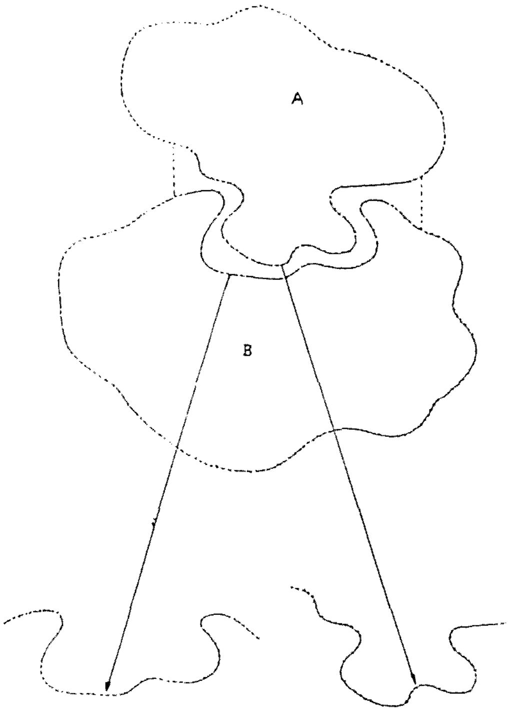

Fig. 5. If the structure of a complex is available then surface points are generated for each individual molecule (by editing the single file of the complex coordinates and producing two separate files for each molecule). A distance criterion is then used to obtain points on the interface regions.

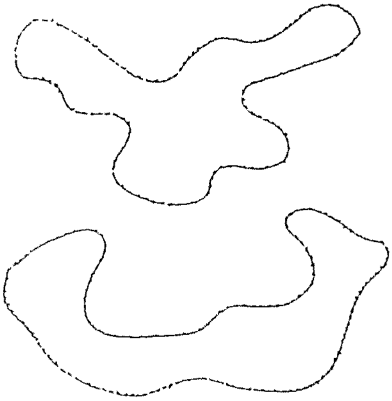

Fig. 6. Atoms forming the regions of interest over a molecule are selected. Connoly points are then generated for these atoms thus "closing off" an otherwise open region of the surface.

region does not play a role in shape complementarity and should not therefore be used in generating shape descriptors. In particular the effect of this closing off is rather different between hollows and protrusions and for this reason this method would seem more suited to comparing similarities in shapes between either a group of receptor sites or a group of ligands.

In either case the result is a set of points or Cartesian coordinates corresponding to the solvent accessible surface. This surface is then the shape for which it is desired to calculate shape descriptors.

7.2. MAPPING TO THE SURFACE OF A SPHERE - THE MAPPING PROGRAM

Connolly's program outputs a list of x y z coordinates and it is evident that the surface must be represented as a function of polar coordinates in order to form an expansion in terms of spherical harmonics. This may be viewed as a process of mapping the molecular shape to the surface of a sphere and would appear to be a straightforward exercise in converting from Cartesian to spherical polar coordinates. However, there are two problems regarding this. The first point is that a direct conversion from Cartesian to polar coordinates is unlikely to produce points at uniform intervals of polar angles, as is required for carrying out the integration. The second problem is to resolve reentrant regions such that from a single multivalued function $r(\theta, \phi)$, several single valued functions $r^k(\theta, \phi)$ are obtained.

To obtain values at uniform intervals requires some method of interpolation and the strategy chosen was the simplest, i.e., that of bilinear interpolation [24]. In order to carry out this operation surface points are converted to polar coordinates and then sorted into "boxes". These boxes are formed by a grid of θ and ϕ values at one degree intervals, the step value used in the Simpson integration. This will generally give the result illustrated in fig. 7 since no assumptions are being made concerning the distribution of points.

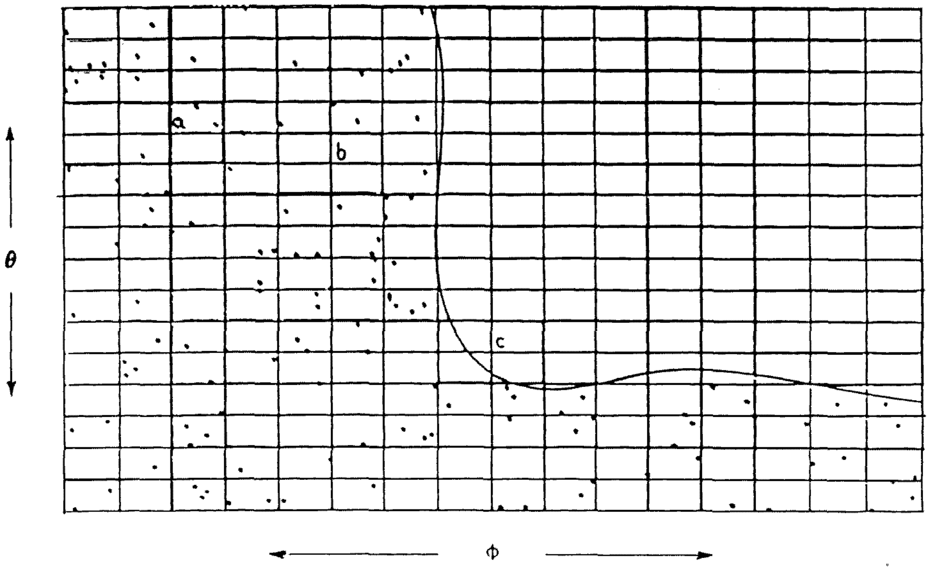

Fig. 7. Mapping to the surface of a sphere. Points surrounding grid intersections (i.e. integer θ and ϕ) are used to interpolate to values at these intersections. See text for further explanation.

Once the points have been sorted into their respective boxes, interpolation to integral values of θ and ϕ can take place. To do this, at each integral value of θ and ϕ, points in the surrounding four boxes are obtained and the values $r(\theta, \phi)$ of these points are used to estimate the value r, at the particular integral θ and ϕ.

Unfortunately the situation is somewhat more complicated because of how the points may be distributed. Hence it may occur that there are many points in the surrounding four boxes (coordinate a in fig. 7), in which case interpolation will work well, or there may be no points at all in the surrounding boxes (coordinate b). In this latter case the strategy is to widen the search so that boxes further away are also considered. At the next level there are twelve boxes to be considered. This process can in principle go on indefinitely and the further afield the search then the less accurate is likely to be the interpolated value. It also may be the case that the value to be interpolated is actually zero so that at some stage the search for points should be abandoned and the interpolation value set to zero (coordinate c which lies outside the piecewise continuous surface being considered).

Because of these uncertainties there is no criterion for deciding how far out the search should be made before setting an interpolated value to zero. The ideal search level or radius will depend on how densely the points are distributed over the molecule (a parameter which can be set in Connolly's program). Also, having converted these points into polar form and sorted into boxes there will be differences in the number of points per grid box over various regions of the sphere. This will also include the possibility of boxes having no points. This difference will be most pronounced if "polar" regions are compared with "equatorial" regions. Hence, even if it is supposed that the number of points per unit area is uniform over the sphere, since the number of grid boxes per unit area is higher at the poles than at the equator the result would be more points per box at the equator than at the poles. This is a consequence of the non-Euclidean nature of the surface of a sphere; one degree grid boxes do not cover the same area from one region of the sphere to another. Note also that the number of points per grid box is influenced by the size of a molecule. The most suitable way of overcoming the problem is by having as high a density as possible and a low search radius and this was the general strategy adopted. It is possible to change the point search strategy according to the region of the sphere being considered but this idea was not implemented.

The second even more difficult problem is that caused by reentrants. In converting to polar coordinates and then sorting into boxes, points that lie on separate regions of surface (i.e., $r(\theta, \phi)$ is multivalued giving a number of separate surface regions) will be placed into the same box (fig. 8).

Hence, iterations using all such points will result in an error with the corresponding consequences as shown in fig. 8. The problem is therefore how to decide first whether there is a reentrant region and secondly select those points on a particular region of the surface for interpolating a value on that region of surface. As there is no way of knowing where reentrants lie or indeed their severity, the difficulty in this task becomes obvious.

This problem has been overcome with some success by considering the radial spread of points within each box. The points are sorted into increasing order of distance from the origin and then sorted into groups based on their separation. The process is one of "running" through points separated by increasing radius (i.e., increasing distance from the origin) until a point is encountered whose radial value

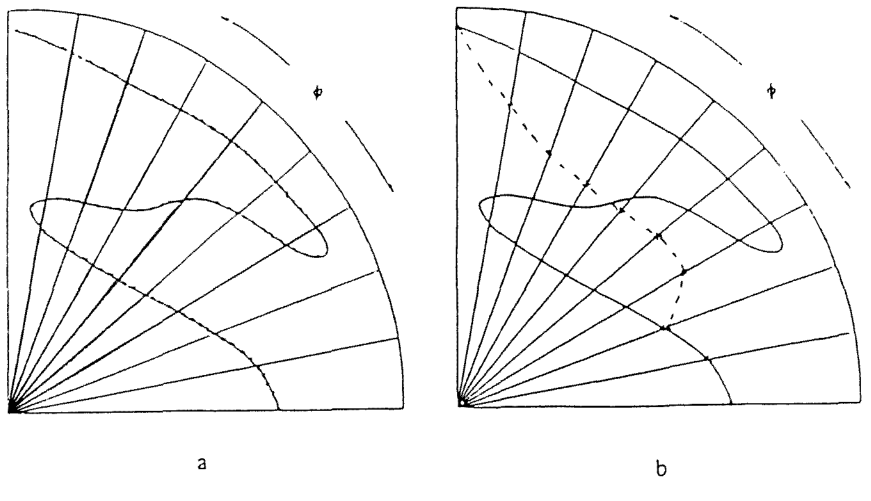

Fig. 8. The reentrant problem. a) Points on a multivalued contour. b) The result of direct interpolation to integral values of ϕ. c) Step distance l allows points to be resolved to separate contour regions. d) However reentrant edges remain unresolved.

is greater than the previous point by more than a set value. All previous points are then taken to lie on a separate surface region and the large separation between the current and previous surface point indicates the beginning of a new surface region (see fig. 8). Again, there is the difficulty of deciding on the set value for beginning a new surface region; too large a distance and separate surfaces will not be resolved and too small and any local undulations in the surface will be treated as separate regions. A value of 1.4 Å was chosen since this certainly will resolve reentrants capable of letting a water molecule in. However, even with this choice it is still impossible to cope with the edges of reentrants without any error since the edge of a reentrant corresponds to where surfaces meet (the radius vector at such a point is tangential to the surface) and it would require rather more sophisticated techniques to be able to deal with this situation without any problems.

A less serious problem is the number of separate subsurfaces that are required when dealing with reentrants. Recall that a multivalued function is resolved into several single valued functions for each of which shape descriptors are calculated. For molecules with a great many complicated reentrants, the corresponding function will be highly multivalued and many separate subsurfaces will be required. However, the situation is generally unknown so that the worst possible case, within reason, must be allowed for. Therefore it may happen that some of the separate subsurfaces will in fact be entirely zero and of course the shape descriptors for these will be zero. In such a situation these will clearly make no contribution to any shape difference measures.

8. From raw descriptors to shape difference

8.1. USING ROTATIONAL INVARIANTS

Previously it was shown that expansion coefficients, i.e., the shape descriptors could be used to calculate new coefficients that were independent of the orientation of the original coordinate system. In other words, they are rotationally invariant. A program has been written to evaluate these descriptors and calculate a shape difference measure using eq. (44). With these descriptors normalisation is now unnecessary but information has of course been lost and it is not possible to regenerate the original shape from these modified descriptors. In fact, in calculating the modified descriptors phase information is lost.

8.2. NORMALISATION

Normalisation as already explained involves finding solutions such that the distance measure (eq. (43)) becomes a minimum. This is then a full normalisation. Other normalisation strategies were not considered, i.e., suboptimum strategies [14]. These approaches may be suitable for automating recognition process but they do not generally give a minimum measure of the distance function (recognition can be based on a cut-off value) and for this work it is preferred to have a definite shape difference value and not some upper bound.

Inserting the rotation matrices into eq. (43) means the most general normalisation is therefore to find the values of alpha, beta and gamma which minimise the equation

$$D = \left\{ \sum_{k=0}^{K} \sum_{l=0}^{L} \sum_{m=-l}^{+l} \left| a_{lm}^{k} - \sum_{q=-l}^{+l} D_{mq}^{(l)}(\alpha, \beta, \gamma) b_{lq}^{k} \right|^{2} \right\}^{1/2} . \tag{50}$$

As far as translation is concerned, the minimum occurs when the centroids of the two shapes coincide and this is a straightforward task. Actually in all cases the origin of the polar coordinate system is set at the centroid of the molecular surface being dealt with. In view of the comments at the end of subsection 5.3, this means that the centroid of each subsurface will not lie at the origin and the $l = 1$ shape descriptors will in general not be zero and are therefore considered along with the other shape descriptors that are being used for the analysis. It is possible to minimise the shape difference measure by obtaining the minimum of each subsurface contribution. Hence, this would involve translating each subsurface so that the origin coincided with the centroid of that subsurface and using the descriptors calculated under these circumstances. This however does not preserve the spatial relationships between the individual subsurfaces and is therefore inappropriate. The same argument applies when considering rotations also; the same rotation is applied to each subsurface (so subsurfaces do not rotate and "slide" over each other).

There are three approaches that may be chosen in attempting to find values of the rotation angles that give a minimum to the above equation. The first is simply to perform a grid search over the three Euler angles. Note that if translations had to be included in the normalisation expression, such a search using a reasonable grid size would not be possible. However, even with just three degrees of freedom this approach is somewhat expensive. The second method is to differentiate with respect to the three Euler angles to obtain three simultaneous equations. These are then set to zero and the solutions give turning points in the function. This method has been used with success in the two dimensional situation [29] but is not recommended for higher dimensional cases for the reasons outlined in Press [24]. The third method is to use a multi-dimensional minimisation technique (as opposed to multi-dimensional root-finding just described).

It is required to find the global minimum of eq. (50) and this will depend on having initial estimates of alpha, beta and gamma "close" to the global minimum. To achieve such estimates would most likely require the use of interactive graphics since as has been seen the minimum distance measure corresponds to the two shapes fitting as closely together as possible (according to eq. (4) in a two dimensional situation).

A second possibility is to consider locating minima for low resolution surfaces, i.e., low order descriptors only are considered. Since such low order surfaces will be relatively smooth, the resulting shape difference measure will also be relatively smooth. Hence, there will be few turning points or local minima and it should not be a difficult task to locate the global minimum. Further, the global minimum for a low order representation should also be close to the global minimum for a high order representation since the overall shape of the molecules and hence the overall shape of the difference measure function are contained in the low order descriptors. From this basis, an approach seems possible for the location of global minimum, even to high order shape representations, by first considering low order terms.

A program was written which locates minima given initial estimates using the downhill simplex method [24]. The program worked satisfactorily when considering ellipsoidal functions and was able to find the best positions for fitting ellipsoidal surfaces together (involving rotations of 90 degrees) but otherwise it appeared unreliable even for low order shape representations and a program of obtaining low order minima and then moving to higher orders has so far not been tested.

9. Quantitative molecular shape difference

Of itself a quantitative shape difference measure between two bodies holds little meaning without there being any standard of reference. One possible way to overcome this is to introduce some standard range of shape difference so that, for example, a value of zero corresponds to identical shapes and a value of 100, say, corresponds to completely dissimilar shapes. The latter may be defined by producing random coordinates of atoms, obtaining Connolly surfaces for these and calculating shape descriptors in the usual way. The shape difference given by such shape descriptors may then be taken to define completely dissimilar shapes and thus a suitable normalisation (i.e. scaling) for a standard shape difference scale is obtained. It may also be advantageous to introduce certain constraints when random coordinates are produced in order to produce more realistic "random molecules" but the main point is that there should be no correlation between completely dissimilar shapes.

However, the need for such a scale is bypassed when a number of molecules are being compared with respect to each other. This can be done by taking the molecules of interest, obtaining shape difference measures between all the molecules and constructing a distance matrix or dissimilarity matrix containing the quantitative shape difference information. By means of classical scaling [30] it is often possible to obtain a visual display in two or in three dimensions of this distance matrix. In this way, relationships such as clusters of molecules can be directly observed. Results obtained using this technique will be presented in a subsequent paper.

10. Conclusion

In this paper, a method of obtaining shape descriptors for molecules or regions of molecules based on the theory of spherical harmonics has been described. In addition a number of difficulties have been discussed and ways of overcoming these problems presented. Computer programs have been written to obtain shape descriptors and the method adopted for calculating the shape descriptors has been seen to provide accurate shape descriptor values. It has been demonstrated that these values do indeed contain the required shape information and it has also been shown how these values can be used to calculate a quantitative measure of shape difference between molecules. For the raw descriptors, this difference must be minimised with respect to coordinate operations which for large numbers of molecules is an expensive task. However, further descriptors can be calculated which do not require normalising and these used to define a shape difference measure. For a large number of molecules, a distance matrix may be obtained and a visual representation of the distance matrix can be obtained using the techniques of classical scaling.

References

[1] P.G. Mezey, J. Comp. Chem. 8 (1987) 462.
[2] A. Verloop, W. Hoogenstraaten and J. Tipker, Drug Design 7 (1976) 165.
[3] G.A. Arteca and P.G. Mezey, J. Comp. Chem. 9 (1988) 554.
[4] S.E. Leicester, Quantitative molecular surface shape analysis using spherical harmonic Fourier shape descriptors, Ph.D. Thesis, University of London (1989).
[5] S.E. Leicester, J.L. Finney and R.P. Bywater, J. Mol. Graph. 6 (1988) 104.
[6] N.L. Max and E.D. Getzoff, IEEE Comp. Graphics Appl. 8 (1988) 42.
[7] T. Wallace and P.A. Wintz, Fourier descriptors for extraction of shape information, in: *Image Understanding and Information Extraction*, eds. T.S. Huang and K.S. Fu, School of Electrical Engineering, Purdue University, Indiana, USA TR-EE 77-35 (1977).
[8] F. Etsami and J.J. Uicker, Comp. Vision, Graphics Image Process. 29 (1985) 216.
[9] D.L. Fritzsche, A systematic method for character recognition, Ohio State Univ. Res. Found., Columbus, Rep. 1222-4 ASTIA AD 268-360 (1961).
[10] K. Gotoh and J.L. Finney, Powder Tech. 12 (1975) 125.
[11] R.L. Cosgriff, Identification of shape, Ohio State Univ. Res. Found., Columbus, Rep. 820-11, ASTIA AD 254 792 (1960).
[12] C.T. Zahn and R.Z. Roskies, IEEE Trans. Comp. C21 (1972) 269-271.
[13] P. Dennery and A. Krzywicki, *Mathematics for Physicists* (Harper & Row, New York, 1967).
[14] O.R. Mitchell, P. Soc. Photo. 442 (1983) 38.
[15] T. Wallace and P.A. Wintz, Comp. Graphics Image Process. 13 (1980) 99.
[16] B. Shridhar and A. Baldredin, Pattern Recognition 17 (1984) 515.
[17] K.S. Park and N.S. Lee, Comp. Biomed. Res. 20 (1987) 125-140.
[18] R.A. Crowther, The fast rotation function in the molecular replacement method, in: *The Molecular Replacement Method: A Collection of Papers on the Use of Non-crystallographic Symmetry*, ed. M.G. Rossman (Gordon & Breach, New York, 1972).
[19] E.W. Hobson, *Spherical & Ellipsoidal Harmonics* (Cambridge University Press, 1931).

[20] J.P. Elliot and P.G. Dawber, *Symmetry in Physics*, Vol. 1 (Macmillan, London, 1979).

[21] A.W. Joshi, *Elements of Group Theory for Physicists* (Wiley Eastern, New Delhi, 1973).

[22] J.D. Davis and P. Rabinowitz, *Methods of Numerical Integration* (Academic Press, London, 1987).

[23] R.A. Wiggins and S. Masanori, Bull. Seismolog. Soc. Am. 61 (1971) 357.

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

[25] 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.

[26] Dewar Research Group and J.P.P. Stewart, AMPAC 2.0, QCPE Bull. 6 (1986) 24 (QCPE No.: 506).

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

[28] F.M. Richards, Ann. Rev. Biophys. Bioeng. 6 (1977) 151.

[29] E. Persoon and F. King-Sun, IEEE Trans. Syst., Man & Cybern. SMC-7 (3) (1977) 170.

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