Molecular volumes and surfaces of biomacromolecules via GEPOL: A fast and efficient algorithm

E. Silla and F. Villar

Departamento de Química Física, Universidad de Valencia, Burjasot (Valencia), Spain

O. Nilsson, J. L. Pascual-Ahuir and O. Tapia

Department of Physical Chemistry, Uppsala University, Uppsala, Sweden

A triangular tesselation approach to build up surfaces has been adapted to the study of biomolecules. By using a data-coded generic pentakisdodecahedron each atom is assigned a particular sphere whose radii are chosen according to any suitable property. Different types of surfaces have been adapted to this method: van der Waals, surface accessible, and Richard's molecular surface. A simple method is used to eliminate all triangles found at the intersection volume of the atomic spheres and a fast algorithm is employed to calculate the area of the envelope surface and the volume therein. The data about the surface are given by the coordinates of the center of each triangle, elementary surface value, and vector coordinates of the normal to the surface. Color coding of standard properties such as charge densities, potential energy, or any scalar property can be easily done with standard graphics libraries. Fairly detailed information on vector properties, *such as electric field and atom velocity, can also be graphically represented by using projections along the normals with adequate color coding.*

Keywords: accessible surface, van der Waal surfaces, molecular volume, fractal indeces, Molecular Dynamics data display

INTRODUCTION

Characterization of biomolecular surfaces is an important problem in molecular biology. Accurate determinations of these surfaces may help us to understand the relationships between structure and function.[1] In the early 1970s, Lee and Richards proposed a solvent-accessible model surface (SAS).[2,3] In the 1980s, a number of algorithms have been proposed for estimating the area and volume of molecules.[2-11] Among them, Connolly's molecular surface (MS) approach has been widely used. None of these methods has yet provided a fully satisfactory solution for this problem. In particular, Connolly's method has been criticized for oscillatory behavior with respect to its internal parameters.[1] In this paper, an algorithm based on sphere tesselation which was designed to get surface and volume of simple molecules[13,14] is extended and graphically implemented to calculate biomolecular surfaces.

The tesselation method was coded in program GEPOL and is available from QCPE.[15] The basic procedure has been previously reported.[16] Here, we describe those complementary algorithms used for (1) selecting the portions of the spherical surfaces that form the envelope surface, (2) determining of the coordinates of tessels, and (3) calculating the actual molecular area and volume.

The results obtained are compared with the MS (so-called MSDOT) scheme.[17] Special attention is given to the accuracy of the tesselation procedure; this is gauged with simple models whose surface and volume can be exactly determined. Spatial invariance of the algorithm is numerically tested for various cases. Computer efficiency measured by CPU time has been checked.

A Fortran program has been written by using the Silicon Graphics Personal Iris graphics library to display the surface types calculated by tesselation. GEPOL may produce different types of surfaces. Van der Waals surfaces (VWS) and SAS are available. Richards MS is also implemented. This latter method excludes all molecular surfaces that are not accessible to a spherical solvent molecular probe. The data used for display consist of triangle vertex and the coordinates of the normal vector to each tesselar surface. To

Address reprint requests to Dr. Tapia at the Department of Physical Chemistry, Uppsala University, Box 532, 75121 Uppsala, Sweden.
Received 19 April 1990; accepted 1 May 1990

Dr. Pascual-Ahuir's present address is the Department of Physiology and Biophysics, Mount Sinai School of Medicine, New York, NY 10029, USA.

illustrate the power of this procedure, the surface information is used to display the atom velocity projection along each normal and color-coded inward or outward components; the velocities are obtained from molecular dynamics simulations. This is a unique device to communicate kinetic and dynamical information in a molecular graphics display manner.

THE ENVELOPE SURFACE

The model starts by setting on each atom a sphere. The radius is chosen according to the atom type. van der Waals radii are used here, but any other type of data can be entered in the program. When looking for the equivalent of Richards MS, the solvent-unaccessible surface is eliminated by creating new spheres between those initially defining the model.[15,16]

Each sphere is tesselated with 60 spherical triangles, all equal in area. To do this a pentakis dodecahedron is projected onto each of the previously generated spheres (including smoothing spheres). This division may be too coarse-grained, in which case we may increase the tesselation; thus, each triangle may be subdivided in 4, 16, etc. smaller triangles. In Figure 1 we illustrate the result of increasing the tesselation parameter (NDIV in the figure and FORTRAN code). For two overlapping spheres, the finer the tesselation, the less the calculated surface is hidden by the overlap, since it is obtained by summing up all triangles whose centers are not in the intersection volume.

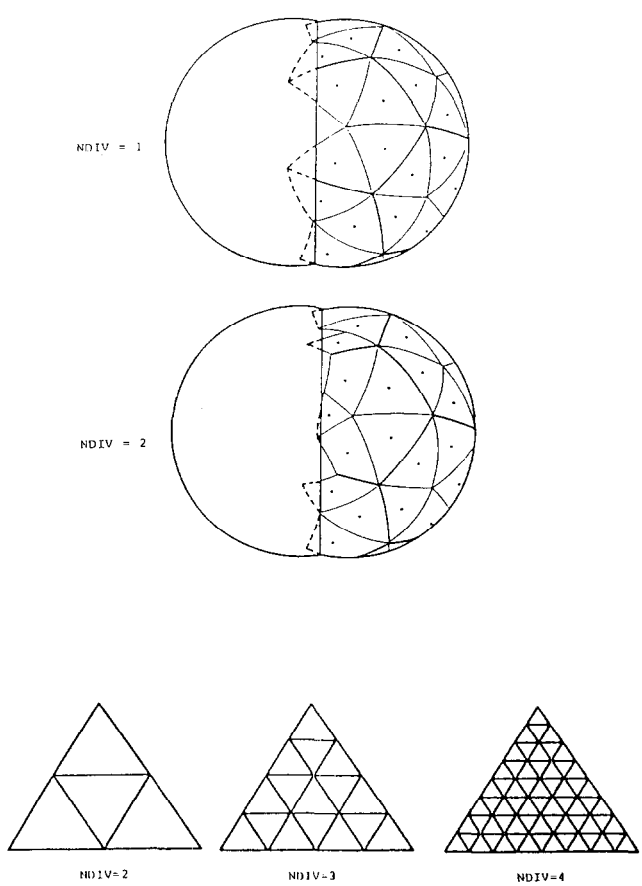

Figure 1. Model of two intersecting spheres. On the right hemisphere the remaining triangles are indicated. Those having a part into the overlap region are drawn with the region inside the overlap in broken lines. It can be seen how the fine-graining technique (NDIV = 2) improves the boundary between the intersecting spheres. Once the boundary has been refined at a given degree of granularity to avoid an excessive number of tessels, the algorithm makes it possible to come back to the initial triangles with appropriate surfaces. Note that after this some tessels no longer have triangular shape

Triangle division

To divide a spherical triangle into four (see Figure 1, NDIV = 2, for instance) the coordinates of the middle point along the arc are needed. Let $\mathbf{x}_1$ and $\mathbf{x}_2$ be the coordinates of two vertices, and $\mathbf{x}_m$ the one corresponding to the midpoint (on the surface). Calculate the midpoint of the chord (straight line passing by $\mathbf{x}_1$ and $\mathbf{x}_2$) $\mathbf{x}_c$. Now let $\mathbf{R}_c$ be the radius from the origin of the sphere to $\mathbf{x}_c$ and $\mathbf{R}$ its radius. The coordinates of the midpoint on the spherical surface are the components of the vector: $\mathbf{x}_m = \mathbf{x}_c\,(|\mathbf{R}|/|\mathbf{R}_c|)$.

Coordinates of the triangle center

The procedure requires a knowledge of the triangle center and the normal vector. First, calculate the coordinates of the barycenter formed by the three vertices: $\mathbf{R}_b = (\mathbf{x}_1 + \mathbf{x}_2 + \mathbf{x}_3)/2$; the coordinates of the triangle center on the spherical surface are the components of the vector: $\mathbf{x}_m = \mathbf{x}_b\,(|\mathbf{R}|/|\mathbf{R}_b|)$. The vector normal to the triangle surface is calculated and stored with the corresponding coordinate center; positive direction is toward the center of its sphere.

Once the tesselation has been achieved at a given level of granularity (NDIV) a set of triangle center and normal vectors (Fig. 1) is obtained. If the distance between the center of a triangle and the origin of another sphere is less or equal to the radius of the sphere where it belongs, the triangle is discarded. The envelope area is easily obtained by summing up the areas of all remaining triangles.

Volume

Let us define a molecular origin and let $\mathbf{R}_i$ be the position of the ith triangle center and $\mathbf{n}_i$ the corresponding normal vector. The volume is obtained by summing up all solid volumes made by the triangles vector surfaces and the origin of the molecular system: $V = \frac{1}{3}\,\Sigma_i S_i \mathbf{n}_i \cdot \mathbf{R}_i$; where S_i is the surface of the ith triangle. It is easy to see that internal volumes are canceled out.

EXACT TEST CASE: TEN LINEARLY ARRANGED SPHERES

Surface

The system consists of 10 spheres, each of radius 1.8 Å, and 1.5-Å nearest-neighbor center-to-center distance. The granularity parameter NDIV was varied from 1 to 5. Comparisons with MSDOT program varying the point density parameter from 1 to 25 are made as a function of CPU time on a Microvax 2000.

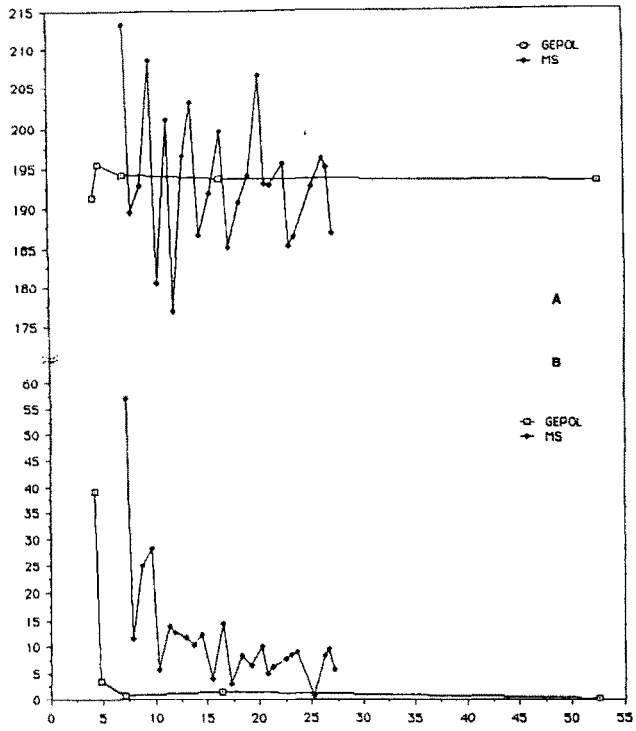

Figure 2. Ten-sphere surface. (a) In ordinates the area in arbitrary units, in abscissa the CPU time (seconds) of a MicroVax 2000 required for the actual calculation with different values of granularity. MS stands for MSDOT in the text. (b) In ordinates the rms deviation in area for the same abscissa

The surface is compared in Figure 2a. GEPOL results converge fairly fast, while MSDOT outcomes oscillate significantly. The relative error (not shown) for GEPOL is never larger than 1%, while MSDOT goes up to 10% with large downward fluctuations.

Spatial invariance

Quite generally, the algorithms are dependent on the orientation of the reference frame to a more or less acute degree. To test the invariance level, the area is calculated with three orthogonal orientations. The root mean square (rms) deviation is calculated and shown in Figure 2b. The result for NDIV = 1 has a rather large rms deviation, and a clear improvement is seen for NDIVs greater than one. MSDOT has a poor behavior. These results confirm the oscillatory nature of MSDOT.[1]

Volume

Another powerful feature with GEPOL can be seen with the calculation of volume. In Table 1 the average over three different orientations of the reference system for the 10-sphere volume is given as a function of granularity (NDIV). The result with NDIV = 1 is apparently good, but the rms is the largest by a factor of 10 compared to the others. The volume for NDIV = 2–5 converges from above to the exact value with decreasing error and rms. Note that for

Table 1. Volume of 10 spheres calculated with GEPOL as a function of granularity[a]

NDIV	Volume	Deviation%	Rms
1	152.96	0.60	19.28
2	155.50	1.05	1.70
3	155.03	0.74	0.39
4	154.85	0.62	0.68
5	154.67	0.51	0.14
Exact volume = 153.890			

[a]The exact value is also indicated. The units are arbitrary

NDIV = 1 the volume is smaller than the exact volume, which suggests that this level of granularity is not the best choice to describe the volume of the system. This and other examples allow us to conclude that GEPOL yields fairly good values for the volume embodied by the envelope surface for NDIV larger than one.

BIOMOLECULES

The calculation of molecular surfaces and volume presents no particular problems when large molecules such as proteins are used. To illustrate the quality of the present method a comparison is made with the MSDOT program. Two systems examined are (1) a 185-atom section of the coenzyme binding domain of liver alcohol dehydrogenase, which includes the mobile loop made by residues 294–298[18]; (2) the carboxyterminal fragment of the L7/L12 ribosomal pro-

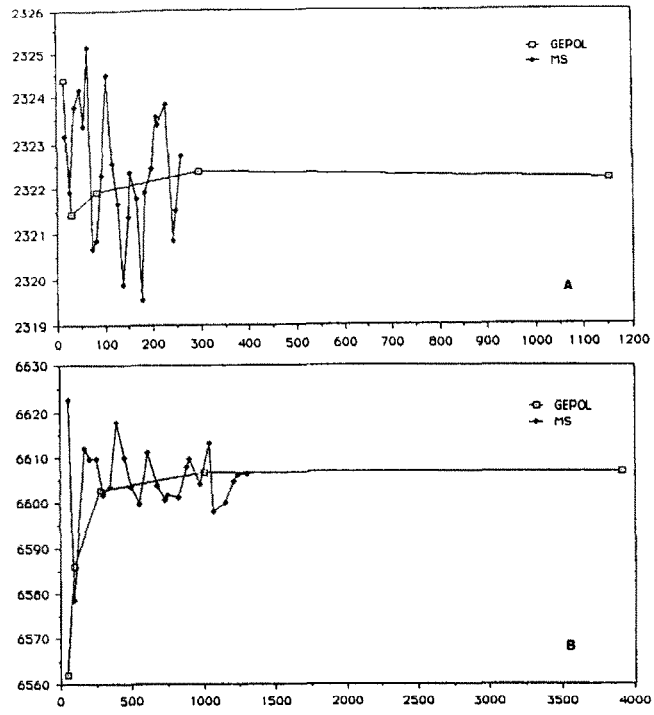

Figure 3. Area as a function of granularity and CPU time. (a) Corresponds to the liver alcohol dehydrogenase coenzyme domain loop 294–298; (b) describes the results for CTF

tein from *E. coli*.[19] These systems are presently subjected to molecular dynamics studies. The first illustrates a fairly elongated section and is basically used to test the reference frame orientation invariance of our method; the second system represents a well-folded protein.

In Figure 3a the area is plotted as a function of CPU time. Each point is calculated with the same set of parameters as in the 10-sphere case. A similar plot for CTF is presented in Figure 3b. As was found for the 10 spheres, the MSDOT method tends to oscillate, while GEPOL displays a smooth variation. The amplitudes of the fluctuation for MSDOT decrease with increase of the size. The relative error made with MSDOT is larger than with GEPOL for LADH-loop. With NDIV = 2, GEPOL is within a 2% error only. For the larger system reported here both methods have fairly low relative errors as one increases granularity. This result has consistently been found for larger proteins, such as retinol binding protein (RBP).

Since most of interesting properties depend upon local surfaces and not on the global one, it is important to realize that local surfaces are more accurately calculated with GEPOL than with MSDOT. This is valid for all molecular sizes.

Fractal index calculation

In view of the smoothness shown by GEPOL, this method is better adapted to determine the variation of molecular surface as a function of the probe size. In Figure 4 a log–log plot for RBP is presented. A linear relationship holds in the probe radius range from 1.41 Å up to 2.5 Å, which is in good agreement with a previous study.[20] As the probe radius decreases below 1.26 Å, internal surfaces become accessible until one gets into the range of 0.5 Å down to 0.25 Å, where the area changes only marginally. In this region we are consistently calculating the limiting value of the surface compatible with the algorithm. The slope of the linear regime κ is related to the topological dimension $D_t = 2$ and the fractal dimension D by the relationship $\kappa = D_t - D$. In this probe radius range, the fractal dimension and the topological dimension are equal, i.e., the area does not depend upon the size probe. From Figure 4 it follows that this fractality concept[21] is well defined (a linear relationship) in the standard range of probe radius. Therefore, one has to be careful when applying numerical procedures designed to obtain molecular surfaces and test for the linear region in the log–log plot. The fractal dimension of the apo-RBP surface is 2.54, which is larger than for holo-RBP (2.18)[20] as one would expect.

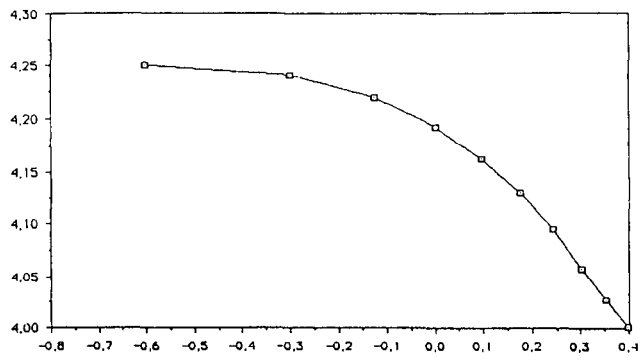

Figure 4. Log–Log plot of area as a function of probe radius for retinol binding protein in apo form

From the studies reported above, it can be concluded that the method used here offers an accurate, stable, spatially invariant, and rather fast algorithm to obtain both surfaces and volume of molecules.

GRAPHICS DISPLAY

The basic information obtained from GEPOL concerns triangles (spherical or planes), which makes it easy to use with present graphics display workstations.

In Color Plate 1a the van der Waals surface of the heme molecule is depicted. NDIV has been taken equal to 2. The planar triangles built up from the vertex of the spherical ones are displayed using the standard graphics subroutines from the Silicon Graphics workstations. Z-Buffer hidden-surface elimination is used. The polyhedral representation is apparent. This feature can be hidden by using Gouraud's shadowing method[22]; the shading discontinuities near the triangle edges disappear with this technique. The polyhedral representation is still underneath; the edges are still made of straight lines. If NDIV is increased, the borders become smoother; however, this procedure might be too CPU-time demanding in computation and graphics display resources, and is not very rewarding. In Color Plate 1b a tilted view of the heme is shown where NDIV = 2 and shadowing is switched on. A fairly good picture is obtained.

In Color Plate 2a and b two views of the backbone of the αA helix of CTF are shown: In Color Plate 2a shadowing is off; in Color Plate 2b it is on. Top left in Color Plate 2a there are two carbonyl oxygens. The second makes part of a peptide unit whose amino group is making an H bond with a carbonyl oxygen on the upper ridge of the picture. Descending from the previous position one can see the presence of two rows of H bonds. In the second view the molecular fragment was rotated around the helix axis (almost) and the H-bonding pattern is clearly visible.

Graphical representation of molecular dynamics data is still in the early stages of development insofar as "dynamics" properties are concerned. Recently, interactive graphical analysis of conformational space has been outlined by one of us.[23] Here, a simple way to convey vector properties is suggested. Among vectorial observables there is the atom velocity obtained from a MD simulation. By calculating the scalar product of the normal to a given triangle with the corresponding atom velocity, the direction of motion across the tessels (a velocity "flux") can be color coded according to the sign of the scalar product. This projects an instantaneous view of the atomic motions. In Color Plate 3a atom velocities for the helix studied above are represented. For the green color, the atom is moving toward the center of the sphere and for the red color it is moving in the opposite direction. A gray color means that the velocity of the atom is perpendicular to the normal. The positions of the atoms correspond to those shown in Color Plate 2b. A red arrow is pointing toward the hydrogen atom in the H-bond system found in the upper right of the previous plate. Note that the motion in the bridge is antisymmetric in the sense that both

proton donor and acceptor are moving with the same phase (color), while hydrogen has the opposite direction (red color). The hydrogen atom in the H bridge below the preceding one is approximately moving away from the helical axis, since the ball has all three colors. The α-carbon just above the first H bridge is moving toward the axis.

It is, of course, difficult to discuss in detail the motion of an alpha helix or any other secondary structure from a single static picture. In this sense, the use of a graphics display workstation is essential. A great deal can be learned from careful analysis of a vectorial pattern.

A sort of kinetic energy flux can also be defined by taking the square of the projected velocity multiplied mass surface density (mass of the atom divided by the number of triangles). Using the same color coding introduced for the alpha helix, in Color Plate 2b we present the backbone of CTF. Although, it is difficult to get the idea of motion from a snapshot, it should be noticed that when working at the screen, the picture can be rotated and some of the concerted motions emerge from such scrutiny. In the present case the first feeling one gets by looking at different frames is best described as stochastic-like motion. Nevertheless, attentive observation shows regions of local concerted motions.

With the advent of high-performance graphical workstations that can handle more than one million vectors and polygons, the surface method presented here offers a significant improvement in analyzing molecular shape fluctuations, volume fluctuations, and solute–solvent interactions. It also provides a simple model for parameterizing complex stochastic interactions by using the discrete representation of molecular surface.

NOTE

For the academic community, the computer program can be obtained directly from us by electronic mail: ORLANDO@SEMAX51.

ACKNOWLEDGMENTS

O.T. is most grateful to NFR for financing this work and to the University of Valencia for inviting him to Spain for a short-term visit where this work was completed.

REFERENCES

1 Meyer, A. *Chem. Soc. Rev.* 1986, **15**, 449, and references given therein
2 Lee, B. and Richards, F. M. *J. Mol. Biol.* 1971, **55**, 379
3 Richards, F. M. *Annu. Rev. Biophys. Bioeng.* 1977, **6**, 151
4 Hermann, R. B. *J. Phys. Chem.* 1972, **76**, 2754
5 Miertus, S., Scrocco, E. and Tomasi, J. *Chem. Phys.* 1981, **55**, 117
6 Connolly, M. L. *J. Appl. Crystallogr.* 1983, **16**, 548
7 Connolly, M. L. *Science* 1983, **221**, 709
8 Connolly, M. L. *J. Am. Chem. Soc.* 1985, **107**, 1118
9 Meyer, A. Y. *J. Chem. Soc. Perkin Trans. 2* 1985, 1161
10 Meyer, A. *J. Comput. Chem.* 1986, **7**, 144
11 Drummond, M. L. J. *J. Chem. Phys.* 1988, **88**, 5021
12 Voorinthold, R., Kosters, M. T., Vegter, G., Vriend, G. and Hold, W. G. H. *J. Mol. Graphics* 1989, **7**, 243
13 Pascual-Ahuir, J. L., Silla, E., Tomasi, J. and Bonacorsi, R. *J. Comput. Chem.* 1987, **8**, 778
14 Pascual-Ahuir, J. L. and Silla, E. Quantum chemistry—Basic aspects, actual trends. In *Studies in Physical and Theoretical Chemistry,* (R. Carbó, Ed.), Vol 62. Elsevier, Amsterdam, 1989, p. 597
15 Program GEPOL can be obtained from Quantum Chemistry Program Exchange, Indiana University, Bloomington, IN 47405. Program 554, QCPE Bull (1988)
16 Pascual-Ahuir, J. L. and Silla, E. *J. Comput. Chem.* submitted
17 Connolly, M. L. Pgm. No 75, *QCPE Bull.* 1981
18 Tapia, O., Eklund, H. and Branden, C. I. *Esteric Aspects of Biomolecular Interactions* (G. Naray-Szabo and K. Simon, Eds.). CRC Press, Boca Raton, FL, 1986
19 Aquist, J., van Gunsteren, W. F., Leijonmark, M. and Tapia, O. *J. Mol. Biol.* 1985, **183**, 461
20 Aquist, J. and Tapia, O. *J. Mol. Graphics* 1987, **5**, 30
21 Peitgen, H.-O. and Saupe, D. (Eds.) *The Science of Fractal Images.* Springer, New York, 1988
22 Gouraud, H. *IEEE Trans. Comput.* 1971, **C-20**, 623–628
23 Nilsson, O. *J. Mol. Graphics,* submitted
24 Tapia, O., Nilsson, O., Campillo, M., Åqvist, J. and Horjales, E. In *Structure and Methods. DNA Protein Complexes,* Vol. 2 (R. H. Sarma and M. H. Sarma, Eds.). Adenine Press, New York, 1990, 147–170