Description of molecular surface shape using Fourier descriptors

Steve E. Leicester, John L. Finney and Robert P. Bywater*

Department of Crystallography, Birkbeck College, London, UK
Molecular Biophysics Department, Pharmacia AB, Uppsala, Sweden*

We have developed a method for describing the three-dimensional (3D) shape of molecules based on the Fourier shape descriptor technique. Our method has proven valuable in two-dimensional (2D) shape description and recognition in a number of important areas. In this paper, we discuss the 2D method briefly and explain how we adapted it to the 3D description of molecular surfaces. Our method is based on representing a molecular surface in terms of spherical harmonics, and some results on the accuracy of such a representation are shown. We discuss the use of this representation for two interacting molecules for quantifying the goodness of fit based on the topography of the relevant interfaces.

Keywords: shape, molecular surface, Fourier descriptors

Received 6 January 1988
Accepted 2 February 1988

It has long been recognized that the successful interaction of one biomolecule with another depends upon, among other things, the shape complementarity of the two molecules.[1] Computer graphics has proved to be an important tool in displaying molecules and has enabled possible interactive sites to be located based not only on the required chemistry but also on recognizing sites with shape complementarity. This makes full use of the human cognitive powers in shape matching, but, of course, no quantifiable measure of shape difference or the goodness of fit can thereby be obtained. Ideally, a fully automated system would exist that facilitates the location of possible geometrically complementary sites and that also yields a measure of this complementarity. Unfortunately, such a system does not exist.

Although the automatic location of geometrically complementary sites is at present unfeasible, the immediate problem of measuring shape difference can be tackled as a step toward this goal. Shape difference measurements do, in fact, have wide application in computer systems requiring recognition of objects based on an image of their shape. Examples of this are automatic machine part recognition[2] and aircraft recognition,[3] in which the computer must recognize an object from an image of its outline. Such techniques have a record of the original, so the recognition process is one of comparison of shapes. Because of the nature of computers, shapes must be quantitative for an automated recognition decision to be made. Such two-dimensional (2D) shape-matching problems as these have been solved within the framework of Fourier analysis using the so-called Fourier descriptor technique.[4] This work describes how we have adapted the 2D technique for the analysis in three dimensions of the shapes of molecules. A brief description of the 2D problem begins the discussion.

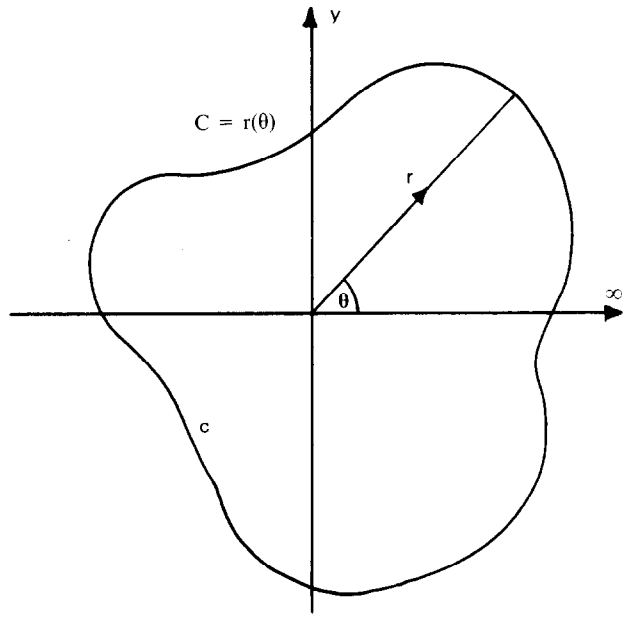

Figure 1. Parametric representation of a shape contour c

TWO-DIMENSIONAL FOURIER DESCRIPTORS

To obtain a quantitative measure of shape difference, the outline of the shape is described as a parametric function, as shown in Figure 1. This function is periodic of period 2π and can therefore be represented in terms of a Fourier series:

$$r(\theta) = \sum_{n=-\infty}^{+\infty} a_n e^{in\theta} \qquad (1)$$

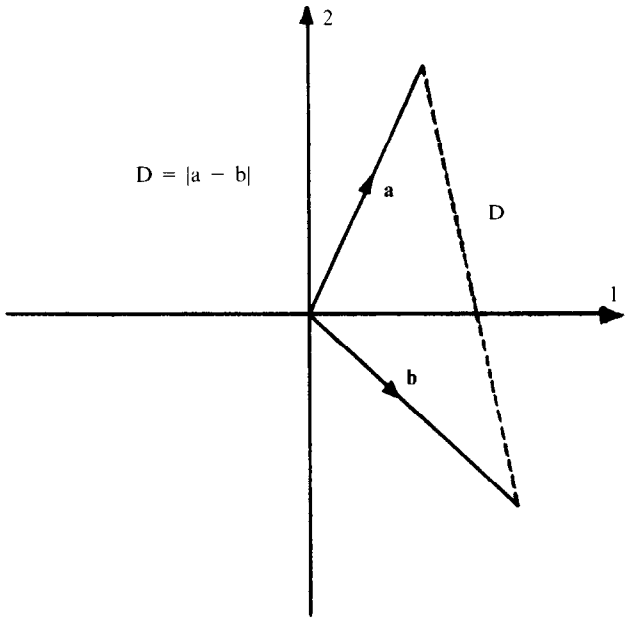

Figure 2. Descriptors are considered as components of a vector; in this case, there are two descriptors for shapes a *and* b. D *then gives a quantitative measure of shape difference*

The expansion coefficients a_n are obtained from the expression

$$a_n = \int_0^{2\pi} r(\theta)\, e^{-in\theta}\, d\theta \qquad (2)$$

For two such shapes, two sets of Fourier descriptors are obtained that can then be interpreted as n-dimensional vectors if the expansion is taken to n terms. The shape difference, D, is then defined as the magnitude of the vector difference of the two vectors. Essentially, then, a shape is represented by a position vector in an n-dimensional vector space, and the distance between end points of vectors is a measure of shape difference for the shapes that the vectors represent. This is illustrated in Figure 2.

Unfortunately, there is an additional complication, since the Fourier descriptors will depend on how the coordinate system is set up, and various coordinate operations are necessary such that the difference function becomes a minimum. This is usually referred to as normalization;[5] after normalization, we have an absolute measure of shape difference. Generalizing these ideas to three dimensions is now considered.

THREE-DIMENSIONAL FOURIER DESCRIPTORS

Describing the shape of molecular surfaces requires a 3D Fourier technique; a parametric description of the surface — which will depend on two parameters — is therefore needed. The ideal would be a parametric representation that always gives a single valued function, no matter what the surface; no such representation capable of handling reentrants has yet been found. Our choice of parameterization was guided by the need to consider the effect of rotations on the expansion coefficients — that is, the descriptors — in the processes of minimizing the distance function or of normalization. This leads naturally to the choice of polar coordinates and spherical harmonics as orthonormal basis functions instead of the trigonometric functions usually used in the 2D case. This parameterization also has the advantage of simplicity.

METHOD

For its starting point, we require a polar representation of the surface or region of a surface of a molecule. The surface model we used takes the molecular surface generated by the Connolly MS program.[6] This program produces $x\ y\ z$ coordinates of points lying on that surface which is mapped out when a sphere of a given radius is rolled over the van der Waals surface of a molecule. The molecular coordinates are obtained from the Brookhaven Protein Data Bank file.[7] If the sphere radius is 1.4 Å, then $x\ y\ z$ coordinates that sample the so-called solvent-accessible surface are produced.[8] To convert these $x\ y\ z$ coordinates into polar coordinates, we wrote a program that also performs interpolation calculations to obtain a function $r(\theta, \phi)$, defined at integer values of the polar coordinates θ and ϕ. Effectively, the surface description produced by MS is mapped to a unit sphere.

In order to represent this function in terms of spherical harmonics, it must be, among other things, single valued. At this preliminary stage, this has been contrived by choosing a suitably large rolling probe sphere, although work is in hand in overcoming this problem due to reentrants. This function can be represented by a convergent series of spherical harmonics[9] Y^m_l ($l = 0, 1, 2, \ldots$, $m = -1, -1+1 \ldots 0, 1, \ldots 1$), which in a complex numbered representation is given by

$$r(\theta, \phi) = \sum_{l=0}^{+\infty} \sum_{m=-l}^{+l} a_{lm}\, Y^m_l(\theta, \phi) \qquad (3)$$

From this expression the expansion coefficients or descriptors a_{lm} are given by a double integral over the unit sphere as follows:

$$a_{lm} = \int_0^{2\pi} \int_0^{\pi} r(\theta, \phi)\, Y^{*m}_l(\theta, \phi) \sin\theta\, d\theta\, d\phi \qquad (4)$$

We wrote a program to evaluate the descriptors from the above expression. We used Simpson's rule with a first-order correction for the inner integral and Simpson's rule without correction for the outer integral. We calculated the Legendre polynomial part of spherical harmonics once only by backward recursion,[10] and these values are read from a file and stored in arrays at the beginning of the main integration program. Our present work has limited maximum l value to 50, although in principle there is no limit to the number of terms that can be calculated. However, with l equal to 50 and m taking values 0 to 1, there are $(50 + 1)(50 + 2)/2$ integrals to be calculated. (Note that $a_{lm} = (-1)^m a^*_{l-m}$ where * indicates complex conjugation so that the number of integrations is $(l + 1)(l + 2)/2$ and not $(l + 1)(l + 1)$.) With one degree step intervals, this takes on a VAX 8600 about 9 hours for 1 326 integrations. This time will be considerably reduced using a vectorized

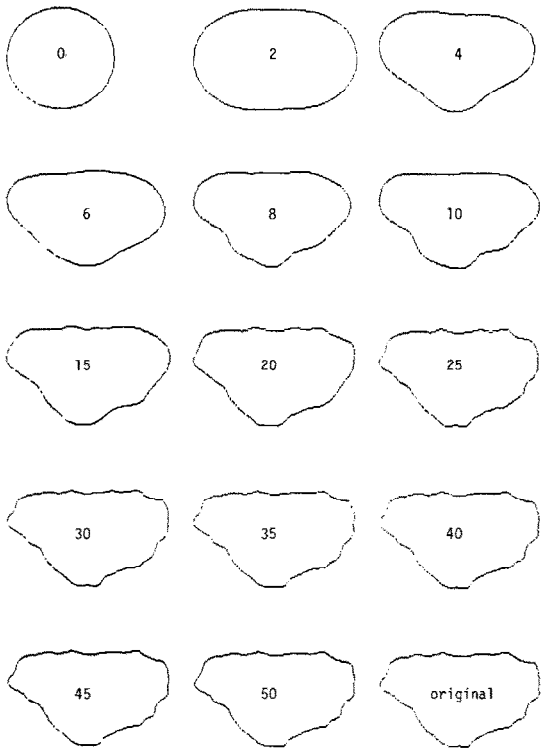

Figure 3. Cross-sectional plots of the test molecule hemerythrin B for differing l *values. The* l *values are shown within each contour*

version written to run on a CONVEX C1. It may, of course, be possible to reduce computational time by developing a discrete transform formulation and an optimized summation technique along the lines of the Fast Fourier Transform.[11] Hence, we do not see CPU time taken as a serious limitation to the technique.

As a check on the accuracy with which shapes can be represented by this method, we calculated coefficients for a function for which they are zero for $l > 2$. The integrals can be evaluated analytically for the remaining low-order terms. In this case, and working to double precision, low-order terms were accurate to 7 decimal places; this was also in general the case for higher-order terms, though for some particular linear terms the accuracy fell to 4 decimal places. This may be due to the highly oscillating nature of the high-order basis functions for which Simpson's rule may not be the most suitable evaluation procedure. Perhaps Filon's rule would be more appropriate, though so far its use has not been necessary.

In order to determine the number of terms required to give a reasonable approximation to the original surface, we wrote a program that regenerates the original surface from a specified number of spherical harmonics using expression (3). This procedure was carried out using hemerythrin B as a test molecule, with a probe sphere of diameter 6.4Å. There are various ways of displaying this regenerated surface graphically. You can display cross-sectional plots using DEC RGL software (Figure 3); alternatively, you can use a vector graphics

Table 1. Average rms error between original and expanded surface

lmax	Average rms error(%)
0	21.3
2	10.3
4	8.7
6	4.7
8	3.3
10	2.4
20	1.0
30	0.8
40	0.7
50	0.7

system. Color plates 1–10 show reconstructions of the shape of the test molecule at different l values (lmax) taken on an Evans and Sutherland PS390. From these examples, it is clear that for relatively smooth functions a reasonable representation can be obtained after $l = 20$. Table 1 shows the average root mean square error between the expanded surface and the original surface. The average was calculated for surface values taken at intervals of 5° in the polar coordinates.

SHAPE COMPARISON

This method has so far produced an alternative representation of a molecular surface, and for two surfaces we should like to make a quantitative comparison of the surface shapes. In exactly the same way as was done for the 2D contours, we can define a shape difference measure as follows:

$$D^2 = |\mathbf{a} - \mathbf{b}|^2 = \sum_{l=0}^{l\max} \sum_{m=-l}^{+l} (a_{lm} - b_{lm})(a^*_{lm} - b_{lm}^*) \quad (5)$$

which would therefore quantify the difference in shape between the two surfaces. As for the 2D case, this measure must be minimized with respect to rigid body coordinate operations. The translation is straightforward, since a minimum occurs when the centroids of the surfaces coincide. For rotations, we wrote a program to find local minima by iteration from initial values. These initial values are Euler angles through which one surface must be rotated with respect to the second to give a low value for the distance function. For this to work, it is necessary to consider the effect of rotation on the expansion coefficients so that expression (5) is a function of three Euler angles. For an absolute shape difference measure, the global minimum must be found, and hence orientation angles must be chosen so that the solution converges to this value. To obtain such values at this stage would require the use of interactive graphics so that geometrical shapes could be docked together manually. The change in orientation in doing this would give the Euler angles close to the values that would give the global minimum of expression (5). The minimization program will then converge to give Euler angles such that the distance function is at

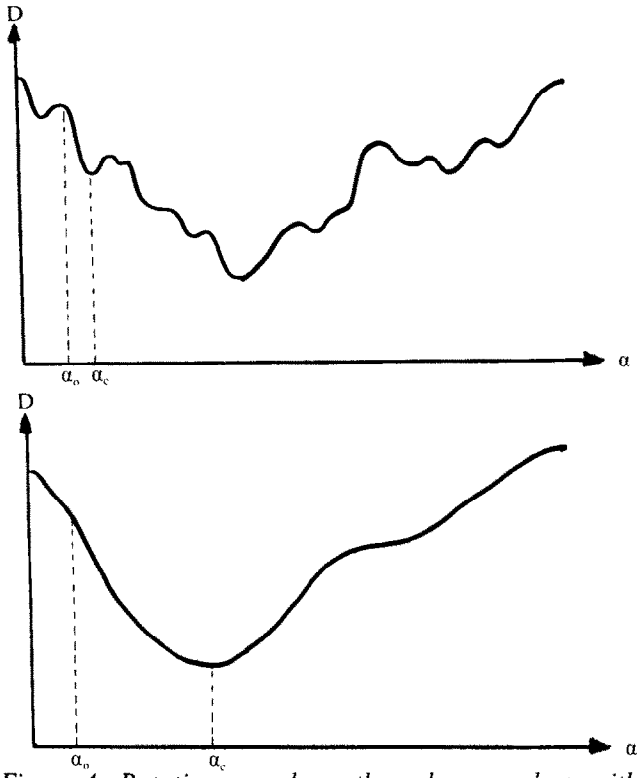

Figure 4. Rotating one shape through an angle α with respect to a second shape alters the value of the difference measure D. *Normalization means finding the angle* α_c *through which one shape must be rotated to give the global minimum of* D. *From initial value* α_0 *convergence will be more satisfactory for few descriptors (lower graph) than many (upper)*

the global minimum; hence, an absolute measure of shape difference can be obtained.

A noniterative alternative to this approach is to consider locating the best fit for low values of *l* for which there will be fewer turning points. The global minimum for low *l* will be close to that for higher values of l, so that with this approach automation may be possible (see Figure 4). Note that, considering the distance function as a surface for high *l* values, there will be many ripples on this surface. Close to the global minimum, such local changes in the value of the distance function may not be important, so convergence to any one of these local minimum should give a value reasonably close to the actual global minimum. This is clearly a problem concerning the level of detail that is important in shape complementarity for interacting systems.

DISCUSSION AND CONCLUSION

The technique of molecular shape description adapts an already successful technique of shape description and recognition for 2D contours. The advantage of this method is that shape description can be treated rigorously, thus allowing a quantitative measure of shape difference. Also, it is possible within this approach to change the level of detail to be considered by altering the number of descriptors to be used, a procedure that ultimately may lend itself to automated docking. The method describes a shape by building it up out of mathematically defined functions or shapes so that the descriptors measure the contribution from these various ideal shapes.

There are, however, a number of difficulties with this method of shape measurement. The first is that in the form described, it is limited to treating single valued functions. This problem has been solved within the current framework, but work is still required before our solution can be fully implemented. This will then allow us to treat the realistic problem of the solvent-accessible surface that in general will have reentrants producing a multivalued function when mapped to a sphere. A particular difficulty is that rather more coefficients are needed to obtain a similarly good representation to those shown in this work; consequently, greater computer time is required. An approach not yet explored may be to map a multivalued function of the polar coordinates onto a higher dimensional sphere so that in this way, for example, a multivalued function of two parameters would become a single function of three parameters. Again, this would require more computer time.

There is a second drawback. Our method is a global shape description method, so that from the shape descriptors of a particular surface it is not possible to obtain any information about local shape features. It is for this reason that, in order to get useful numbers for interacting molecules, the interacting interface of the two molecules must be obtained and defined before a comparison of the shape can be given. Regions of the molecules not involved in any interaction must therefore be left out of the shape description. It is also for this reason that this technique cannot be used to automatically locate complementary shape regions of an entire protein surface, say. However, other shape description methods[12] may be of use toward this aim. Once such regions are located, the Fourier descriptor technique would be of great value in giving an accurate quantitative measure of shape difference.

Although we have taken the Connolly molecular surface as the model of a molecule, there is no reason why it should be restricted to such a model. You could also take energy surfaces as the input model; in this way, you could obtain a measure of complementarity based also on the physical chemistry of the interactions. Of course, the shapes of molecules are constantly undergoing change and will certainly change during ligand binding interactions. This is the origin of current ideas on induced fit that supplant the lock and key concept.[1] However, the ideas described here should be fully developed first, since the method's usefulness has been demonstrated for 2D situations, and in principle there are no reasons why such a powerful technique should be limited to flatland.

ACKNOWLEDGEMENTS

We thank John Lee for useful discussions on the Fourier descriptor technique, Alwyn Jones for use of the com-

puter graphics facility at the Department of Molecular Biology, Uppsala University, and Mark Harris for supplying graphics display routines for the PS390. S.L. was supported by a CASE Research Studentship funded by the Science and Engineering Research Council and Pharmacia, Ltd.

REFERENCES

1 Fischer, E. "Einfluss der configuration auf die wirkung der enzyme." Berichte der Deutschen Chemischen Gesellschaft, 1984, **27**, 2985–2993

2 Etesami, F., and Vicker, J. J. "Automatic dimensional inspection of machine part cross-sections using Fourier analysis." *Computer Vision Graphics and Image Processing*, 1985, **29**, 216–247

3 Wallace, T. P., and Wintz, P. A. "An efficient three dimensional aircraft recognition algorithm using normalised Fourier descriptors." *Computat. Graph. and Image Proc.*, 1980, **13**, 99–126

4 Wallace, T., and Wintz, P. A. "Fourier descriptors for extraction of shape information" in *Image Understanding and Information Extraction*, Huang, T.S., and Fu, K.S. School of Electrical Engineering, Purdue Univ., Indiana, 47907, USA. TR-EE 77–35, September 1977

5 Profitt, D. "Normalisation of discrete planar objects." *Pattern Recognition*, 1982, **15**, 3, 137–143

6 Connolly, M. L. *Quantum Chemistry Program Exchange Bulletin*, 1981, **1**, 75

7 Bernstein, F. C., *et al.* "The protein data bank: a computer-based archive file for macromolecular structures." *J. Mol. Biol.*, 1977, **112**, 535–542

8 Richards, F. M. "Areas, volumes, packing and protein structure." *Ann. Rev. Biophys. and Bioeng.*, 1977, **6**, 151–176

9 Hobson, E. W. *The Theory of Spherical and Ellipsoidal Harmonics*. Cambridge University Press, 1931

10 Wiggins, R. A., and Masanori, S. "Evaluation of computational algorithms for the associated Legendre polynomials by interval analysis." *Bulletin of the Seismological Society of America*, 1971, **61**, 375–381

11 Cooley, J. W., and Tukey, J. W. "An algorithm for the machine computation of complex Fourier series." *Math. Comp.*, 1965, **19**, 297–300

12 Connolly, M. L. "Measurement of protein surface shape by solid angles." *J. Mol. Graph.*, 1986, **4**, 3–6