Measurement of protein surface shape by solid angles

M L Connolly

Department of Molecular Biology, Research Institute of Scripps Clinic, La Jolla, CA 92037, USA

A method for measuring the gross shape of local regions of protein surface is presented and applied to an examination of three proteins. Any point on the protein surface can be assigned a number that measures the degree of convexity or concavity of the surface in the vicinity of the point. This number is computed by centring a sphere at that point and measuring how much of the sphere lies inside the protein. The sphere radius is a parameter chosen according to the scale of the features that are being analysed. The amount of sphere intersecting the protein is interpreted as a solid angle, denoted omega. Three proteins are analysed by this method: lysozyme, superoxide dismutase and chymotrypsin. The resulting omega values are used to colour code the protein surfaces displayed on a colour raster graphics terminal. The method can be seen to reliably identify protrusions and depressions. The difficulty of developing a generally useful method for measuring protein surface shape is discussed. Possible applications of the solid-angle method include the analysis of shape complementarity at protein interfaces, the development of computer algorithms for predicting complexes between proteins, or between proteins and ligands, the identification of homologous epitopes on different proteins that might be immunologically crossreactive, and the determination of correlations between surface geometry and chemical properties.

Keywords: shape, protein, surface, raster graphics, algorithm

received 22 April 1985, accepted 20 May 1985

The function of proteins and the formation of protein aggregates depends upon the ability of proteins to recognize small molecules and each other. The specificity of molecular recognition is based on complementarity between the interacting molecules. This complementarity takes several forms; juxtaposition of groups with opposite charges, pairing of hydrogen bond donors and acceptors, opposition of hydrophobic regions of the protein surface and steric complementarity. The work described here develops a geometric, rather than a physical chemical, approach to evaluating the steric component of molecular recognition.

Although the human mind can intuitively identify topographical features of proteins by inspection of physical or computer graphical models, rigorous science requires the development of precise definitions and mathematical algorithms for measuring protein surface shape from crystallographic coordinates. This task is made difficult by the complex and varied nature of protein topography. Previous methods for analysing biological shapes[1] are able to compare only members of a family of homologues, and so are not appropriate for this application.

The method described below measures the concavity or convexity of protein surface regions. It is valuable because:

- it agrees with our intuitive perceptions of protein topography;
- it produces accurate numbers that vary smoothly over the protein surface;
- it is coordinate-system invariant; and
- it is general enough to be applied to objects other than molecules.

METHODS

The basic idea of the shape measurement method is illustrated in Figure 1 for the two-dimensional case of a contour in the plane. For a circle centred at a point on a convex part of the contour, only a small part of the circle lies inside the contour. For a circle centred at a point on a concave part of the contour, a large part of the circle lies inside the contour. The amount of the circle that lies inside the contour may be measured by its angle. By considering circles centred at all points on the contour, a function varying continuously between 0 and 2π may be defined. In fact a whole set of functions may be defined, one for each different circle radius. The radius defines the scale of the features that are being described.

For macromolecules, a similar approach is used to define a shape function. The circle is replaced by a sphere, the ordinary angle by a solid angle[2], and the contour by a closed surface. This work uses the solvent-accessible molecular surface[3-4] as computed by the AMS program[5-6]. This surface is well-suited to shape analysis because it is smooth, has both convex and concave regions, and precisely defines the boundary of the solvent-excluded volume.

A number quantitating shape is computed for a particular point on the protein surface as follows. First, a sphere of a given radius is created centred at that point (see Colour Plate 1). The radius of the sphere is chosen according to the scale of the topographical features that are being investigated. All the work described here used a 6 Å radius sphere, which identifies features involving a few amino acids.

 0263-7855/86/010003-04 $03.00

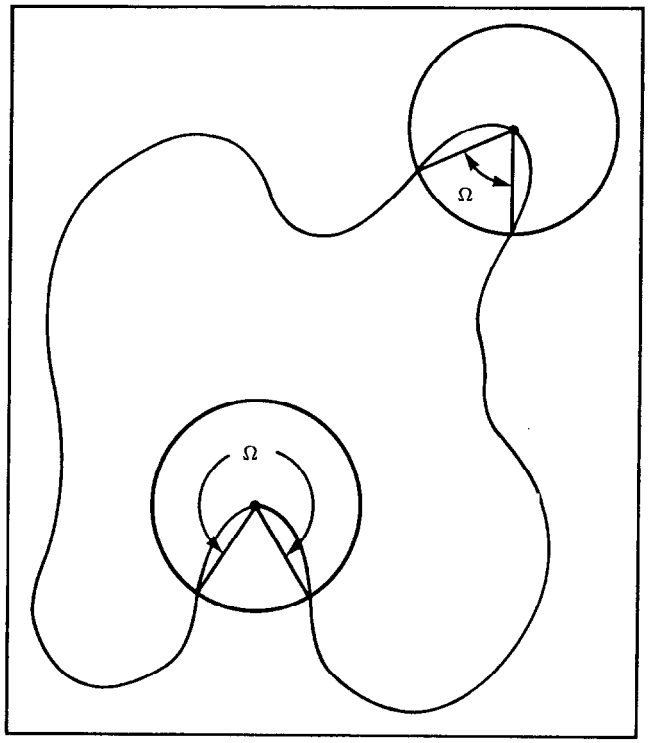

Figure 1. Two-dimensional representation of solid angle method; the fraction of the circle inside the contour is low for convex regions and high for concave regions

Next, the intersection contour of the sphere with the surface is calculated (see Colour plate 2). Usually the sphere-surface intersection consists of only a single contour, but sometimes there are several, depending on the complexity of the protein topography. Each contour is constructed as a chain of arcs and each arc is the intersection of the sphere with a face of the surface.

The calculation of the intersections of the sphere with the convex and concave faces is straightforward. These faces are parts of spheres, and the intersection of two spheres is a circle, so the intersection of the sphere with a convex or concave face is an arc of a circle. The intersections with saddle faces are more complex, because saddle faces are regions on tori, and the intersection of a sphere with a torus is a curve described by a pair of fourth degree equations. To simplify the problem, the intersection of a sphere with a saddle face is approximated by an arc of a circle.

The final step consists of measuring the area of the part of the sphere lying inside the protein. This area is computed from the geometry and topology of the sphere–surface intersection contours by applying the Gauss–Bonnet formula[7] in a manner analogous to its application in computing areas of convex spherical faces of the analytical molecular surface[6]. The area is divided by the sphere radius squared to give a solid angle in the range between 0 and 4π steradians, which is denoted by Ω. The equation for Ω, derived from the Gauss–Bonnet formula is:

$$\Omega = 2\pi\chi - \int k_g dl - \sum\delta$$

The Euler characteristic, χ, measuring the connectivity of the area, the geodesic curvature, k_g, in the line integral over the arcs, and the point curvatures, δ, resulting from small kinks where arcs join, are computed by formulae analogous to those presented in Ref 6. The solid angle omega distinguishes convex and concave regions of the protein surface.

Several omega values are calculated for each face, with sample point spacing of about 1 Å. Each face is colour-coded by its average omega value and the surface is displayed using the RAMS and MCS computer graphics programs[8]. Low omega values (convex regions) are coloured yellow, middle values, red and high values (concave regions), blue.

The computer program for computing omegas is written in FORTRAN and is run on a VAX-11/750 with floating point accelerator operating under VMS. The execution time is approximately proportional to the number of atoms. Lysozyme, which has about 1000 atoms took 20 hrs of CPU time. Surfaces were displayed on an AED 767 raster graphics terminal, which has a resolution of 575 by 768 pixels. Readers interested in acquiring a copy of the program should contact the author.

As a check on the analytical algorithm, a numerical procedure for calculating omega was developed. Although this method is slower and less accurate than the analytical algorithm, it has the advantage of being very simple to program. The method begins with a molecular dot surface computed by the MS program[9], which may be obtained from the QCPE*. This program produces not only the coordinates of points on the protein surface, but also an area and outward pointing normal vector associated with each point. A second dot surface is created for the intersection sphere, using the Genun subroutine of the MS program. Counting the number of points on the intersection sphere lying inside the protein surface gives a rough way of calculating omega (see Figure 2).

The only difficult part is determining whether a particular point lies inside the protein surface. Let M denote the number of points on the protein surface. Each surface point p_j has a unit normal vector $\boldsymbol{n}_j$ and an associated area a_j. Define a solid angle Θ at a point q by:

$$\Theta(q) = \sum_{j=1}^{M} \frac{(p_j - q)\cdot n_j}{|p_j - q|^3} a_j$$

It is shown in Ref 10 that this angle is approximately 4π when q is inside the surface and approximately 0 when it is outside the surface. Let a related function $I(q)$ be defined which has the value 1 when q is inside the protein surface, and 0 when it is outside. Using a solid angle cutoff of 2π steradians the following applies:

$$I(q) = \begin{cases} 1, \text{ if } \Theta(q) \geqslant 2\pi \\ 0, \text{ if } \Theta(q) < 2\pi \end{cases}$$

If N is the number of points on the intersection sphere, then omega is given by:

$$\Omega = \frac{4\pi}{N}\sum_{\kappa=1}^{N} I(q_\kappa)$$

When the analytical and numerical algorithms were tested on the same data, they produced omegas which differed by about 3%. This was with surface and sphere points spaced 1 Å apart. The accuracy of the numerical algorithm increases with more finely spaced points.

*The Quantum Chemistry Program Exchange is located at the Department of Chemistry. Indiana University, Bloomington 47405, USA. MS is program No. 429

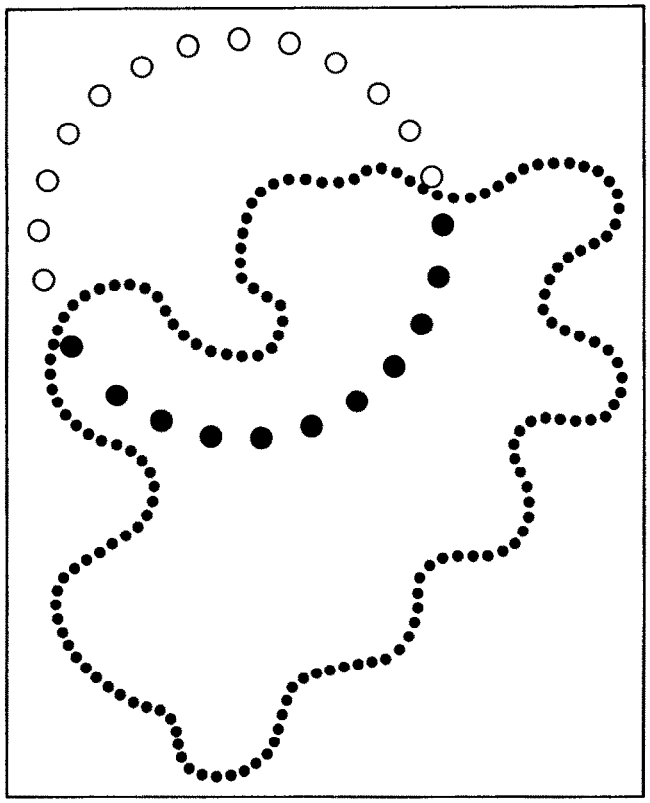

Figure 2. Two-dimensional representation of numerical algorithm for evaluating omega. The protein surface is represented as a dotted contour, the intersection sphere as a dotted circle. Each point of the intersection sphere is checked to determine whether it is inside the molecular dot surface. The circle's solid dots represent points in the protein interior, open dots represent points in the solvent

Protein coordinates were obtained from the protein data bank at Brookhaven National Laboratory[11]. A 1.5 Å radius probe was used and van der Waals radii were taken from Ref 12.

RESULTS AND DISCUSSION

Colour plates 3–6 show three proteins whose surfaces have been coloured according to the average omega value of each face. The computer algorithm produces results that agree with intuitions about what is a depression and what is a protrusion. The amino acid side chains sticking out into the solvent are consistently identified as being convex (orange and yellow) and the grooves, depressions and pockets are consistently identified as being concave (magenta and blue). The polysaccharide binding cleft of hen egg-white lysozyme[13] can be seen to contain a blue region indicating a concavity (Colour Plate 3). In fact, omega attains its maximum in this region. The superoxide binding pocket of superoxide dismutase[14–16] is blue and magenta (Colour Plate 4). Chymotrypsin's[17] binding pocket for bulky hydrophobic amino acid side chains is a large blue feature on its surface (Colour Plates 5–6). Kuntz *et al.*[18] observed that for the proteins they studied (myoglobin and prealbumin), the binding site was the largest invagination. By examining other views in addition to those shown in Colour Plates 3–7, this generalization has been seen to hold for lysozyme and chymotrypsin, but not for superoxide dismutase, which has a small substrate.

The development of quantitative methods for analysing protein topography is not as simple as it may seem at first glance. Methods for analysing the topography of the Earth depend on the fact that the planet is almost perfectly spherical and topographical features such as mountains and valleys are much smaller than the Earth itself. Locally, the Earth's surface is approximately flat, with a well-defined upward direction. Topographical maps may be made with contours for different elevations above sea level.

For a protein, however, the topographical features are not sufficiently smaller than the protein itself that a 'sea level' can be properly defined. The drawback with using z-coordinates as a measure of protein topography is that there is a dependence on the coordinate system used. For a protein there is no globally defined upward direction. Even locally there is substantial ambiguity about which way is up. Even if an upward direction could be defined, there would still be the problem of defining where $z = 0$. The heights of topographical features would be defined only relative to each other, not absolutely. A spherical polar coordinate system, where height is measured radially from the centroid of the molecule, suffers from similar disadvantages (Figure 3). For an approximately flat interface, the haemoglobin $\alpha_1\beta_1$ interface, a measurement of protein surface topography based on a global rectangular coordinate system has proved to be useful for certain applications[19]. Nevertheless, a generally useful method for characterizing protein surface shape should be coordinate-system invariant.

Another consideration is that of the size of the features one is seeking to recognize. For atomic-sized features, the solvent-accessible surface gives direct information on topography, because it consists of convex, concave and saddle faces of atomic size. It is not practical to identify larger-scale features by calculating a surface using a large probe sphere for several reasons:

- most concavities are not well-approximated by the underside of a large sphere,
- most ligands larger than a water molecule are not well-described by a sphere;
- the convex and concave faces would have substantially different radii; and
- the surface itself has been changed in the attempt to measure its shape.

The solid angle method presented in this work is coordinate-system invariant, and the scale of the features it recognizes may be determined simply by adjusting the intersection sphere radius. While the technical details of the solid-angle calculation are dependent on the object's surface being described as a molecular surface, the basic idea of intersecting a sphere with a solid object and measuring the area of the sphere contained within the object is not limited to chemical applications. In principle, any physical object whose surface could be mathematically defined could have its shape characterized by the solid-angle approach.

There are many biochemical applications where a protein shape analysis method would be useful. For structures of protein–protein complexes or multimeric proteins which have been solved, the molecular inter-

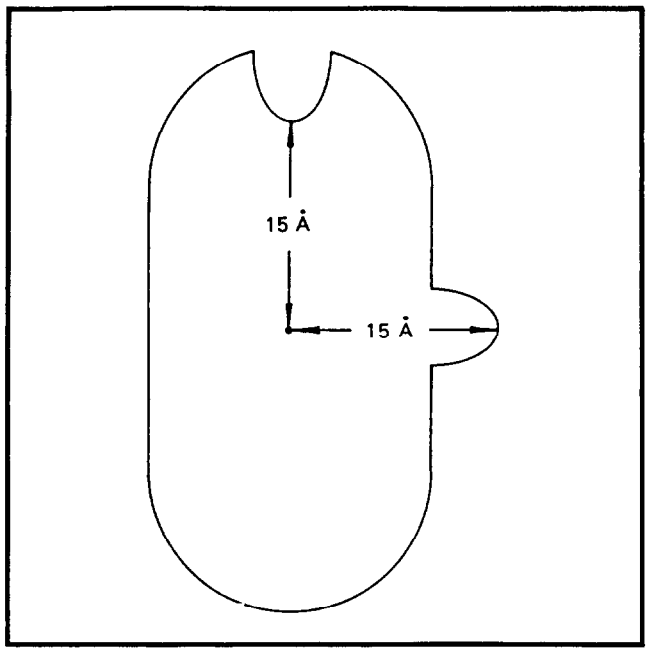

Figure 3. Two-dimensional representation of a protein surface. In a spherical polar coordinate system, each point of the surface is assigned an r *value equal to its distance from the protein centre of mass. If the protein is even moderately oblong, the bottom of a pocket and the top of a protrusion may be assigned the same value of* r

faces could be analysed. Solid angles computed for opposing surfaces should have complementary values, summing to approximately 4π. This is a necessary (but not a sufficient) condition for a close fit. Omega maxima on one surface should correspond to omega minima on the other. If this proves to be the case, it may be possible to develop an automatic algorithm for predicting protein–protein associations based upon docking knobs into holes.

Regions assigned the same omega value by the solid-angle method may differ in shape in more subtle ways. A future direction would be to develop methods for describing the shape of a protein surface region that would produce not just a single number, but a set of numbers that would uniquely identify each shape. The simplest way to compute several numbers is to compute omega for several different radii. If omega is regarded as a function of radius, then the integral $\int_0^R \Omega r^2 dr$ gives the volume of the part of the protein that lies inside a sphere of radius R. This local volume reflects the shape of all the local surface, not just that located at a particular radius.

A sophisticated set of shape measurement functions could be used to categorize all surface regions of all proteins of known three-dimensional structure into a comprehensive taxonomy. This taxonomy could be used to identify surface regions in different proteins with homologous shapes, indicating that the proteins might be cross-reactive immunologically. Of course, antigen shape is just one component of recognition by the immune system.

The solid-angle method for quantitating molecular surface shape should make possible correlations with a number of biological and chemical properties: charge, electrostatic potential, hydrophobicity, accessibility to various ligands, antigenicity and evolutionary variation.

ACKNOWLEDGMENTS

The author would like to thank Art Olson for the use of his computer graphics facility, and John Tainer and Elizabeth Getzoff for their useful comments on the manuscript. This work was supported in part by National Institutes of Health Grant GM 34338 and by the Helen Hay Whitney Foundation.

REFERENCES

1 **Bookstein, F L** *J. Theor. Biol.* Vol 107 (1984) pp 475–520
2 **Gellert, W, Küstner, H, Hellwich, M and Kästner, H** *The Van Nostrand-Reinhold Concise Encyclopedia of Mathematics* Van Nostrand-Reinhold, New York (1975) p 186
3 **Lee, B and Richards, F M** *J. Mol. Biol.* Vol 55 (1971) pp 379–400
4 **Richards, F M** *Annu. Rev. Biophys. Bioeng.* Vol 6 (1977) pp 151–176
5 **Connolly, M L** *Science* Vol 221 (1983) pp 709–713
6 **Connolly, M L** *J. Appl. Crystallogr.* Vol 16 (1983) pp 548–558
7 **Carmo, M P do** *Differential Geometry of Curves and Surfaces* Prentice-Hall, Englewood Cliffs, New Jersey (1976) pp 264–283
8 **Connolly, M L** *J. Mol. Graphics* Vol 3 (1981) pp 19–24
9 **Connolly, M L** *QCPE Bull.* Vol 1 (1981) p 75
10 **Connolly, M L** *Ph.D Thesis,* University of California at Berkeley, (1981) pp 44–46
11 **Bernstein, F C, Koetzle, T F, Williams, G J B, Meyer, Jr., E F, Brice, M D, Rodgers, J R, Kennard, O, Shimanouchi, T and Tasumi, M** *J. Mol. Biol.* Vol 112 (1977) pp 535–542
12 **McCammon, J A, Wolynes, P G and Karplus, M** *Biochem.* Vol 18 (1979) pp 927–942
13 **Diamond, R** *J. Mol. Biol.* Vol 82 (1974) pp 371–391
14 **Tainer, J A, Getzoff, E D, Beem, K M, Richardson, J S and Richardson, D C** *J. Mol. Biol.* Vol 160 (1982) pp 181–217
15 **Tainer, J A, Getzoff, E D, Richardson, J S and Richardson, D C** *Nature* Vol 306 (1983) pp 284–287
16 **Getzoff, E D, Tainer, J A, Weiner, P K, Kollman, P A, Richardson, J S and Richardson, D C** *Nature* Vol 306 (1983) pp 287–290
17 **Cohen, G H, Silverton, E W and Davies, D R** *J. Mol. Biol.* Vol 148 (1981) pp 449–479
18 **Kuntz, I D, Blaney, J M, Oatley, S J, Langridge, R and Ferrin, T E** *J. Mol. Biol.* Vol 161 (1982) pp 269–288
19 **Greer, J and Bush, B L** *Proc. Natl. Acad, Sci. USA* Vol 75 (1978) pp 303–307