New determinations and simplified representations of macromolecular surfaces

G. Perrot and B. Maigret

Laboratoire de R.M.N. et de Modélisation Moléculaire, Institut Lebel, Université L. Pasteur, Strasbourg, France

Several new methods or improvements of older algorithms determining the different pieces of molecular surface are presented. Their improvement in time and their complexity are discussed. Only the indexes of the atoms on which the pieces are relying are, in fact, determined, since their explicit representation from these numbers varies according to the 3D capabilities of the graphic workstation (dots, grid, etc.), and this generation is not C.P.U. consuming. To have a simplified representation of the surface of macromolecules, a polyhedron with planar triangular faces is then introduced: Each concave triangular surface piece is replaced with planar triangles relying on its three atomic centers, while saddle-shaped rectangles and convex pieces are wholly ignored. A minimal data structure of the polyhedron is then proposed, which contains only topological informations, since no coordinates have been generated. If the atomic radius is then considered to be constant (independent of atomic type), the surface of a set of N points is now defined by the choice of a subset with a topology. This choice is controlled by a parameter of rugosity (the atomic radius). Contrary to Voronoi polyhedrons partition, which gives a topology for a set of N points, our approach gives a topology only for the exterior points of this set. A few applications of this very simple definition of molecular surface are then discussed: the 3D interactive manipulation of macromolecules, the steric intermolecular recognition, and the determination of local and global properties of the surface.

Keywords: algorithm, surface topology, 3D interactive visualization, shape recognition

INTRODUCTION

As larger and larger molecular systems are investigated, the reduction of the cost of molecular representation is a necessary goal. This cost is mainly due to the determination of the different molecular surface pieces (main processor time, MPT), and to an important amount of generated data (graphic processor time, GPT). With only a few approximations, a lot of MPT can be saved, as shown under Methods.

Saving of GPT requires a few explanations. The molecular surface has already been described very accurately (see Ref. 1 for its analytical determination). Such a precise definition of surface is not always necessary for macromolecules such as proteins. Besides, even if a precise description of the active site of a macromolecule is necessary, it is important to have a rough representation of the remaining parts. To have efficient shape recognition and 3D interactive raster visualization algorithms, it is attractive to reduce the number of surface points and to connect them in a grid as regular as possible in terms of vertices order, faces order, and faces size. A quasi-regular triangular grid has already been generated.[2] When trying to minimize the cost of the algebric determination of molecular surface, a very simple grid appears naturally. It is not, by far, as regular as the previous one, but can still give a lot of information about the surface, and is much simpler to determine: It does not need any generation of points!

The present study gives a useful molecular representation which can be obtained with a limited amount of CPU time, at low memory and disk storage cost. Besides, it is easy to use this data structure to recover any part of the molecule and visualize it at any rate of precision, provided that removal of self-intersecting surfaces is not needed (see Ref. 1). Opposite to the description of surface by a set of calculated points, the proposed data structure contains only topological information, which is essential to completely describe a surface.

METHODS

1. Determination of the surface

Let us analyze briefly the molecular surface determination. For small proteins (150 residues), the determination of molecular surface is rather time-consuming when compared to van der Waals surface calculation. During the determination of van der Waals surface, only collisions between bound atoms have to be considered, and the topology is the connectivity of the molecule. Otherwise, it is necessary to build a list of neighboring atoms (LNA), which usually costs $O(n^2)$ (n, total number of atoms). Then, the determination of the concave triangles (DCT) costs $O(n * k^3)$ (k, average number of neighbors per atom). Indeed, it is necessary, for

Address reprint requests to Dr. Perrot at the Laboratoire de R.M.N. et de Modélisation Moléculaire, Institut Lebel, Université L. Pasteur, 4 rue Blaise Pascal, 67000 Strasbourg, France.
Received 13 March 1990; accepted 27 March 1990

each triplet (I, J, K) of neighboring atoms, to test if the lying probe sphere intersects a neighboring atom L. The determination of the saddle-shaped rectangles and of the convex pieces are immediate, since they are adjacent to concave triangles except for entire tori that are neglected, because they can be considered as one-dimensional objects, and appear rarely at usual radii.

The total complexity seems therefore to be $O(n^2)$, and one could expect this article to present new methods for LNA determination, which seems to be the most time-consuming step.

But LNA cost can be easily reduced (see Ref. 3, for example), for a small protein, building LNA is about ten times faster than DCT, and k increases with atomic radii and probe sphere radius. Let us consider all atomic radii (including probe radius) to be equal, to simplify our discussion. The surface now depends on R, the constant radius, and the n given centers. Thus, for $R = 1.5$ Å, $k \approx 40$ for a small protein, which should not be too far from its asymptotic value. This means that for $n = 2000$, $n^2 << n * k^3$. Besides, if we decide to increase R, we have $k \approx 200$ for $R = 3$ Å!

That is why our interest focused on DCT. To have reasonable time costs for the requested goal, we suggest a few possible modifications of the DCT:

(1) Sorting the neighbors speeds up the intersection search considerably. Proximity tables other than the LNA do not seem necessary.

(2) A seed method: If a triplet (I, J, K) is known, a new triplet can be found easily by exploring the possible triplets (I, J, L) classically, which leads to an $O(n * k^2)$ algorithm, or by sorting according to the $[(I, J, K), (I, J, L)]$ angles, which leads to an $O(n * k)$ algorithm (provided the polyhedron generated by all valid triplets does not intersect itself, which is the case with the previous approximation on radii). A seed for the exterior surface can easily be found, if the interior surfaces of a molecule (defects) are not needed. If a previous surface is known near the one to be studied (which is the case during molecular dynamic or Monte Carlo simulations), seeds for every connexe pieces can also be found.

(3) The last method is derived from a Voronoi approach and from accessible surface definition. Instead of looking for concave triangles, we could look for convex pieces first. Let us assign at each atom I, a sphere $S(I, 2 * R)$ centered at I, with a radius of $2 * R$. Let Λ be the LNA of I ($J \in \Lambda \Leftrightarrow S(I, 2 * R) \cap S(J, 2 * R) \neq \emptyset$). If $J \in \Lambda$, let P_{IJ} be the plane perpendicular to IJ and containing $(I + J)/2$ and P_{IJ}^{+}, the half-space defined by P_{IJ} and containing I, $P = \cap_{J \in \Lambda} P_{IJ}^{+}$ and $S = S(I, 2 * R)$. If $P \cap S = \emptyset$, I is not accessible. If not, let A_k be an edge of P intersecting S. $A_k \cap S$ gives the centers of a probe sphere lying on three atoms (I, J, K). The determination of the surface is then reduced to the determination of P, which is a limited Voronoi problem (see Ref. 4, for example).

2. Determination of surface topology

We now simplify the surface definition by restricting it to a polyhedron with planar triangular faces based on triplets of atomic centers determined in part 1. (see Color Plate 1(a)). We consider then that two points A,B of the polyhedron are "near" if the rolling probe sphere can find a short way from A to B. Provided the discontinuities at the vertices and at the edges of the polyhedron normal are smoothed, this polyhedron is, in fact, topologically well defined as a two-dimensional C^∞ manifold, compact, with no edge. The polyhedron can therefore be represented by a map, which is composed of a set of integer couples (edges), a set of integer circular permutations (vertices), and real triplets (coordinates, here atomic centers positions) (see Ref. 5 for a precise definition of maps). Since precise representation of the different pieces (concave, convex, and saddle-shaped), and usual cusp trimming[2] is not needed anymore, other issues are encountered when embedding the manifold into space.

Though it is easy to see that triangles do not intersect each other at interior points (with approximation on radii), three types of degenerated cases appear:

(1) Two faces with the same atomic numbers and opposite orientations
(2) An edge with more than two adjacent faces
(3) A vertex with several normals

Dealing with these singularities is the tricky part of a mapping algorithm: There is no bijection between atomic centers and polyhedron vertices, since several of these can coincide with the same center. A lot of good algorithms ($O(n)$) can be found to solve this problem. We choose to build each vertex, by having the probe sphere rolling around an atom. We thus obtain permutations on the edges of the triangles, which are then numbered and updated to point at vertices.

Because this polyhedron is a rough approximation of the molecular surface, which reduces significantly the interior volume, another polyhedron, which is, in a way, the dual of the previous one, has been defined (see Color Plate 1(b)). Its edges are joining two centers of the probe sphere generating two adjacent triangles. This new polyhedron is thus composed of nonplanar faces with any orders. The order of its vertices is now constant and equal to three, except for rare cases where it is equal to two. The embedding and the graphic representation of this polyhedron are more difficult, but it can still be proved useful (see Results).

RESULTS

For practical reasons, we dissociate the calculation of the surface (1) and the determination of its topology (2). It is obvious that a sophisticated algorithm could gather 1.2 and 2. Programs have been implemented on a UNIX 6150 workstation (IBM), using V.S. FORTRAN. They have been

Table 1. CPU times for different methods, with an atomic radius $R = 1.5$ Å and the PDB conformation of phage T4 lysozyme

	Method			
	Connolly DOTS	USURF	A	B
Time (s)	3900	1350	1400	190

Table 2. CPU times for methods A and B, for different values of the atomic radius R and the PDB conformation of phage T4 lysozyme

	R						
	1	1.5	2	2.5	5	50	5000
Time (s) B	250	190	230	280	730	1200	920
Time (s) A	360	1400	5900	17500			
d	86	2	0				
s	13	0					

Note: d, number of defects; s, number of complete tori

tested on different conformations of phage T4 lysosyme[6] (a typical small protein, 164 A.A., 2657 atoms), obtained by molecular dynamic at increasing temperatures.[7]

1. Determination of molecular surface

For an atomic radius of R = 1.5 Å, CPU times of the algorithms 1.A and 1.B together with the Connolly DOTS program and a QCPE program USURF[8] are shown in Table 1. Algorithm C has not been implemented, since algorithm B is quite sufficient if defects are not needed, and algorithm A is acceptable when R is small. When increasing R, algorithm A is prohibitive, while algorithm B is quite good, and defects disappear quickly, which leads algorithms A and C to be useless (see Table 2 and Color Plates 3(a)–3(c) to see the effect of variation of R). The convex hull of the molecule is obtained for very large R ($R \approx 5000$ Å in the studied case). Algorithm C, though, could be useful for very large molecules, with R = 1.5 Å and a need for the defects. CPU times of DOTS methods are presented, and must be used carefully since a little extra CPU time is needed to determine dots (though cusp trimming has been suppressed in the Connolly program for a significant comparison), and the USURF methodology is totally different.

2. Determination of surface topology

CPU time can be ignored for the considered range of molecules. Some practical applications of the obtained topology are immediate: the determination of the connected parts (c parts), and their isomorphic class indexes g_i. It is surprising that c and g_i are not affected today by the suppression of complete tori. Surface and volume values are not precise enough to be mentioned here. The number of accessible atoms a, c, g_0, and w, the total number of degenerescences, for three conformations of phage T4 lysosyme (folded (f), semiunfolded (su), unfolded (u), R = 1.5 Å) are presented in Table 3. Even in small molecules such as cryptates, g_i can be correctly calculated (see Color Plate 4 and Table 3).

3. Applications

Graphic applications The calculation of vertices and of the normals to each face is immediate, and the polyhedron can be manipulated by a 3D graphic workstation with shading and illumination (see Color Plates 5(a) and 5(b)). It is also easy to select interactively a piece of surface, using a locator and a valuator. Such a selection is even more important than clipping, because it is then possible to rotate the selected piece easily and visualize any interactions.

Molecular recognition A new definition of steric recognition (SR) can be given using the surface polyhedrons. Let (M1) and (M2) be two sterically interacting molecules. SR can be defined as the best steric fit between the polyhedron surface (S1) of (M1) and the dual polyhedron surface (S2′) of (M2) (see Color Plates 6(a)–6(d)). To quantify this fit, in well-known complexes, the number of couples (I, J), $I \in S1, J \in S2'$ and distance $(I, J) < 1$ Å has been calculated for various R (see Table 4). For R = 1.3 Å, a very good fit is obtained. Such a definition of steric recognition leads to automatic recognition algorithms of high complexity, since nobody knows which good bijection between atomic indexes of (S1) and atomic indexes of (S2′) has to be applied. A good way to lower the complexity is to assign to each atom a set of values that describes its local or semilocal environment.

Table 3. Different global parameters of molecular surface for atomic radius R = 1.5 Å, and three different conformations of phage T4 lysozyme obtained by molecular dynamic simulations, while increasing the temperature

Conformation	f	su	u
a	900	1000	1400
c	8	10	8
g_0	0	1	8
w	50	80	150

Table 4. Number of fits between ligand grid and subtrate dual grid, for two PDB complexes and different values of the atomic radius

Atomic radius	Lysozyme + substrate	Trypsin + inhibitor
1.2	17	24
1.3	23	25
1.4	22	25
1.5	16	23

Local properties A tool based on the topology of the polyhedron has been developed to study the local properties of molecular surface. Starting from an arbitrary point on the surface, the algorithm can be seen as Icare, willing to get out of this labyrinth (the surface grid). He decides to explore it methodically and at each new visited crossing, he leaves a sign of his visit, near the passage that has just led him to it and finds a way as far left as possible, which leads to a nonvisited crossing. After removal of the dead ends, the result is a spiral rolling the starting point up (see Color Plate 2(a)). A natural path on the surface (one-dimensional parameterization) controlled by a starting point has been generated. It is then easy to calculate a local convexity, a local density of points, or an interatomic distance map. As an example, the spirals of two sterically interacting atoms of the phage T4 lysozyme and its substrate are shown (see Color Plate 2(b)).

CONCLUSION

The simplified representation of a molecule, described above, is based on a limited number of approximations. It depends only on atomic coordinates and a roughness parameter, and is determined at low CPU cost, with a good complexity. Its graphic visualization is easy to program on any 3D workstation, and proves to be useful for manual docking.

A "natural" one-dimensional parameterization of the surface, depending only on the choice of a surface atom, and a definition of the steric fit of two molecules have been proposed. An automatic fitting procedure could be implemented by using this parameterization as a description of the local environment of each surface atom to guide the search. To obtain significant results, it would be necessary to include at least electrostatic interactions.

The defined surface appears to be quite general, and the algorithms could be applied to any set of points to describe its exterior surface, in a finer way than the convex hull, and could be used to recognize any nonconvex solid.

ACKNOWLEDGMENTS

The authors thank IBM for lending a graphic workstation and its technical support. A. Dietrich helped to obtain the pictures. This work was supported by the Groupement scientifique between IBM and the French National Research Center (CNRS).

REFERENCES

1 Connolly, M. L. *J. Appl. Crystallogr.* 1983, **16**, 548–558

2 Connolly, M. L. *J. Appl. Crystallogr.* 1985, **18**, 499–505

3 Yip, V. and Elber, R. *J. Comput. Chem.* 1989, **10**, 921–927

4 Finney, J. L. *J. Comput. Chem.* 1979, **32**, 137–143

5 Lienhardt, P. Proc. Eurographics RFA. Sep. 1989, pp. 439–452

6 Phage T4 lysozyme of the Protein Data Bank, 2LZM

7 Perrot, G. and Maigret, B. Unpublished results

8 Moon, J. and Howe, W. *J. Mol. Graphics* 1989, **2**, 109