Fast and Robust Computation of Molecular Surfaces

Michel F. Sanner[1], Arthur J. Olson[1], Jean-Claude Spehner[2]

[1]The Scripps Research Institute, 10666 North Torrey Pines Road, La Jolla, California 92037, USA.
[2]Université de Haute Alsace, Laboratoire MAGE, F.S.T., 4 rue des Frères Lumière, Mulhouse 68093, France.

Abstract

In this paper we define the *r*-reduced surface of a set of *n* spheres representing a molecule in relation to the *r*-accessible and *r*-excluded surfaces. Algorithms are given to compute the outer component of the *r*-reduced surface in O[*n*log*n*] operations and the analytical description of the corresponding *r*-excluded surface in time O[*n*]. An algorithm to handle the self-intersecting parts of that surface is described. These algorithms have been implemented in a program called MSMS, which was tested on a set of 709 proteins. The CPU time spent in the different algorithms composing MSMS are given for this set of molecules.

1 Introduction

Molecules are often represented as a set *S* of overlapping spheres, each having the van der Waals radius of its constituent atom. The solvent accessible surface [1] or *r*-accessible surface (*Ar(S)*) and the solvent excluded surface [2,3] or *r*-excluded surface (*Er(S)*) of a molecule are defined using a spherical probe of radius *r* representing a solvent molecule. *Ar(S)* is the topological boundary, denoted $\delta(S')$, of a set of spheres *S'* obtained by adding *r* to the radius of each sphere of *S*. It is also the trajectory of the center of a sphere of radius *r* rolling over the atoms. *Er(S)* is made of spherical and toroidal patches lying on the atoms and the probes. For $r = 0$, $Er(S) = \delta(S)$ and for $r = \infty$, *Er(S)* is the convex hull of *S*, denoted conv(*S*). The complexity in size of the set of *Er(S)* obtained for all possible values of *r* is in $\Theta[n^2]$ and these surfaces can be found with a worst case complexity in $O[n^2 \log n]$ [4]. *Ar(S)* and *Er(S)* can be made of several closed components and the outer one circumscribes all others. Several algorithms to compute analytical models of the *Ar(S)* [5,6] and triangulations of *Er(S)* [8,9] have been described, but we are aware of only two other programs which compute explicitly an analytical model of the *Er(S)*: PQMS [7] and AMS [3]. One of the main difficulties in computing this surface is the detection and handling of the self-intersecting parts of the surface called singularities [10,11].

After defining the *r*-reduced surface of *S*, denoted *Rr(S)* we present an algorithm to compute its outer component in O[*n*log*n*] operations and the corresponding *Er(S)* component in O[*n*] operations. Then the way singularities are treated is presented and the results of our program MSMS are presented and compared with those obtained with PQMS for our set of 709 proteins.

2 The *r*-Reduced Surface; *Rr(S)*

Rr(S) has been defined [11] as follows. The center of each sphere of *S'* holding *p* (*p*>0) *Ar(S)* faces defines a vertex of *Rr(S)* which has multiplicity *p*. For each circle *C*, which is the intersection of two spheres of *S'* and which holds *q* (*q>0*) edges of *Ar(S),* the open line segment joining the centers of these spheres is an edge of *Rr(S)* which has multiplicity q. If the *Ar(S)* edge *c* has no vertex, i.e. $c = C$, this edge of the *Rr(S)* is said to be free. For each vertex *v* of the *Ar(S)* there are *k* spheres $s'_1, s'_2, ..., s'_k$ of *S'* such that *v* belongs to all these spheres and the circles $s'_k \cap s'_1$ and $s'_i \cap s'_{i+1}$ for *i* in {*1, 2,...,k-1*} hold an *Ar(S)* edge ending in *v*. If $c_1, c_2, ..., c_k$ are the centers of the spheres $s'_1, s'_2, ..., s'_k$, a surface homeomorphic to an open disk bounded by the polygonal line $c_1c_2 ... c_kc_1$ and included in conv($c_1, c_2, ..., c_k$) is a face of *Rr(S)*.

For any calibrated set *S* of *n* spheres, and for any value of *r* adequate for *S*, the number of edges and faces of *Rr(S)* is in O[n] [11, 12]. A calibrated set of spheres is such that the ratio r_{max} / r_{min} of the maximum radius and the minimum radius in *S* is smaller than a constant T and the distance between the centers of any pair of spheres in *S* is larger than r_{min} / T. An adequate value of *r* for *S* is such that $r_{min} \le r \le r_{max}$.

Rr(S) is closely related to alpha-shapes [13]. The set of spheres *S* representing a molecule is calibrated, and the usual probe radius (1.4-1.5Å) is adequate for *S*.

3 Computing *Rr(S)* and *Er(S)*

• *Rr(S)*: The vertices of the first face *F* of *Rr(S)* are: the center of the left-most atom s_1 (min{*x* coordinate - radius}), the center of the atom s_2, such that the circle $s'_1 \cap s'_2$ holds the left-most point, and the center of any atom s_3 such that the probe tangent to s_1, s_2 and s_3 has no overlap with any atom of *S*. From such a position, the probe can roll over each pair of atoms, whose centers c_i, c_j are connected by an edge of *F*, until it hits a third atom whose center defines with c_i and c_j a new face of *Rr(S)*. This is done until all edges have been treated. Pairs of *Rr(S)* vertices which aren't connected by an edge and are closer than $2r + r_1 + r_2$,where r_1 and r_2 are the atom's radii, form free edges of that component of *Rr(S)*. Computational complexity: Because *S* is calibrated the number of atoms to consider when treating an edge is independent of *n* [9]. These atoms can be retrieved in O[log*n*] operations using a hierarchical subdivision of space [17] obtained by recursively dividing the bounding box of the set of atom's centers into two subsets of comparable sizes. Since *Rr(S)* has O[*n*], edges *Rr(S)* can be computed in O[*n*log*n*].

• *Er(S)*: For every *Rr(S)* face a spheric reentrant face of *Er(S),* lying on the probe, is created. For each *Rr(S)* edge, a toric reentrant face of *Er(S)* is created. Cycles of convex edges of toric reentrant faces define the contact faces. The complexity to build *Er(S)* from *Rr(S)* is a linear function of the number of faces and edges of *Rr(S)* which is in O[*n*].

4 Handling singularities

• Radial singularities occur when the radius of the torus, described by a probe σ while rotating around an edge *e* of *Rr(S),* is smaller than the radius of the probe. These singularities are detected during the computation of *Rr(S)*. The singular points ($\sigma \cap e$) are

stored with the *Rr(S)* edge. The toric reentrant face created for such an edge is made of two triangular faces connected by a singular vertex or a singular edge.

• Non-radial singularities occur when a part of a spheric reentrant face is inside another (non *Ar(S)* neighbor) probe. These singularities are handled in three steps:

1 - For each pair of probes tangential to the same three atoms and which both intersect the plane *P* defined by the centers of the three atoms, clip both spheric reentrant faces with *P*. The complexity to handle these singularities is a linear function of the number of *Rr(s)* faces having identical vertices.

2 - For each singular edge *e* created by a radial singularity, if *e*, or part of it, is inside a probe, it is entirely or partially "eaten". We add singular vertices at the intersection of the probes eating the biggest part of *e*, and *e*. The eaten part of *e* is replaced by singular edges lying on the circles intersection of the probes on which *e* was lying and the ones eating *e*. The created singular edges are added to the list of singular edges to be checked.

3 - For each pair of probes generating the spheric reentrant faces *F* and *F*', not tangent to the three same atoms and having an intersection circle *C* which doesn't intersect the edges of *F* or *F'*, clip *F* and *F'* with the plane holding *C*.

The complexity of steps 2 and 3 is a function of the number of singular edges and probes generating singularities. In the case of molecules these numbers are small and the experimental results show that the time spent in this part of the program is negligible and behaves as a linear function of the number of atoms (Fig. 1).

5 Results and Discussion

The outer component of *Rr(S)* and *Er(S)* of a set of 709 molecules have been computed by MSMS and PQMS for a probe radius of 1.5 Å. These molecules have been selected from the Brookhaven Protein Data Bank to provide a wide range of protein sizes (33 to 43632 atoms) and types. The *Er(S)* outer components were triangulated with a vertex density of 1.0 vertex/Å^2. Simple tests were performed on the triangulation to verify that every vertex belongs to at least two edges and that each edge belongs to exactly two triangles. All 709 MSMS surfaces successfully passed this test. This compares to 603 surfaces (85%) for PQMS.

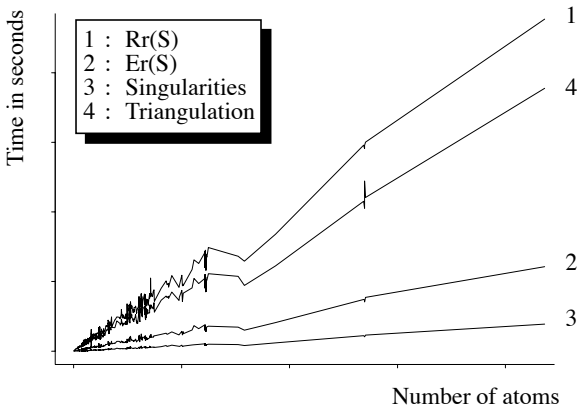


Fig. 1 : Time spend in different parts of MSMS measured for 709 proteins.

Figure 1 shows the CPU time in seconds used by the different algorithms composing MSMS for all molecules in our test set. These times were obtained on an SGI Challenge with 256 Mbytes RAM, running a MIPS R4400 processor at 150 MHz. MSMS computed a triangulation of the outer component of the *Er(S)* of glutamine synthetase (2gls entry in the Brookhaven Data Base, 46,632 atoms, 494,586 triangles) in 47.7 + 12.14 + 3.87 + 37.7 = 101.48 seconds.

MSMS computes the genus of the analytical *Er(S)* and of its triangulation. The comparison of these two numbers has proved to be a very inexpensive and powerful test for surface correctness.

For visualization purposes, it is important to avoid very thin triangles which are poorly rendered. The percentage, over all molecules, of such triangles dropped from 0.29% for PQMS to 0.078% for MSMS. Although MSMS produces fairly regular triangulations, the assumption that the spheres of *S* are in general positions generates a number of very thin toroidal faces which could be avoided by using an *Ar(S)* multiple vertex approach as proposed by Eisenhaber [6].

6 Conclusion

The *r*-reduced surface provides a very compact way to store geometric information which can be used to build several molecular surfaces: van der Waals, *r*-excluded and *r*-accessible. It enables the checking of atom or residue exposure to the solvent and the computation of accurate molecular surfaces and surface areas. *Rr(S)* contains all the information needed to test if a given point is inside or outside a specific molecular surface. The implementation of the described algorithms was shown to be very reliable and efficient. The practical performance of MSMS fit the theoretical complexity. The speed achieved allows interactive recomputation of dynamic molecular surfaces for small proteins undergoing molecular dynamics or normal mode motion. Higher speed could probably be achieved by optimizing the code. The program behaves remarkably well with large molecules where the complexity and the implementation are crucial. The surfaces computed with MSMS are widely used at The Scripps Research Institute to illustrate publications, study protein-protein and protein-ligand interactions, describe molecular surfaces by spherical harmonics [15] and study molecular properties.

Acknowledgment

This work was supported, in part by Sandoz Pharmaceutical Ltd., Basel Switzerland and by NIH grant number P01GM38794.

References

[1] Lee B.and Richards F.M. J. Mol. Biol. 55:379-400, 1971.
[2] Richards F.M. Ann. Rev. Biophys. Bioeng. 6:151-176, 1977.
[3] Connolly M.L. J. Appl. Cryst., 16:548-558, 1983.
[4] Boissonnat J.D., *et al*. J. Mol. Graphics, 12:61-62, 1994.
[5] Perrot G., *et al*. J. Comp. Chem., 13(1):1-11, 1992.
[6] Eisenhaber F., Argos P. J. Comp. Chem., 14(11):1272-1280, 1993.
[7] Connolly M.L. J. Mol. Graphics, 11:139-141, 1993.
[8] E. Silla *et al*. J. Mol. Graphics, 8:168-172, Sept. 1990.
[9] Varshney A., Ph D Dissertation. UNC. 1994
[10] Connolly M.L. J. Appl. Cryst., 18:499-505, 1985.
[11] Sanner M.F. *PhD thesis*. UHA, (France), 1992.
[12] Halperin D. and Overmars M. Proc. 10th ACM Symp. Comp. Geom., pp.113-122, 1994.
[13] Edelsbrunner H. and Mücke E. ACM Transaction on Graphics, 13:1:43-72, 1994.
[14] Barnes J. and Hut P. Nature 324:446-449, 1986.
[15] Duncan B.S., Olson A.J. Biopolymers, 33:219-229, 1993.