GEPOL: An Improved Description of Molecular Surfaces II. Computing the Molecular Area and Volume

Estanislao Silla* and Iñaki Tuñón

Departmento de Química Física, Universidad de Valencia, 46100 Burjassot, Valencia, Spain

Juan Luis Pascual-Ahuir†

Department of Physiology and Biophysics, Mount Sinai School of Medicine, New York, New York 10029

Received 19 October 1990; accepted 21 May 1991

The algorithm used by the program GEPOL for a finer description of molecular surface (for a fast calculation of molecular area and volume and for an efficient selection of sampling points) is presented in detail. Different types of surfaces such as van der Waals and Richard's molecular surfaces can be computed. As we described in the first article (J.L. Pascual-Ahuir and E. Silla, *J. Comp. Chem.*, **11**, 1047(1990)), GEPOL begins by building a set of spherical surfaces which fill the space which is not solvent accessible. In this second article, a triangular tessellation approach to select the parts of these spherical surfaces which form the molecular surface is described. By using a data coded generic pentakisdodecahedron, each spherical surface is divided in triangular tesserae. A simple method is used to eliminate all triangles found at the intersection volume of the spheres. The center coordinates and the surface of the remaining triangles are used in order to calculate the molecular area and volume and as starting point of the graphic representation of scalar and vector properties. We study the behavior of the method, presenting several examples of application. Special attention is given to the accuracy, spatial invariance and computer efficiency measured by CPU time. Some models of aligned spheres whose area and volume can be found exactly allow us to do a comparative study with a well-known method, analyzing their behavior in line with their respective graining parameters. A fragment of protein is used as an example of the application of the method for characterizing biomolecular surfaces. Aqueous solubility of organic compounds is studied as an example of the experimental property that depends on the molecular area obtaining a good correlation between the logarithm of the solubility and the area calculated using GEPOL.

INTRODUCTION

In order to gain a better understanding of structure-function relationships, considerable emphasis has been placed upon accurate calculations of the outer surface of molecules within the last few years.[2] The method in which to define and quantify this surface has been relegated to a question of convenience. Aside from very simple geometrical models such as spheres,[3] ellipsoids,[4] and cylinders[5] to represent molecular surfaces, there are two principle models. The first model, in which molecular surfaces are represented by van der Waals surfaces (WMS) or Lee and Richards solvent accessible surface (AMS),[6] consists of the envelope surface of a set of interlocking spheres (of appropriate radii) centered on each atom of the molecule. The second model, in which the definition of molecular surface by Richards[7] is applied (MS), the molecular surface is divided into two distinct parts: a contact surface and reentrant surface. The former is part of the WMS of each atom which is accessible to the solvent when the solvent molecule is modeled as a hard sphere (probe-sphere), while the latter is the inward-facing part of the probe-sphere surface which is in simultaneous contact with more than one atom.

In the last decade a number of algorithms have been proposed for estimating the molecular surface area and the volume enclosed within it;[8-17] yet none provide a fully satisfactory solution to this problem. Although Connolly's molecular surface approach has been widely used, it has also been criticized for its oscillatory behavior with respect to its internal parameters.[1-2] We have developed a set of efficient and reliable algorithms to calculate the molecular surface as well as the molecular volume.[1,18,19] These algorithms have been incorporated into GEPOL,[19] a program which computes the molecular surface as a distribution of points and then calculates the corresponding area and volume. The program begins by

*Author to whom all correspondence should be addressed.

†This work was performed during a stay of J.L.P. at the Universidad de Valencia.

CCC 0192-8651/91/091077-12$04.00

using a set of spheres centered on the atoms or group of atoms, whose radii are specified in the input; this is sufficient to compute the WMS and AMS. If the MS is the desired surface, the program will fill the spaces inaccessible to the probe sphere by creating a new set of spheres among the original set. This part has been explained in the previous article.[1]

Once the set of spheres has been defined, their spherical surfaces are divided into a set of triangular tesserae by using a data-coded generic pentakisdodecahedron. A simple geometrical procedure is used to eliminate all triangles found at the intersection volume of the spheres and a fast algorithm permits an efficient calculation of the actual molecular area, volume, and the coordinates of the remaining tesserae. The data concerning the surface are provided by the coordinates at the center of each triangle, the elementary surface value and the components of the normal vector to the surface. This facilities color coding of standard properties with standard graphic libraries.[20] In this article, we present the procedure employed to choose the fragments of the spherical surfaces that constitute the envelope surface, the method in which the surface area is calculated, and the set of selected points that describe it as well as its corresponding volume. Finally, we made a detailed analysis of the results obtained from several selected examples.

In the following section we will explain the method used for: (1) tesselling the spherical surfaces; (2) selecting the portions that form the envelope surface; (3) determining the surface of tesserae; (4) calculating the actual molecular area and volume; and (5) reducing the number of sampling points.

Often, it is difficult to estimate the merit of these methods, as there are no experimental quantities available for direct comparison. We did a comparative study with the MSDOT method[9-11,21] created by Connolly, because it is the most widely used method at the present time.[2,18] The comparative study was made using models of aligned spheres that serves as exact test cases. The correlation between the graining parameters and quality of the results was investigated. By calculating the surface of the retinol binding protein (RBP) we illustrated the application of this method in determining the area and volume of the biomolecules, as well as its possible use in the graphic display of proteins surfaces.

Finally, we present an example of how the GEPOL program is used in studying the existing relationship between the molecular surface and a macroscopic property determined experimentally: solubility. This relationship has been put forth by diverse authors using different methods to calculate the molecular surface. An exhaustive study carried out on 67 organic molecules shows the efficiency of GEPOL in establishing empirical relationships between calculated and experimental results.

COMPUTATION METHOD

The program begins by building a set of spheres in accordance with the model selected in the input WMS, AMS, or MS. Technical details on this procedure are reported elsewhere[1] and will not be repeated. Here, interest is focused on the procedure for calculating the resultant surface by means of a tessellation approach. The algorithm works as follows:

Tessellation

Once the set of spheres that define the chosen surface have been selected, the program divides each sphere in 60 spherical triangles of equal area, projecting on each sphere a pentakisdodecahedron, $\{3,5+\}1,1$ in the geodesic symbolism.[22] This can be visualized as inscribing a dodecahedron in the sphere and then defining additional triangles having two vertices in common with those of dodecahedron pentagons and the third vertex on the sphere surface on the radius passing by the pentagon center. (Fig. 1).

The division may be too coarse-grained, but the tessellation can be increased by means of the NDIV parameter given in the input. If the NDIV parameter is equal to 2, each initial triangular tesserae is divided into four triangles by linking the middle points of the sides with the maximum arcs. If NDIV is equal to 3 each one of the resulting triangles will be divided in four more and so on. Figure 1 shows the procedure carried out to calculate the coordinates of the vertexes of the new triangles. Let $\mathbf{x}_1$ and $\mathbf{x}_2$ be the co-

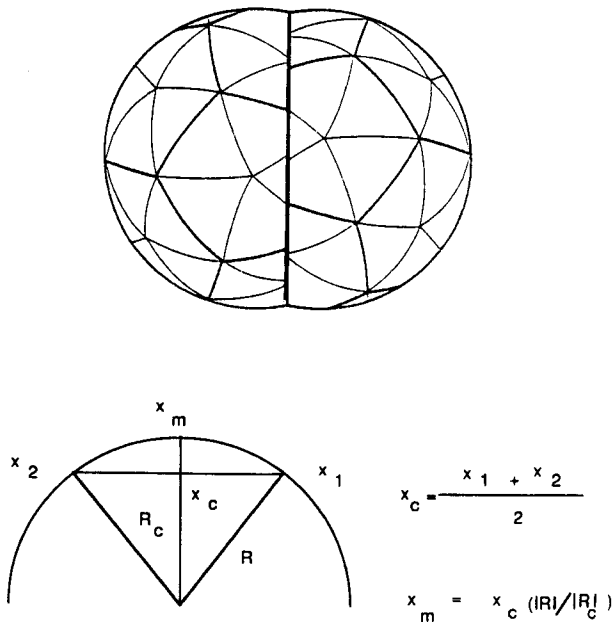
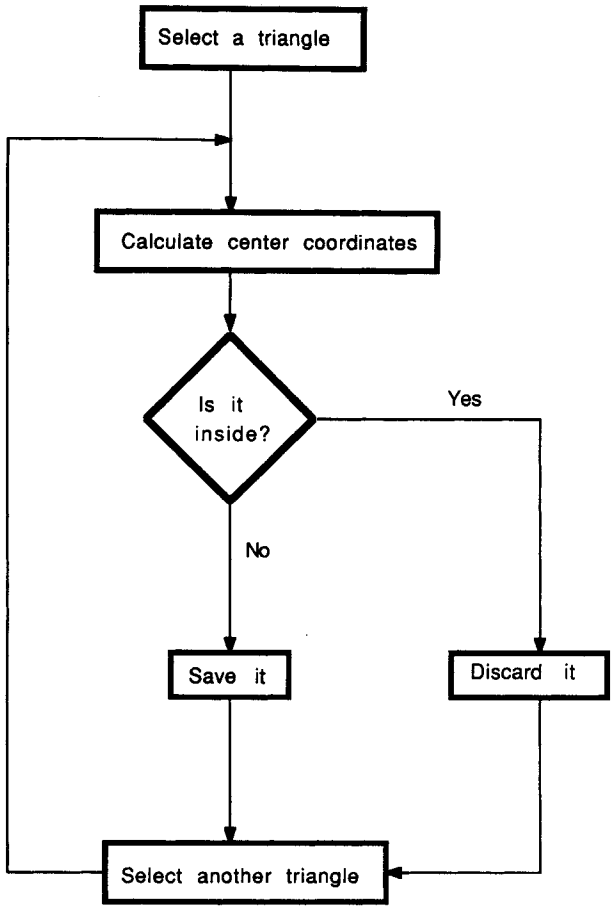


Figure 1. A schematic representation of two spheres tesselation obtained by projecting a pentakisdodecahedron on each one and procedure to divide the tesserae (see text).

ordinates of two vertexes of the initial triangle and $\mathbf{x}_m$ the coordinates corresponding to the midpoint along the arc (on the sphere surface). The midpoint of the chord (straight line passing by $\mathbf{x}_1$ and $\mathbf{x}_2$) $\mathbf{x}_c$ is calculated. If R_c is the distance from the origin of the sphere to $\mathbf{x}_c$ and R is the radius of the reference sphere, the coordinates of the midpoint on the spherical surface are the components of the vector: $\mathbf{x}_m = \mathbf{x}_c\ (R/R_c)$. The division process is repeated NDIV-1 times. Therefore, if NDIV is 1, 2, 3, 4, or 5, the number of spherical triangles per sphere will be 60, 240, 960, 3840, and 15360 respectively.

Selection of the Envelope Surface

When the triangular division is completed, the program calculates the coordinates of the center of each triangle and eliminates the triangles that have their centers within the surface, so that the remaining triangles now form the envelope surface. This procedure requires the calculation of the triangle's center. The program computes the coordinates of the barycenter: $\mathbf{x}_b = (\mathbf{x}_1 + \mathbf{x}_2 + \mathbf{x}_3)/3$, then the coordinates of the triangle center on the spherical surface are the components of the vector: $\mathbf{x}_n = \mathbf{x}_b\ (R/R_b)$. In order to determine which triangle is over or inside the surface, the program calculates the distance between the center of a triangle and the center of all other spheres to which the triangle does not belong to. If this distance is less or equal to the radius of this sphere the triangle is discarded; otherwise, the triangle belongs to the envelope surface. This process is repeated for all triangles forming the spherical surfaces set as shown in Scheme 1.



Scheme 1. Flux diagram to analyze which spheric triangles belong to the envelope surface.

Area and Volume Calculation

The envelope area is trivially obtained by summing up the areas of all remaining triangles.

$$S = \sum S_i \qquad (1)$$

where S_i is the area of the ith triangle.

In order to calculate the molecular volume, we first define a molecular origin. Let $\mathbf{r}_i$ be the position vector of the ith triangle center and $\mathbf{n}_i$ the corresponding normal vector to the surface at this point. The volume is obtained by summing all solid volumes made by the triangles vector surfaces and the origin of the molecular system:

$$V = (1/3) \sum S_i\ \mathbf{n}_i\ \mathbf{r}_i \qquad (2)$$

Distribution of Points

In addition to calculating area and volume, this kind of method can be used for other applications such as displaying the molecular surface on graphic terminals[20] and as a guide for developing models in which to calculate properties such as solute-solvent interactions.[18,23] For these applications, a reduced set of points on the surface is preferable. Therefore in order to avoid an excessive number of points, GEPOL (after calculating the area and volume) regroups the subtriangles into the larger original tesserae, resulting in a maximum of 60 points for each sphere.

STUDY OF THE VALUES OF NDIV

As we described in the previous section, the NDIV parameter controls the level of fine-graiiness by specifying the number of disivions each triangle in the original tesserae underwent. In order to study the behavior of the results as a function of NDIV, we have calculated the area and volume of a series of systems varying the graining parameter from 1 to 5. The first structures studied consist of a number (6, 8, 10, 12) of linearly arranged spheres, each of radius 1.8 Å and a distance of 1.5 Å between neighboring centers. All calculations were made in three orthogonal orientations and the average and root mean square (r.m.s.) deviation were computed. Comparisons of the MSDOT program were made by varying the point density parameter (DEN) from 1 to 25 as

a function of the CPU time on a Microvax 3200. The following three items were considered: *Accuracy*, a comparison of GEPOL with MSDOT with exact values permits conclusions to be drawn; *spatial invariance*, an important aspect of these models is that the result should not vary according to the spatial orientation of the structure under study with respect to the reference frame (spatial invariance of the algorithms is tested); *computer time*, computer efficiency measured by CPU time consumed is important, therefore the results were plotted vs. the CPU time.

The average area values are compared in Figure 2. All the examples demonstrate the same behavior: GEPOL results converge fairly quickly while MSDOT results oscillate significantly. We noticed that these are exact test cases in which it is possible to extract the exact value of the area. In Figure 3, we have represented the relative error of the values computed for the area of a structure comprised of 12 spheres. The relative error for GEPOL is never greater than 1.1% and decreases rapidly, while MSDOT increases to 10% with large downward fluctuations. Figure 4 shows the r.m.s. deviation for the same structure. These results can be taken as a measurement of the dependence of the orientation with the reference frame. As can be seen, an obvious improvement in the r.m.s. deviation of GEPOL results is produced when NDIV is greater than 1, while MSDOT results confirm their oscillatory nature. Convergency of the GEPOL results allow us to anticipate and to select the level of spatial invariance. In all those cases studied the CPU time necessary for obtaining acceptable results is less with our program. When one generally makes a calculation, it is necessary to know which parameter has to be used and the numerical reliability of the results. From all our calculations we can deduce that GEPOL is suitable enough when NDIV is greater or equal to 3, and we believe that this value allows for a good balance

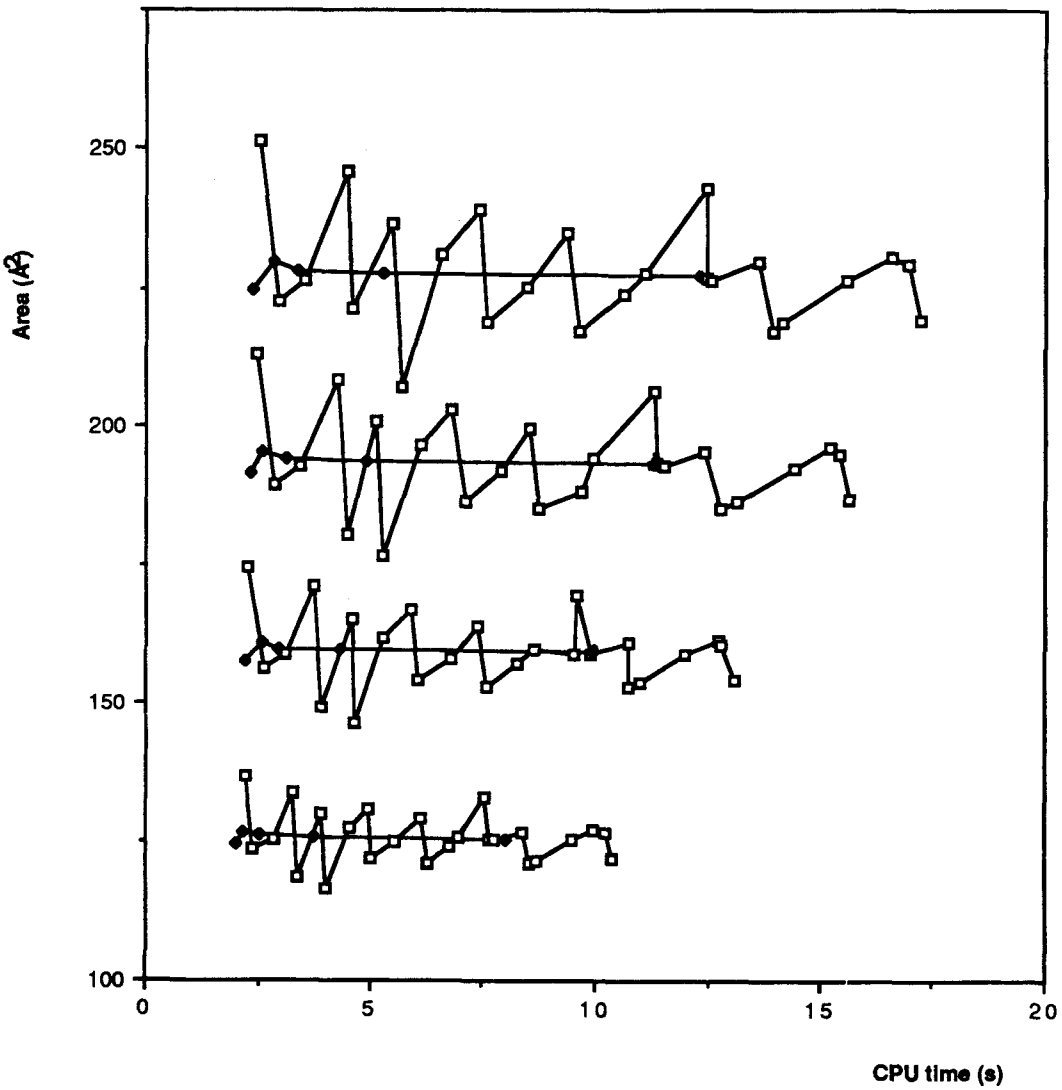


Figure 2. Surface area values for 12, 10, 8, and 6 spheres models versus CPU time with GEPOL (♦) and MSDOT (□) methods.

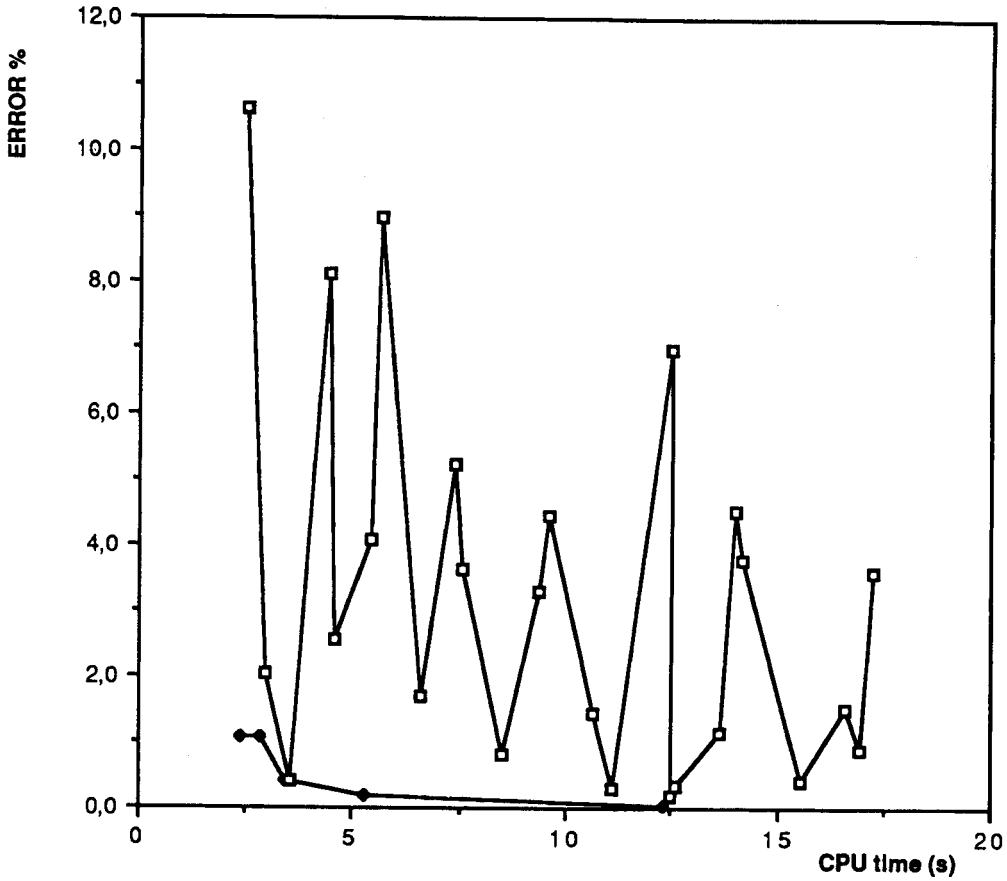


Figure 3. Relative error for 12 spheres model surface area versus CPU time with GEPOL (♦) and MSDOT (□) methods.

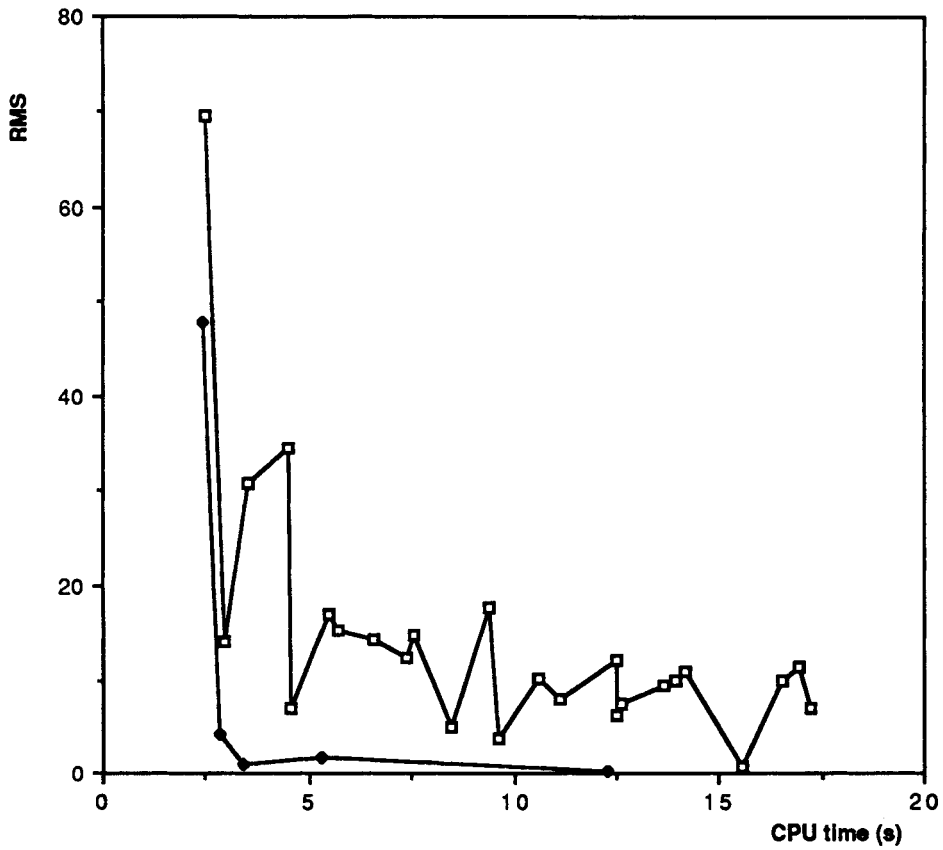


Figure 4. rms deviation for 12 spheres model surface area versus CPU time with GEPOL (♦) and MSDOT (□) methods.

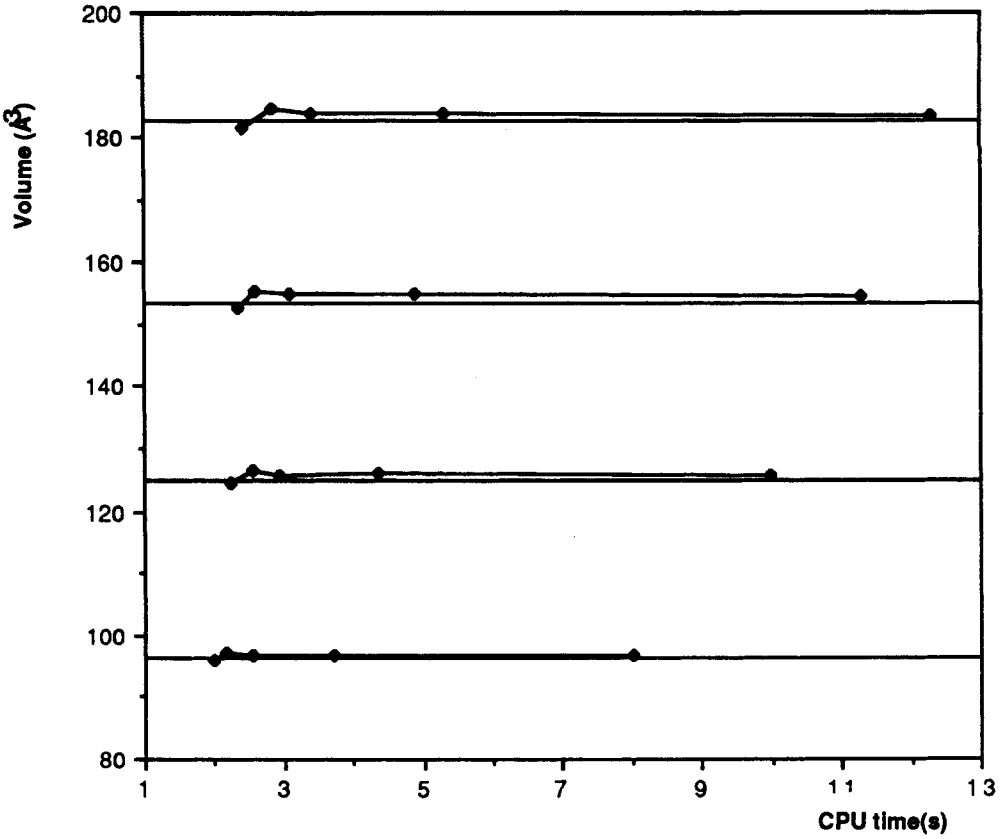


Figure 5. Volume values for 12, 10, 8, 6 spheres models calculated with GEPOL method versus CPU time. Straight lines show the exact values.

between accuracy and CPU time.

An additional advantage of our program is that we are given the value of molecular volume in the same calculation. Figure 5 indicates the averaged volume values of the structures formed by linearly arranged spheres. Once again, we observe both a good behavior in GEPOL for determining the volume of the cavity and a rapid convergence towards the exact value. We point out that the maximum error is only 1.08% and that the r.m.s. deviation decreases rapidly. The slight error obtained for NDIV = 1 seemed to be a consequence of the compensation between the errors of the three orientations, since a large value for r.m.s. is obtained in this calculation. Therefore, GEPOL is a good method for calculating both the area of a given surface and the volume enclosed by it.

PROTEIN SURFACE CALCULATIONS

Characterization of biomolecular surfaces is one of the important problems in molecular biophysics. To illustrate the quality of the present method in this application, calculations have been done for the van der Waals molecular surface of retinol binding protein (RBP). This molecule is comprised of 1438 heavy atoms. The following radii have been used: 1.4 Å for the oxygens, 1.5 Å for the nitrogens, and 1.9 Å for the carbons. The geometry was taken from a molecular dynamics simulation.[24] Figure 6 shows the values of the area obtained with GEPOL and MSDOT. Area calculations have been done with NDIV = 1, 2, 3, 4, 5 and DEN = 1, 10, 20, 30, 40 and 41 (the last one is the more accurate graining parameter possible in MSDOT calculations for this example). In Figure 7 GEPOL volume values are presented. In Figure 8, relative errors for area and volume values obtained by GEPOL are shown. In this case, the relative error is computed with respect to the best value (NDIV = 5), because the exact value is not known. It is clearly seen, no particular problems are encountered when using GEPOL to calculate the molecular area and volume for large systems. The behavior of GEPOL is convergent when NDIV is increased, while MSDOT again shows oscillatory behavior. This permits the GEPOL user to select the appropriate value of NDIV. Oscillatory behavior does not allow for a correct decision. It also appears in this case that GEPOL has a good behavior when NDIV is greater or equal to 3, a value which permits a good balance between accuracy and CPU time.

Once the molecular surface calculation has been computed, one actually has a set of tesserae centers, a set of elementary area values, and a set of normal vectors. Thus, a fairly detailed information based on scalar and vector properties such as the electric field, atom velocity, etc., can be graphically repre-

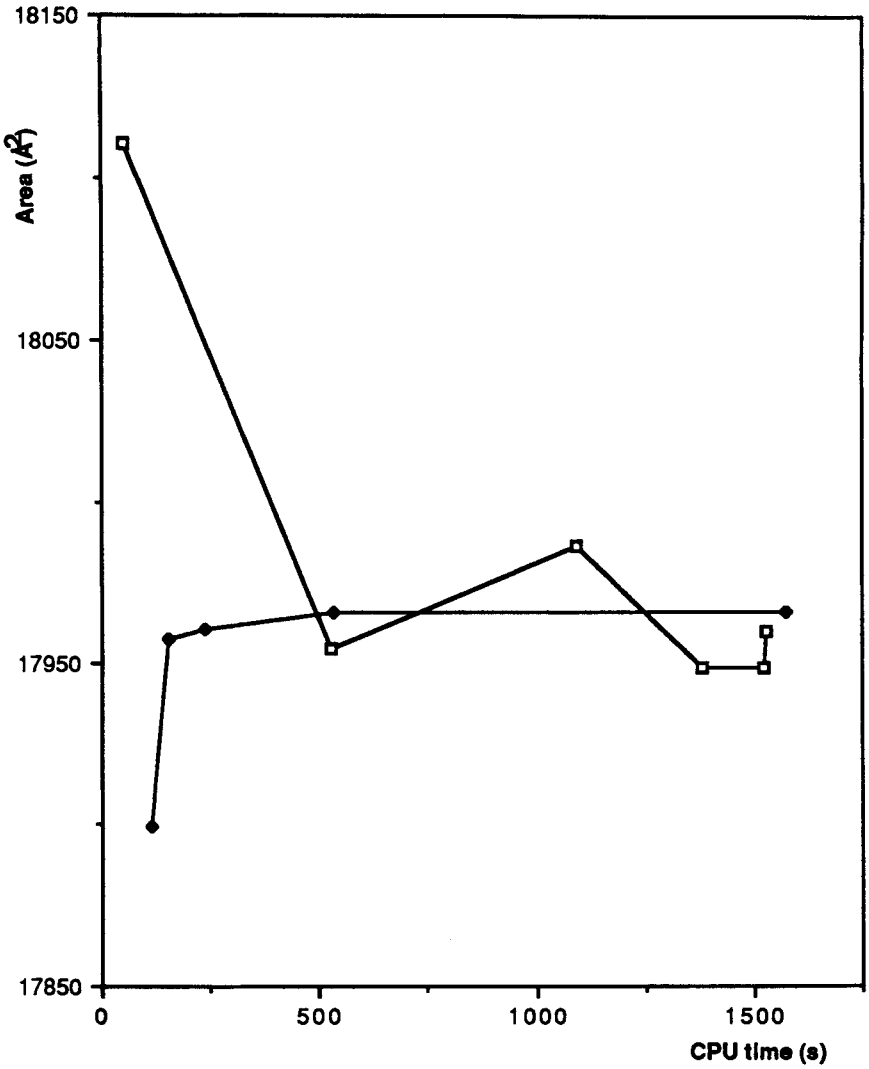


Figure 6. GEPOL (♦) and MSDOT (□) area calculations on retinol binding protein versus CPU time. The graining parameters values used are: NDIV = 1, 2, 3, 4, and 5; DEN = 1, 10, 20, 30, 40, and 41 that is the maximum possible value for this example.

sented by using appropriate color coding. The present representation is an excellent mean to convey vector properties. Among them are the atom velocity obtained from MD simulations. By calculating the dot product of the normal for a given triangle with the corresponding atom velocity, the direction of motion across the tesserae (a velocity "flux") can be color coded according to sign of the dot product. A type of kinetic energy flux can also be defined by taking the square of the projected velocity multiplied by its mass surface density (mass of the atom divided by the number of triangles).[20] When these types of calculations are carried out on biological systems and in building models to calculate molecular properties like solute-solvent interaction,[17,18,23] an excessive number of points on the surface implies a great increase in the CPU time without (necessarily) being matched by an improvement in the results. In Figure 9 we report the number of points obtained with MSDOT and GEPOL for the WMS of retinol binding protein (RBP). As can be observed, the number of points obtained using our program increased slowly when granularity is increased. Recall that GEPOL, after calculating the area and volume, regroups the subtriangles into the larger original tesserae in such a way that a maximum of 60 points is generated for each sphere. Therefore, an improvement in the quality of the surface need not be accompanied by a large increase in the number of points needed to represent it.

MOLECULAR AREA-AQUEOUS SOLUBILITY RELATIONSHIPS

Molecular surface area is one of the determining factors of molecular behavior. Langmuir[25] was the first to indicate that the surface area of a solute determines its thermochemical activity. Langmuir believed that the energy required to build a cavity

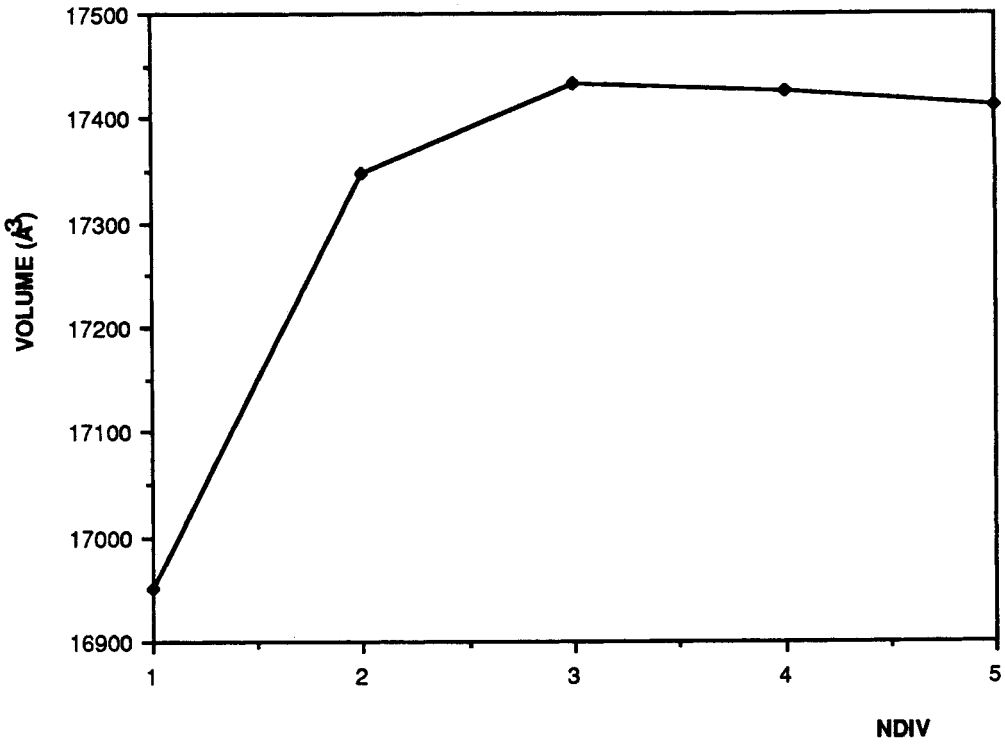


Figure 7. Volume values for RBP obtained with GEPOL versus NDIV.

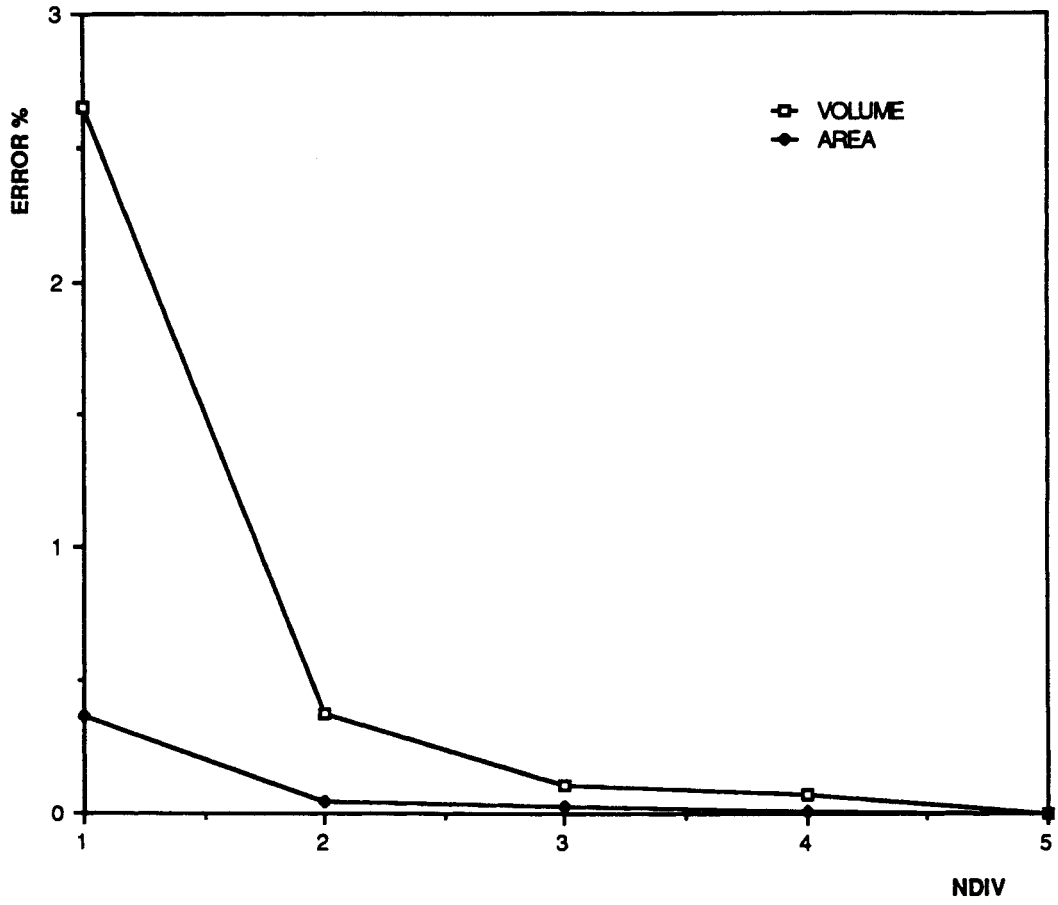


Figure 8. Relative errors of GEPOL volume and area values versus NDIV.

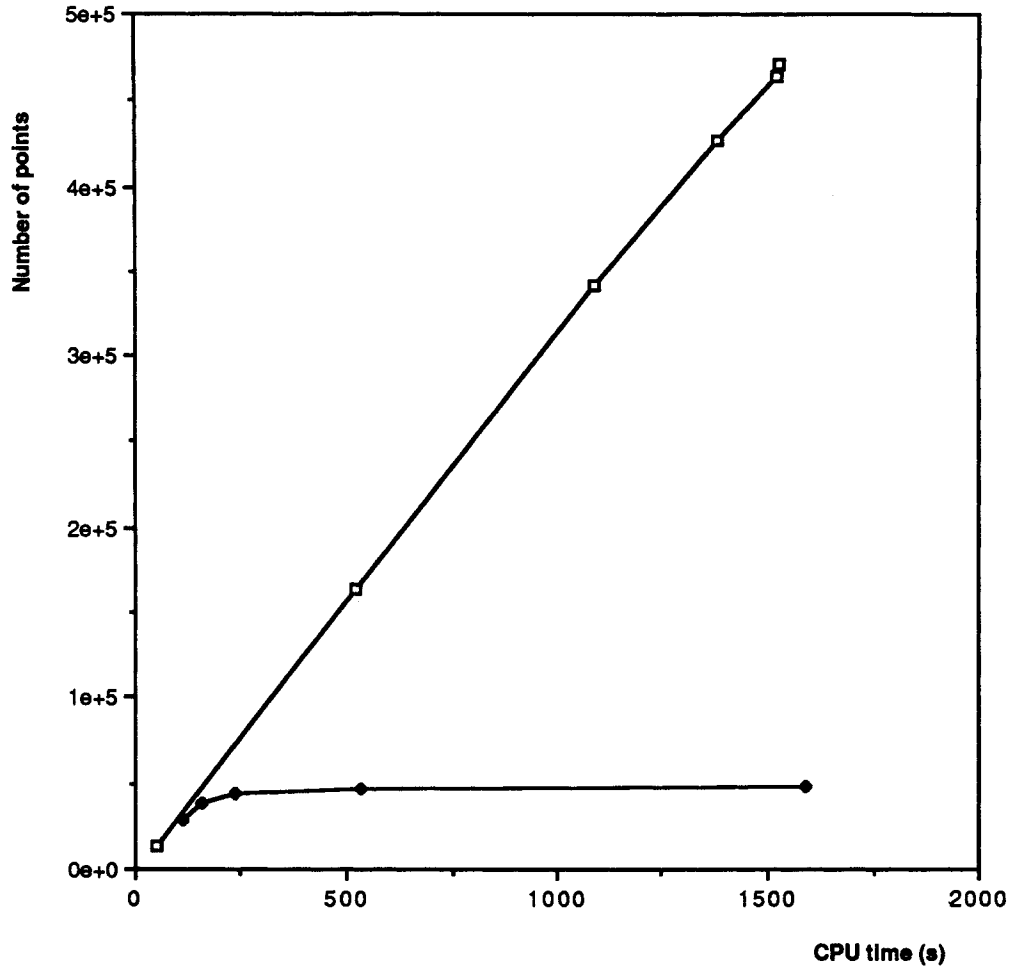


Figure 9. Number of points for RBP surface calculations versus CPU time with GEPOL (♦) and MSDOT (□) methods.

in the bulk of the solvent was a primary factor in solubility and proportional to the surface area of the solute. Since then, several studies showing the relationship between molecular surface area and solubility have been performed.[26-30]

The free energy of solution can be considered as a sum of cavitation, electrostatic and dispersion terms. The cavitation term (ΔG_{cav}) includes the necessary energy to create a cavity in the solvent which shelters the solute. The ΔG_{ele} term (electrostatic) comes from solute-solvent polarization. Finally, ΔG_{dis} is the dispersion energy term between the solute and the solvent. In order to calculate these terms, formulas have been offered illustrating their relationship with molecular surface area, S, of the solute[8,32-35]

$$\Delta G_{sol} = \Delta G_{cav} + \Delta G_{ele} + \Delta G_{dis} = f(S) \quad (3)$$

If a process occurs in ideal solutions: $\Delta G_{sol} = RT\ln X$. However, in real solutions the solute-solvent interactions are different from the solvent-solvent ones, causing the inclusion of a term of enthalpic nature. As derived from eq. (3), we can assume that this term is a linear function of the molecular area thus obtaining

$$\Delta G_{sol} = RT\ln X + aS + b \quad (4)$$

Where a and b are constants. In an equilibrium situation we have $\Delta G_{sol} = 0$ and then $-RT\ln X = aS + b$. For compounds that are not very soluble, such as those studied here: $X = M_s$ (sol)/1000, where sol is the molal solubility and M_s the molecular weight of the solvent. Therefore, if the stated hypothesis are acceptable, the following relationship must be fulfilled

$$\ln \text{sol} = AS + B \quad (5)$$

In our study we have computed the molecular surface area of 67 different molecules of esters, ethers, alcohols and ketones. Linear, branched and cyclic systems also were included (Table I). The molar solubilities[36-38] used correspond to pure supercooled liquid at 25°C. In order to calculate the molecular surface area (as defined by Richards) we had chosen the following geometric parameters. Bond lengths[39] are: H—O = 0.97 Å, O—C = 1.43 Å, C—C = 1.54 Å, C—H = 1.09 Å, C=O = 1.22 Å. Bond angles[40] are: H—C—H = 109.47°, H—O—C = 105.00°, O=C—O = 127.27°, C—O—C = 109.47°.

Table I. Molecular surface area (Å²), observed ln sol (molal) and calculated ln sol (molal) for compounds under study.

Compound	Area	Ln Sol	Ln Sol Calc.
Butanol	119.67	0.09	−0.18
Pentanol	139.69	−1.35	−1.43
Hexanol	159.68	−2.72	−2.67
Heptanol	179.70	−4.07	−3.36
Octanol	199.71	−5.40	−5.16
Nonanol	219.73	−6.91	−6.41
Decanol	239.72	−8.22	−7.65
Dodecanol	279.74	−10.68	−10.15
Tetradecanol	319.75	−12.77	−12.64
Pentadecanol	339.77	−13.80	−13.88
Hexadecanol	359.76	−14.60	−15.13
2-Butanol	117.20	0.07	0.03
2-Pentanol	137.23	−0.63	−1.27
2-Hexanol	157.25	−2.00	−2.52
2-Octanol	197.29	−4.76	−5.01
2-Nonanol	217.32	−6.32	−6.26
3-Pentanol	134.66	−0.49	−1.10
3-Hexanol	154.68	−1.83	−2.35
3-Heptanol	174.70	−3.19	−3.61
3-Nonanol	214.75	−6.12	−6.10
3-Methyl-2-butanol	131.46	−0.40	−0.92
2,4-Dimethyyl-3-pentanol	163.97	−2.80	−2.94
2-Ethyl-hexanol	188.47	−4.99	−4.47
2,2-Diethyl-pentanol	197.01	−5.57	−5.00
2-Methyl-butanol	133.13	−1.06	−1.02
3,5,5-Trimethyl-hexanol	191.55	−5.77	−4.66
Cyclohexanol	145.13	−0.96	−1.77
4-Methyl-2-pentanol	151.07	−1.81	−2.13
7-Methyl-octanol	213.95	−5.74	−6.05
4-Methyl-pentanol	153.98	−2.28	−2.31
2-Methyl-propanol	116.33	0.02	0.03
2-Pentanone	128.11	−0.39	−0.71
2-Hexanone	148.76	−1.80	−1.98
2-Heptanone	168.76	−3.28	−3.24
3-Methyl-2-butanone	125.73	−0.29	−0.56
3-Methyl-2-pentanone	140.38	−1.55	−1.47
4-Methyl-2-pentanone	144.83	−1.63	−1.75
3-Pentanone	131.00	−0.53	−0.89
3-Hexanone	151.07	−1.90	−2.14
4-Methyl-3-pentanone	145.77	−1.87	−1.81
2,4-Dimethyl-3-pentanone	159.58	−2.99	−2.66
4-Heptanone	171.07	−3.32	−3.38
5-Nonanone	211.14	−5.93	−5.88
Methyl-acetate	100.90	1.19	0.99
Ethyl-acetate	122.80	−0.09	−0.38
Propyl-acetate	142.79	−1.69	−1.62
Isopropyl-acetate	137.44	−1.20	−1.29
Isobutyl-acetate	159.43	−2.85	−2.65
Methyl-butyrate	141.00	−1.79	−1.51
Ethyl-butyrate	162.89	−2.94	−2.87
Propyl-butyrate	182.89	−4.39	−4.12
Ethyl-hexanoate	202.96	−5.43	−5.37
Ethyl-heptanoate	223.03	−6.32	−6.62
Ethyl-octanoate	243.03	−7.80	−7.86
Ethyl-nonanoate	263.10	−8.74	−9.11
Ethyl-decanoate	283.09	−9.43	−10.35
Diethyl ether	122.40	−0.15	−0.35
Methyl butyl ether	140.61	−2.28	−1.49
Methyl isobutyl ether	138.66	−2.07	−1.27
Methyl t-butyl ether	128.16	−0.48	−0.71
Methyl s-butyl ether	132.36	−1.69	−0.97
Methyl propyl ether	120.55	−0.86	−0.24
Methyl isopropyl ether	115.51	−0.06	0.08
Ethyl propyl ether	142.40	−1.53	−1.60
Ethyl isopropyl ether	137.29	−1.28	−1.28
Dipropyl ether	162.47	−3.03	−2.85
Propyl isopropyl ether	157.36	−3.07	−2.53

The given radius for spheres centered on each atom[41] were: RH = 1.2 Å, RO = 1.4 Å, and RC = 1.6 Å. Where appropriate the trans conformation was used. The parameters selected to get the MS were: 1.5 Å for the radius of the probe sphere, FRADIO of 0.5 (FRADIO is the parameter that controls the creation of new spheres and is described elsewhere[1]), and NDIV = 5.

The results of the linear relationship between ln sol and molecular surface area are shown in Table II. In general the obtained correlations are very good. Ethers present the worst correlation. We attribute this to the difficulty of finding accurate values for solubility and to the limited number of ethers considered. Table II shows that the assumptions made about molecular geometry are acceptable and therefore we can conclude that the deviations from assumed geometry are small. Herman[27] has used weighted averages for the different conformations in his hydrocarbon surface area calculations. However this did not lead to an improvement of the results because the greatest source of error was due to the method employed to compute surface area and the kind of surface used to describe the molecular behavior in solution.

On the other hand, it is seen that the influence of branching, isomerism, and cyclation on solubility correlates with molecular surface area. Therefore, we can predict that unbranched compounds are less soluble in water than the corresponding branched compounds (these present lower molecular surface area). For the isomeric alcohols we observed that the solubility trend is: 3-alcohol > 2-alcohol > n-alcohol. This can be understood by the corresponding sequence in surface area data: 3-alcohol < 2-alcohol < n-alcohol. Amidon et al.[36] point out that this effect is due to the fact that the association between alcohol molecules in pure liquid decrease in the order primary, secondary, tertiary (by steric limitations). This is confirmed by boiling points. In

Table II. Slopes, intercepts, and coefficients of correlation and standard deviation obtained with equation ln sol = $A^*S + B$ for studied organic families.

Compounds	A^*10^2	B	r	s
Alcohols	−6.399	7.631	0.994	0.455
Ketones	−6.681	8.043	0.994	0.179
Esters	−5.916	6.805	0.997	0.288
Ethers	−6.521	7.329	0.937	0.393

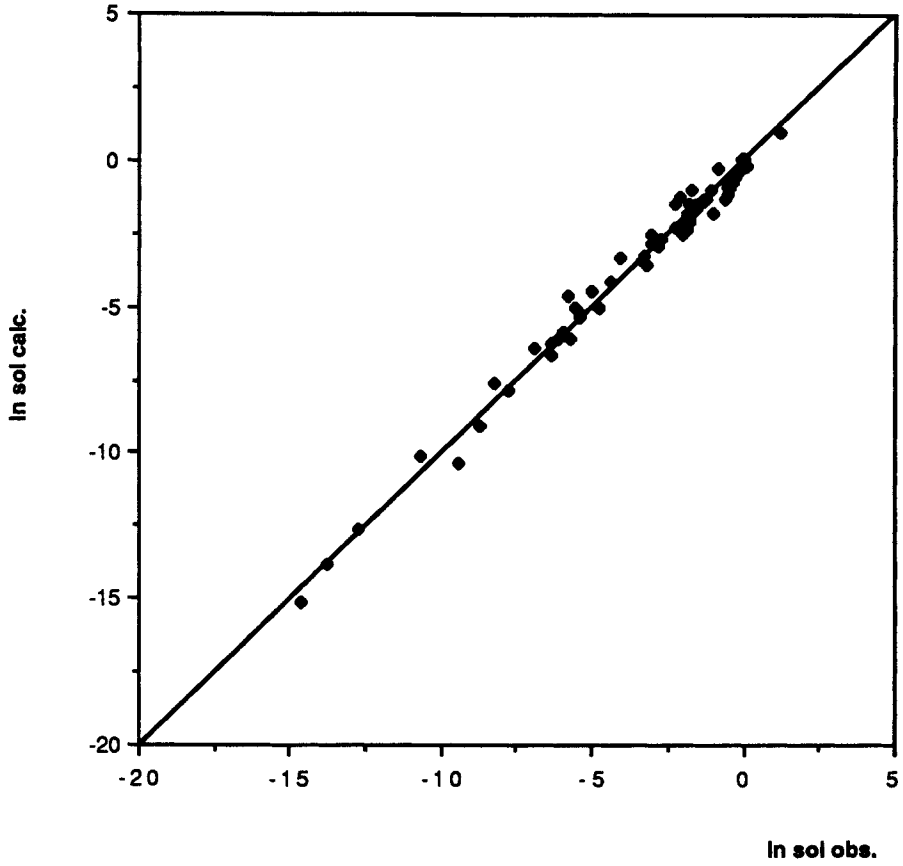


Figure 10. Calculated versus observed ln sol.

addition, this effect can be view as a consequence of molecular surface values (more surface area, more cavitation energy needed, less solubility).

Observing the similarity of slopes and intercepts of the different organic families (alcohols, ketones, ethers, and esters) we attempted to describe them altogether thus obtaining the following equation:

$$\ln \text{sol} = -6.225\ 10^{-2} S + 7.268; \quad r = 0.993; s = 0.441 \quad (6)$$

One can conclude that for these families (alcohols, ketones, esters, and ethers), the solubility is determined mainly by molecular surface area since the influence of functional group is very similar in all of them.

Figure 10 presents the solubility obtained with eq. (6) versus the observed solubility (neperian logarithms). That agreement allows us to conclude that a close relationship exists between solubility and molecular surface area. Hence, it appears that our model (fixed geometry, molecular surface as defined by Richards and calculated by GEPOL) is adequate. The computation method used (GEPOL) is suitable for obtaining good molecular surface area results and these results can be used to predict molecular properties.

CONCLUSIONS

A purely geometrical procedure to compute actual area and volume of molecules has been developed. The molecular surface area is formed by the external parts of a set of interlocking spheres built according to the model of surface area chosen. A fast and efficient tessellation procedure is employed to select the parts of these surface areas which make up the envelope surface area. Using a simple model of aligned spheres, our method has been compared to the most widely used MSDOT. Greater accuracy and spatial invariance as well as a lower cost of CPU time are just some of the advantages of GEPOL. What is more, error and dependence along with the orientation of the results of our program decrease both rapidly and convergently when NDIV is increased, allowing the user to select the quality of calculation in relation to the CPU time employed. When applied to biomacromolecules the results are also satisfactory. When the graining parameter NDIV is increased, the number of points on the surface area does not increase significantly. The agreement between experimental solubility and that computed with eq. (6) shows that solubility is determined principally by the molecular surface and confirms that GEPOL is a good method for computing molecular area.

We would like to thank Mr. Waldo Feng and Mr. Jason

Banfelder for reviewing the language used in the English manuscript. This work was supported in part by Acción Concertada de Investigación (No. 545.4). I.T. acknowledges a doctoral fellowship from Ministerio de Educación y Ciencia (Spain).

References

1. Part I: J.L. Pascual-Ahuir, and E. Silla, *J. Comp. Chem.*, **11,** 1047 (1990).
2. A. Meyer, *Chem. Soc. Rev.*, **15,** 449 (1986), and references given therein.
3. B. Linder, *Adv. Chem. Phys.*, **12,** 225 (1967).
4. B.J. Costa Cabral, D. Rinaldi, and J.L. Rivail, *C. R. Acad. Sc. Paris*, **298,** 675 (1984).
5. A. Horta and I. Fernandez-Pierola, *Macromolecules*, **14,** 1519 (1981).
6. B. Lee and F.M. Richards, *J. Mol. Biol.*, **55,** 379 (1971).
7. F.M. Richards, *Ann. Rev. Biophys. Bioeng.*, **6,** 151 (1977).
8. S. Miertus, E. Scrocco, and J. Tomasi, *Chem. Phys.*, **55,** 117 (1981).
9. M.L. Connolly, *J. Appl. Cryst.*, **16,** 548 (1983).
10. M.L. Connolly, *Science*, **221,** 709 (1983).
11. M.L. Connolly, *J. Am. Chem. Soc.*, **107,** 1118 (1985).
12. A. Meyer, *J. Chem. Soc. Perkin II*, 1161 (1985).
13. A. Meyer, *J. Comp. Chem.*, **7,** 144 (1986).
14. M.L.J. Drummond, *J. Chem. Phys.*, **88,** 5021 (1988).
15. H.R. Karfunkel and V.E. Eyraud, *J. Comp. Chem.*, **10,** 628 (1989).
16. R. Voorinthold, M.T. Kosters, G. Vegter, G. Vriend, and W.G.H. Hold, *J. Mol. Graphics*, **7,** 243 (1989).
17. R.J. Zahuar and R.S. Morgan, *J. Comp. Chem.*, **11,** 603 (1990).
18. J.L. Pascual-Ahuir, E. Silla, J. Tomasi, and R. Bonacorsi, *J. Comp. Chem.*, **8,** 778 (1987). Program GEPOL can be obtained from: Quantum Chemistry Program Exchange, Indiana University, Bloomington, Indiana 47405. USA. Program 554, QCPE Bull., (1988).
19. J.L. Pascual-Ahuir and E. Silla, in R. Carbó, Ed., *Studies in Physical and Theoretical Chemistry*, **62,** 597 (1989).
20. E. Silla, F. Villar, O. Nilsson, J.L. Pascual-Ahuir, and O. Tapia, *J. Mol. Graphics*, **8,** 168 (1990).
21. M.L. Connolly, Program 75, *QCPE Bull.* (1981).
22. H.S.M. Coexter, *Regular Complex Polytopes*, Cambridge University Press, Cambridge, 1974.
23. J.L. Pascual-Ahuir, J. Andres, and E. Silla, *Chem. Phys. Lett.*, **169,** 297 (1990).
24. J. Åquist, P. Sandblom, T.A. Jones, M.E. Newcomer, W.F. van Gunsteren, and O. Tapia, *J. Mol. Biol.*, **192,** 593 (1986).
25. I. Langmuir, *Third Colloid Symposium Monograph*, Chemical Catalog Co. New York, N.Y., 1925, p. 3.
26. G.L. Amidon, S.H. Yalkowsky, S.Y. Anik, and S.C. Valvani, *J. Phys. Chem.*, **79,** 2239 (1975).
27. R.B. Hermann, *J. Phys. Chem.*, **76,** 2754 (1972).
28. G.L. Amidon, S.C. Valvani, and S.H. Yalkowsky, *J. Phys. Chem.*, **80,** 829 (1976).
29. G.G. Hall and C.M. Smith, *J Mol. Struc. (THEOCHEM)*, **179,** 293 (1988).
30. E. Silla, I. Tuñón, F. Villar, J.L. Pascual-Ahuir, *J. Mol. Struc. (THEOCHEM)* (in press).
31. S. Miertus, V. Frecer, and M. Majekova, *J. Mol. Struc. (THEOCHEM)*, **173,** 353 (1988).
32. F. Floris and J. Tomasi, *J. Comp. Chem.*, **10,** 616 (1989).
33. R.A. Pierotti, *J. Phys. Chem.*, **67,** 1840 (1963).
34. O. Sinagoglu and O. Halicioglu, *Ann. New York Acad. Sci.*, **158,** 308 (1969).
35. M.J. Huron and P. Claverie, *J. Phys. Chem.*, **76,** 2123 (1972). **78,** 1853 (1974). **78,** 1862 (1974).
36. G.L. Amidon, S.H. Yalkowsky, and S. Leung, *J. Pharm. Sci.*, **6,** 1858 (1974).
37. C. Hansch, J.E. Quinlan, and G.L. Lawrence, *J. Org. Chem.*, **33,** 347 (1968).
38. J.M. Sorensen and W. Arlt, *Liquid-Liquid Equilibrium*, Chemistry Data Series, Vol. 5, D. Behrens, R. Eckerman, Ed. Dechema, 1979.
39. L.E. Sutton, *Table of Interatomic Distances*, Supplement, Ed. Chemical Society, London, 1965.
40. *Handbook of Biochemistry*, 2nd Edition, CRC 1970.
41. L. Pauling, *The Nature of the Chemical Bond*, Cornell University Press, Ithaca, N.Y., 1960, pp. 260–262.