Distinct Molecular Surfaces and Hydrophobicity of Amino Acid Residues in Proteins

L. F. Pacios[†]

Departamento de Química y Bioquímica, E.T.S. Ingenieros de Montes, Universidad Politécnica de Madrid, E-28040 Madrid, Spain

Received March 20, 2001

Hydrophobicity is a useful concept to rationalize the role played by amino acid residues in terms of buried or exposed conformationd with regard to the aqueous environment in proteins. The relationship of this concept with distinct approaches to represent the molecular surface is analyzed by computing reliable surface areas for three definitions namely the van der Waals, solvent-accessible, and solvent-excluded molecular surfaces. The surface areas are obtained for all of the naturally occurring amino acids by first setting a proper reference standard state and then calculating their values for a database of proteins containing a total of 4297 residues. Despite the great differences in these molecular sufaces, proper indexes are here defined for handling the information of interest to study the hydrophobic behavior of amino acids provided by such surfaces.

1. INTRODUCTION

The distribution of amino acid residues between the surface of a protein in contact with the aqueous environment and its interior where the residues are withdrawn from water to different extents is commonly analyzed in terms of the hydrophobicity concept.[1] Many attempts to classify amino acids according to their interactions with the surrounding water in proteins have been reported in the past leading to the existence of a large number of hydrophobicity scales (see for example the reviews by Cornette et al.[2] and Naray-Szabo and Balogh[3] and references therein). However, these scales display large disparities not only on the placement of specific amino acids but also sometimes on the grouping of amino acids with regard to their hydrophobicity behavior.[4] Why so many scales have been developed may be attributed on one hand to the different nature of the properties used and on the other hand to the difficulty to devise a single classification for describing the complex conformational behavior of amino acids in proteins. It is obvious that many intermolecular effects such as electrostatic interactions, hydrogen bonds, or electron density repulsions as well as steric effects are involved in the location of a particular amino acid in a protein structure. Under the term *hydrophobic effect* it has been frequently meant different interactions and entropic effects.[5] Despite its widespread use in the literature and textbooks of biochemistry to refer to a phenomenon in itself, the hydrophobicity may alternatively be viewed as just a useful term to mean the result of global differences in the intermolecular effects between the amino acid and water and those between the amino acid and other environment.[4] This is the meaning given to the term *hydrophobicity* in the rest of the paper.

Although it seems likely that no single parameter is able to represent such a complex behavior, the rapid accumulation of three-dimensional structures in the Brookhaven Protein Data Bank (PDB)[6] permits one to analyze the conformational behavior of amino acid residues to increasing accuracy. The search for correlations between some physicochemical property (usually the free energy of transfer from water to a solvent that mimics the protein interior) and estimates of surface areas has been a common approach to the understanding of the structural trends exhibited by amino acids. Although most hydrophobicity scales have been constructed following this strategy, the uncertainties in the measurements of thermodynamical properties and the difficulty to obtain molecular surface areas have sometimes hampered their reliability. Moreover, as discussed in the next section there are several ways to calculate the area covered by the irregular surface presented by a molecule to the solvent. The majority of hydrophobicity studies concerning amino acids and protein surfaces have been carried out in terms of a particular definition, namely the so-called solvent-accessible surface area (see below).

Since the publication of most hydrophobicity scales, new developments in the algorithms to compute molecular surface areas[7−14] and new measurements and theoretical studies on solvation parameters[15−19] have been reported for amino acids. Therefore, it seems worthwhile to revise the information relating hydrophobicity and surface areas. We present in this work new results for amino acid residues and explore their relationship with hydrophobicity data following the well-established approach by Rose et al.[20] In section 2 we outline the three definitions of molecular surfaces considered. Then, the procedure to set a reference state for every amino acid as well as the protein structural data set used to compute the different molecular surface areas is presented in section 3. Section 4 is devoted to analyze the information provided by these surface areas, focusing on the difficult to obtain solvent-excluded surface, in terms of hydrophobicity ideas and discuss the relationship with existing evidence on the hydropathy concept as well as recently reported free energies. Our main conclusions are finally presented in section 5.

[†] Corresponding author phone: 34 91 3367084; fax: 34 91 5439557; e-mail: lpacios@montes.upm.es.

10.1021/ci010369n CCC: $20.00
Published on Web 08/07/2001

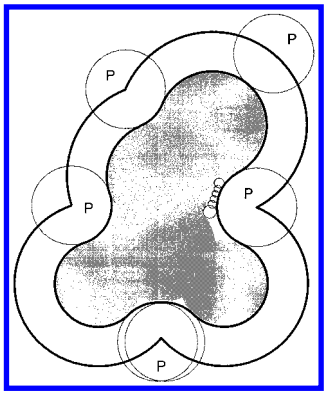

Figure 1. Molecular surface models for a set of spheres. The contour of the shaded area represents the WMS, the outer bold contour defines the SAS, and the inner bold contour is the SES. Circles labeled "P" represent probe spheres simulating the solvent.

2. MOLECULAR SURFACE AREAS

The concept of molecular surface provides an especially useful tool to study the interactions of a molecule with its surroundings. It is commonplace in computational chemistry and one of the most widely used tools in molecular modeling and graphics. The conceptually simplest surface of a molecule is the van der Waals molecular surface (WMS) resulting from a set of interlocking spheres of suitable radii centered at the positions of the atomic nuclei. The WMS should correspond to the external profile of the gray shaded area in Figure 1. Lee and Richards[21] introduced in 1971 the model of solvent accessible surface (SAS) for proteins. The SAS is the surface generated by the center of a probe sphere representing the solvent (labeled "P" in Figure 1) when it rolls over the van der Waals surface (it corresponds to the external bold profile drawn in Figure 1). Richards[22] subsequently proposed a new definition, currently known as solvent excluded surface (SES), as the locus of points traced out by the inward facing part of the probe sphere (internal bold profile in Figure 1). The SES may be viewed as composed of two parts: the contact surface resulting from the WMS which is accessible to the probe sphere and the reentrant surface defined by the inward-facing part of the probe sphere when it is simultaneously in contact with more than one atom (see Figure 1). One concise definition of the SES is the surface envelope of the volume excluded to the solvent represented by a rigid sphere (probe).[8] Some words of caution are in order regarding this terminology. Richards originally named the SES just "molecular surface", which introduced some semantic confusion in the literature. To worsen things, it is frequent to find the name "accessible surface" to refer to the SES. We employ the more rational notation given by the acronyms WMS, SAS, and SES because they are simple and convey concise information about their meaning.

Since the pioneering work of Lee and Richards, the calculation of molecular surface areas and volumes has been the focus of intense research. WMS and SAS are conceptually simple, and many algorithms, both numerical and analytical, exist for obtaining them. Although these surfaces present reeentrancies and cusps, they are trivially simple to program and render graphically. On the contrary, SES is a continuous smooth surface better suited for exploring interactions, but it is much more involved to calculate. The first and most commonly employed procedure to obtain the SES was developed in analytical form by Connolly.[23,24] In 1990 Silla et al. reported the second procedure available to obtain this surface, GEPOL,[25,26] subsequently refined and improved.[8] There have been recently proposed analytical methods such as the MSMS algorithm of Sanner et al.[10,11] which provides precise analytical surfaces and are able to deal properly with surface singularities. While Connolly's and GEPOL methods use discrete point samples to compute the surface, MSMS follows an analytical procedure based on the concept of reduced surface[10,11] to store in a compact form geometric information which can be used to build the molecular surface. The SES can further be triangulated with a vertex density given as input (see below) to handle the molecular surface.[27] For an updated report on molecular surfaces the reader is referred to the excellent review by Connolly available on the Web.[28]

We explored the performance of several algorithms after assembling a computer program which implements many options and methods for computing molecular areas and volumes, ARVOMOL.[29,30] This program includes GEPOL[8,25,26] and Connolly's MSDOT method[23,24] and has been used in applications concerning SAS and SES areas and volumes of proteins.[9,14,31] We found in these studies that GEPOL, especially in its latest implementation GEPOL93[8] included in ARVOMOL3,[30] appears to be more stable with respect to the changes in the internal parameters that allow to improve the accuracy of the SES area. Our results confirmed previous findings suggesting some oscillatory behavior in MSDOT areas when its accuracy parameter (density of points per Å^2) is increased.[25] Therefore, we have used GEPOL to compute all the surface areas reported in this work. A very brief review of the algorithm follows: the reader is referred to the original reports for a detailed account on the method.[8,25,26]

Once the set of interlocking spheres is defined to represent the molecule GEPOL computes the WMS and SAS by dividing their surfaces into 60 spherical triangular tesserae. The triangles found at the intersection volume of the spheres are then removed, and the surface areas and coordinates of the remaining tesserae are calculated with efficient algorithms. The accuracy is increased by dividing the initial set of triangles by 4 and repeating successively this process $60 \times 4^{NDIV-1}$ times, being NDIV a parameter given in input. We have used for all the calculations here reported the maximum value implemented, NDIV = 5, which implies 15360 spherical triangles per sphere. To compute the SES, GEPOL fills the spaces not accessible to the solvent with new spheres, depicted as small circles to the right of the shaded area in Figure 1. The process of creating new overlapping spheres with successively smaller radii goes on checking whether the spheres created contributes or not to the SES. The stop of this iterative process is controlled by two parameters, OFAC and RMIN. The first one measures how overlapped are the new set of spheres, the more overlapped, the smoother and more reliable will be the SES. OFAC varies thus between 0 if the spheres are not overlapped and 1 if they are totally overlapped. The second parameter, RMIN, is the radius of the smallest sphere that may be created, the smallest RMIN giving the best SES. As OFAC tends to 1 and RMIN to 0, the SES generated improves, but the number of new spheres tends to infinity, so that a

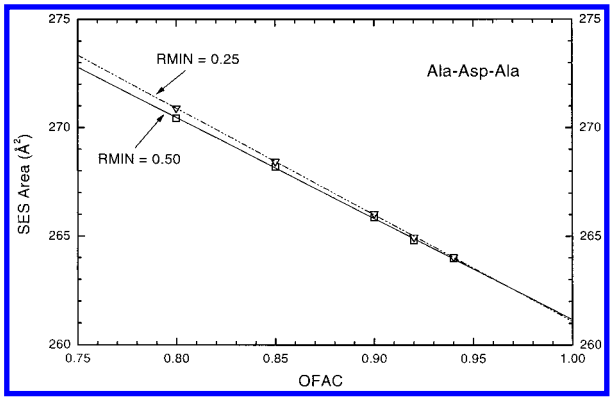

Figure 2. Dependence on the OFAC parameter of SES areas computed with the GEPOL algorithm at the NDIV = 5 level of tessellation for the tripeptide Ala-Asp-Ala using two RMIN values (see the text for the meaning of the parameters). The Richards scale of radii are used for the atomic spheres with a probe sphere of radius 1.40 Å.

compromise between accuracy and computer times must be reached. The default values recommended by the authors, OFAC = 0.80 and RMIN = 0.50 Å, have been shown to represent a good compromise.[8,9] However, we here explore their influence on the SES areas of amino acids in order to minimize the uncertainties inherent to any method used to compute molecular surface areas.

It has been demonstrated that for values of OFAC higher than 0.80, RMIN has very little effect on the computed area provided that NDIV is high enough.[8] We plot in Figure 2 the SES area of the tripeptide Ala-Asp-Ala for the geometry of the reference state discussed in the next section computed with NDIV = 5 and a probe radius of 1.40 Å to mimic the water molecule (see below). As noticed, the difference between surface areas computed with RMIN = 0.50 and RMIN = 0.25 diminishes as OFAC increases. Both sets of points converge to essentially the same result, 261.17 Å^2 for RMIN = 0.50 and 261.08 Å^2 for RMIN = 0.25, areas corresponding to OFAC = 1.0 in the best fit lines displaying correlation coefficients 0.9997 and 0.9999, respectively. This same behavior is found for a number of tripeptides explored so that in order to calculate SES areas for all the systems treated, we have kept RMIN fixed at 0.50 and compute five areas for OFAC 0.80, 0.85, 0.90, 0.92, and 0.94, taking the final result as that of OFAC = 1.0 in the linear regression fit. The value OFAC = 0.94 is the highest one that can be afforded in terms of CPU time when applied to large macromolecules with thousands of atoms.

Because this method is based on a discrete approximation, we test the reliability of SES areas computed with GEPOL by comparing them with the precise analytic surface area calculations performed with MSMS.[27] This comparison is shown in Table 1 for the 20 tripeptides Ala-X-Ala, where X represents the naturally occurring amino acids, in the geometry of the reference state presented in section 3. The GEPOL results calculated with the extrapolation to OFAC = 1.0 is compared with two MSMS values: the numerical area obtained with a triangulation density of 40 vertexes per Å^2 and the reference analytical SES area provided by this algorithm. The average relative error for GEPOL areas in this Table is 0.11%, with only three cases (His, Trp, and Tyr) showing relative errors higher than 0.20%. The numerical MSMS results correspond to a rather high density of

Table 1. SES Areas of Ala-X-Ala Tripeptides in the Geometry of the Reference State[d]

residue	GEPOL[a]	MSMS-Num[b]	MSMS-Anal[c]
Ala	237.75	237.25	237.51
Arg	314.87	314.49	314.83
Asn	259.97	259.38	259.67
Asp	257.53	257.15	257.45
Cys	254.37	254.37	254.65
Gln	278.75	278.34	278.65
Glu	276.60	276.05	276.36
Gly	221.29	220.85	221.09
His	276.94	277.27	277.57
Ile	277.99	278.00	278.29
Leu	278.16	278.29	278.59
Lys	302.66	302.12	302.44
Met	289.23	289.48	289.80
Phe	286.13	285.38	285.69
Pro	252.24	252.22	252.49
Ser	243.82	243.38	243.65
Thr	256.01	255.77	256.04
Trp	291.08	290.14	290.45
Tyr	293.33	293.65	293.97
Val	262.78	262.44	262.73

[a] Numerical GEPOL algorithm, extrapolation OFAC = 1 (see the text). [b] Numerical triangulation with a density of 40 vertices/Å^2 calculated with MSMS. [c] Analytical MSMS algorithm. [d] Set of atomic radii of Richards et al. (ref 34). Probe radius: 1.40 Å. All values in Å^2.

vertices in the triangulation procedure which happen to present the same average relative error (0.11%) than GEPOL. However, these numerical estimates can be systematically improved by increasing the triangulation density so that as it tends to infinity, the computed area tends to the analytical value. The purpose of including them in Table 1 is to illustrate the fact that GEPOL provides surface areas in very close agreement with most reliable numerical procedures as represented by this triangulation stage toward precise MSMS surface areas taken as reference limit. The comparison with the analytical results may be considered as a test of reliability of GEPOL surface areas for amino acid residues presented in the following sections obtained making use of a number of capabilities available in ARVOMOL3[30] to extract information regarding particular residues in proteins.

Another variable to analyze when computing surfaces is the set of radii used for the spheres. As far as conventional van der Waals radii[32] make no distinction among different bond environments for a given atom, several sets of atomic radii have been proposed over the years to take into account the specific atom types in proteins. Two scales proposed by Clothia[33] and Richards[34] are commonly used for protein surfaces. Six classes of atoms are defined in Clothia's set: tetrahedral carbon (1.87 Å), trigonal carbon (1.76 Å), trigonal nitrogen (1.65 Å), tetrahedral nitrogen (1.50 Å), oxygen (1.40 Å), and sulfur (1.85 Å). Nine classes are defined in Richards's set: tetrahedral carbon (2.00 Å), trigonal carbon (1.70 Å), trigonal nitrogen (1.70 Å), tetrahedral nitrogen (2.00 Å), carbonyl oxygen (1.40 Å), carboxyl oxygen (1.50 Å), hydroxyl oxygen (1.60 Å), divalent sulfur (1.85 Å), and sulfhydryl sulfur (2.00 Å). Figure 3 plots the SAS and SES areas computed with these two sets for the reference state of amino acid residues discussed in the next section. Although Richards's set yields slightly larger surface areas due to their overall larger radii, the trends in Figure 3 agree closely. Only lysine deviates about 10% for both surfaces, which is not surprising if one looks at the radius of tetrahedral

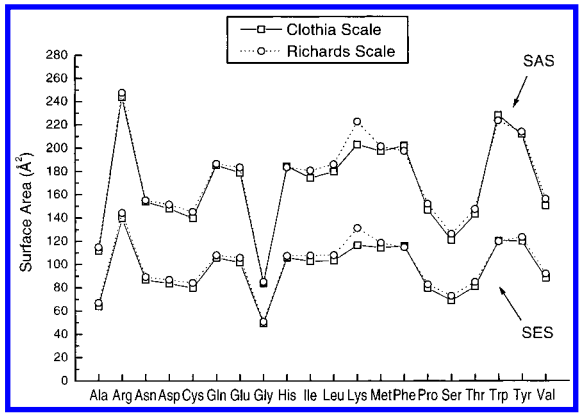

Figure 3. SAS and SES areas for the reference state of amino acids defined in the text computed with two sets of atomic radii: the scale of Clothia[33] and that of Richards.[34]

nitrogen, 2.00 Å in Richards scale and 1.50 Å in Clothia scale.

For the sake of comparison with the results by Rose et al.[20] whose procedure is here taken for reference, we compute the different surface areas of amino acid residues with the radii of Richards used by them. Moreover, the larger radii of this scale are more consistent with the correction proposed by Gerstein et al.[35] to incorporate the behavior of water around proteins. These authors investigated the distribution of interatomic distances involving water around one small protein (pancreatic trypsin inhibitor) in molecular dynamics simulations and found a different behavior of water molecules around apolar, polar, and highly charged atoms in the protein surface. To account for this behavior in a implicit manner they suggested to modify the atomic radii instead of working with three different probe radii, which should complicate considerably the calculations. For the probe sphere we have kept the customary radius of 1.40 Å to mimic a water molecule.

3. SURFACE AREAS FOR AMINO ACID RESIDUES

Following previous work on reference states for amino acids, we specify the standard state for a residue X as the average of the surface areas of X in the tripeptides Gly-X-Gly and Ala-X-Ala (see Figure 4). The influence of the geometry on the computed areas has been explored by optimizing Gly-Ala-Gly and Gly-Asp-Gly test structures in molecular mechanics calculations with two common force fields, AMBER[36] and SYBYL.[37] The optimizations were carried out with a RMS gradient of 0.1 (kcal/mol)/Å starting at initial structures assembled from the standard SYBYL library of amino acids in the molecular modeling program SPARTAN.[38] Both force fields led to optimized structures near to those of the β sheet conformation represented by backbone dihedral angles $\phi = -140°$ and $\psi = 135°$ (see Figure 4), with the values 180° and −120° for the angle ω (dihedral between carbonyls) yielding energetically close conformations. The SAS and SES areas computed for both the initial and optimized structures were very similar, exhibiting the same trends for the tripeptides studied (results not shown). We have accordingly selected the SYBYL geometries of amino acids with backbone dihedral angles $\phi = -140°$, $\psi = 135°$, $\omega = 180°$ for defining the standard state of the tripeptides Gly-X-Gly and Ala-X-Ala.

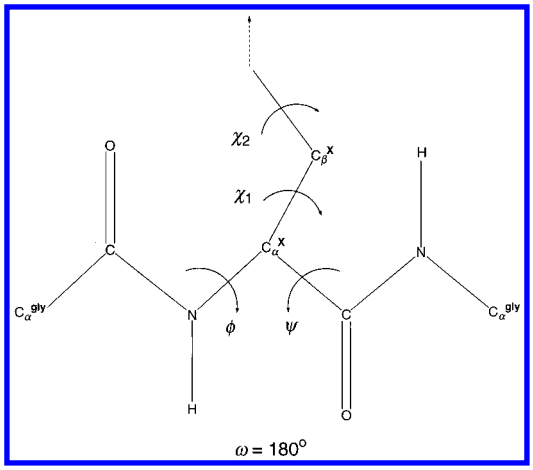

Figure 4. Tripeptide Gly-X-Gly showing the backbone dihedral angles ϕ and ψ for the conformation $\omega = 180°$, being ω the dihedral between carbonyls. χ_1 and χ_2 are side-chain dihedral angles for residue X.

Table 2. Surface Areas of the WMS, SAS, and SES for the Standard State of Amino Acid Residues[b]

residue	A^0(WMS)	A^0(SAS)	A^0(SES)	A^{0a}
Ala	72.29	114.78	67.04	118.1
Arg	156.30	247.59	144.24	256.0
Asn	100.40	155.13	89.16	165.5
Asp	98.95	151.55	86.91	158.7
Cys	93.52	145.14	84.18	146.1
Gln	118.78	186.48	108.05	193.2
Glu	116.71	183.37	105.72	186.2
Gly	53.96	85.05	50.78	88.1
His	121.81	183.43	107.31	202.5
Ile	118.93	180.60	107.75	181.0
Leu	121.51	186.00	108.00	193.1
Lys	142.10	222.64	131.20	225.8
Met	129.32	201.37	118.62	203.4
Phe	133.37	197.57	114.92	222.8
Pro	93.88	151.98	82.91	146.8
Ser	80.16	126.30	73.05	129.8
Thr	96.30	147.41	84.92	152.5
Trp	147.55	223.40	119.89	266.3
Tyr	142.20	213.67	123.25	236.8
Val	102.74	156.13	91.95	164.5

[a] Stochastic standard state of Rose et al., ref 20. [b] All values in Å².

Table 2 collects the WMS, SAS, and SES areas for our standard states of amino acid residues, A^0 as well as the SAS values reported by Rose et al.[20] for their stochastic standard state. These authors defined this stochastic state for residue X as the mean area of X in an ensemble of Gly-X-Gly tripeptides with dihedral angles taken from observed distributions in a database of 23 proteins used to calculate mean SAS areas, $<A>$, of X in the database. Since the work by Rose et al. was published, the crystallographic structures of these proteins have been updated in the Brookhaven data bank. We list in Table 3 the new PDB codes, the number of residues and atoms, and the total WMS, SAS, and SES areas for every protein. The database contains 4297 residues and 36816 atoms.

The average areas are collected in Table 4 where the total number of residues of each type and their percentage of occurrence in the database are also indicated. Two different average SAS and SES areas are given for each residue X, the mean area $<A>_M$ and the weighted area $<A>_W$. The first one is the average over the database of the mean A of X defined for every protein as

Table 3. Database of Proteins Used To Compute Surface Areas for Amino Acid Residues and Total Areas of the WMS, SAS, and SES[a]

PDB code	protein name	num. res.	num. atoms	*A*(WMS)	*A*(SAS)	*A*(SES)
2ACT	actinidin	218	1657	21345	9046	8401
3CNA	concanavalin A	237	1807	21730	10934	10673
8CPA	carboxypeptidase A	307	2437	31159	11704	11249
5CPV	calcium binding parvalbumin B	108	810	10788	5537	5043
5CYT	cytochrome C	103	803	10813	6146	5586
6EST	elastase	245	1822	24005	10568	9943
4FXC	ferredoxin	98	732	9377	5195	4489
2FCR	flavodoxin	173	1330	17408	7725	7020
4GPD	apo-D-glyceraldehyde-3-phosphate dehydrogenase	333	2507	30522	15228	14399
3HIP	high-potential iron−sulfur protein	183	1782	23259	11943	11405
2LDX	apo-lactate dehydrogenase	331	2515	31475	16389	15469
7LYZ	lysozyme	129	1000	12775	6568	5893
5MBN	myoglobin (deoxy)	153	1217	16261	8225	7709
9PAP	papain	212	1652	21319	9145	8578
5PTI	bovine pancreatic trypsine inhibitor	58	462	6059	4038	3362
5PTP	beta trypsin	245	1642	21489	9230	8797
9RSA	ribonuclease A	124	1902	24760	11903	11169
8RXN	rubredoxin	52	397	5091	3277	2737
2SBT	subtilisin	275	1934	23529	10352	9916
2SNS	staphylococcal nuclease	141	1125	13791	8044	7192
3SOD	Cu,Zn superoxide dismutase	151	1094	14139	7198	6395
8TIM	triose phosphate isomerase	248	3734	47569	19640	18750
8TLN	thermolysin	323	2455	31431	12426	11640

[a] Surface areas in Å^2.

Table 4. Average Surface Areas of the WMS, SAS, and SES for Amino Acid Residues in the Database of Proteins in Table 3[b]

	residue		WMS		SAS				SES		
	num.	$(\%)_{occ}$	<%>	<*A*>	<%>	$<A>_M$	$<A>_W$	$<A>_{Rose}$[a]	<%>	$<A>_M$	$<A>_W$
Ala	385	9.0	6.6	79.15	6.6	38.72	69.91	31.5	6.5	23.17	37.36
Arg	130	3.0	4.8	168.42	6.1	86.72	108.32	93.8	6.1	54.44	64.45
Asn	215	5.0	4.7	111.40	6.3	64.85	85.37	62.2	5.9	37.81	46.48
Asp	255	5.9	5.9	106.68	8.4	70.48	96.21	60.9	7.9	40.04	51.14
Cys	94	2.2	2.3	75.94	0.8	13.73	23.69	13.9	1.1	10.58	16.35
Gln	153	3.6	3.9	126.85	5.3	80.05	102.23	74.0	5.0	45.94	55.91
Glu	190	4.4	5.3	128.52	7.4	84.91	109.27	72.3	7.1	49.44	60.38
Gly	419	9.8	4.8	58.42	5.6	35.39	57.66	25.2	5.6	21.02	31.42
His	96	2.2	2.5	124.47	2.0	46.22	63.81	46.7	2.1	29.63	37.72
Ile	239	5.6	5.8	127.07	2.3	22.15	57.60	23.0	2.6	15.29	34.21
Leu	285	6.6	7.4	134.64	3.3	31.57	63.70	29.0	3.7	21.94	39.22
Lys	280	6.5	9.4	155.20	15.6	119.29	133.97	110.3	15.1	69.99	76.47
Met	57	1.3	1.6	125.51	0.9	29.21	40.41	30.5	1.0	19.59	25.43
Phe	135	3.1	4.3	151.78	1.5	28.31	49.67	28.7	1.8	21.46	33.51
Pro	159	3.7	3.7	99.07	4.6	61.68	89.75	53.7	4.6	37.23	50.19
Ser	337	7.8	5.5	87.76	7.4	56.84	85.36	44.2	7.0	32.49	44.70
Thr	273	6.4	5.7	105.97	6.0	55.37	81.60	46.0	6.1	33.84	45.56
Trp	71	1.7	2.5	141.12	1.1	29.77	51.36	41.7	1.3	21.35	32.93
Tyr	180	4.2	5.9	158.94	5.0	52.48	78.19	59.1	5.2	33.56	46.35
Val	344	8.0	7.3	115.02	3.9	35.83	75.92	23.5	4.3	22.80	42.98

[a] Reference 20. [b] Surface areas in Å^2.

$$<A^X>_M = \frac{\sum_i A_i^X}{N^X} \quad (1)$$

N^X being the number of residues X in that protein. The second one is the average over the database of the weighted *A* of residues X defined for every protein as

$$<A^X>_W = \frac{\sum_i f_i^X A_i^X}{\sum_i f_i^X} \quad (2)$$

where the weight factor f_i^X of each area contribution A_i^X is calculated as $f_i^X = A_i^X/A^{0X}$ and is intended to give a measure of the amount of surface exposed with respect to the standard state A^0 so that completely buried residues ($f_i^X = 0$) are left out. Because the WMS represents a simple steric measure, all the $f_i^X \cong 1$ and both averages are nearly identical for this molecular surface. The average SAS areas reported by Rose et al. computed as mean values are also included in Table 4. The <%> values for areas in the table are averages over the database of the percentages of each residue X, that is, the relative contribution of the area of X to the total area of every protein (see below).

As far as SAS areas reported by Rose et al. were obtained with a rather different methodology using a distinct reference state, the comparison of their values with our results provides insight into the own definition of this molecular surface and its validity for exploring solvation issues in proteins. Figure 5 compares the SAS areas for the stochastic state of Rose et

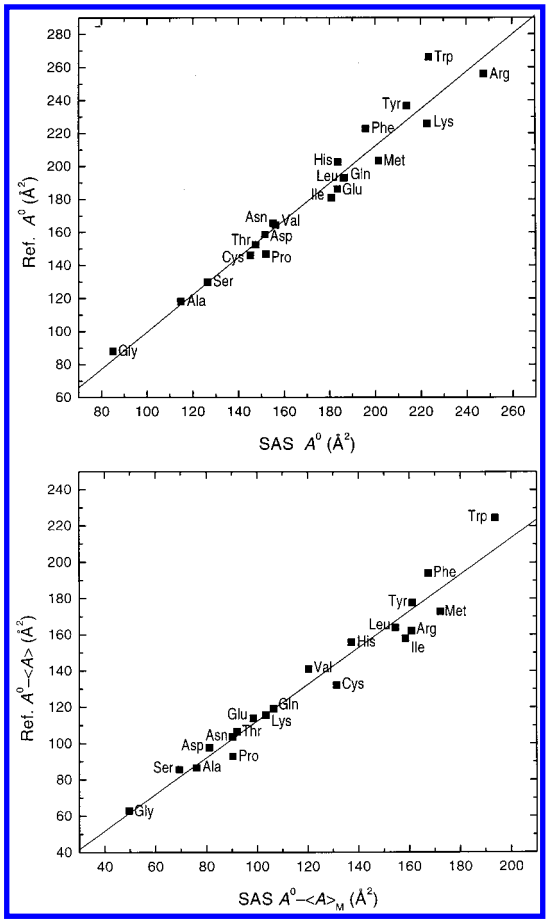

Figure 5. SAS areas A^0 for the reference stochastic state of Rose et al.[20] vs standard state defined in the text for amino acid residues (upper panel). SAS average areas buried upon folding, $A^0 - \langle A \rangle$, of Rose et al.[20] vs average mean $A^0 - \langle A \rangle_M$ values computed in this work (lower panel).

al. with our reference standard state (upper panel) and the average areas buried upon folding estimated as $A^0 - \langle A \rangle$ by the authors with our average mean $A^0 - \langle A \rangle_M$ results. The lines of best fit in Figure 5 are $-12.865 + 1.126x$ and $+11.360 + 1.010x$ with correlation coefficients 0.9765 and 0.9791, respectively. Both sets of results correlate quite well, but while the slope in the fit for buried areas is essentially 1.0, that of A^0 differ about 13% from the unity, which is expected if one takes into account the differences underlying the reference states being compared. The slope for buried areas indicate that the only difference between our results and those of Rose et al. is now a constant factor about 11 Å^2. It follows from this comparison that the buried areas upon folding estimated as $A^0 - \langle A \rangle$ could be viewed as relatively invariant parameters of the residues regardless their reference standard state.

4. HYDROPHOBICITY AND AREAS OF AMINO ACIDS

Data in Table 4 provide information on the hydrophobicity of amino acid residues that can be rationalized in several forms. On the one side, let us consider the percentages of relative contribution to the total areas compared with the percentages of occurrence of every residue in the database of proteins. To make this comparison we take the difference between these two percentages, $\Delta p_{MS} = \langle \% \rangle_{MS} - (\%)_{occ}$, as a simple relative contribution index which is listed in Table 5 for the three molecular surfaces. On the other side, let us consider the fraction of exposure given by $r = \langle A \rangle_W / \langle A \rangle_M$ as an estimate of the amount of buried residues with regard to the two molecular surfaces accounting for the solvent. This ratio happens to be very similar in the SAS and SES and shows a maximum r_{max} in isoleucine (2.60 for SAS, 2.24 for SES) and a minimum r_{min} in lysine (1.12 for SAS, 1.09 for SES). As for the database of proteins used, these results point to isoleucine as the most hydrophobic amino acid, i.e., that presenting the least number of exposed residues, and to lysine as the most hydrophilic amino acid, i.e., nearly all its residues are exposed. This ratio is converted to a scale between 0.0 and 1.0 under the transformation $f = 1.0 - (r_{max} - r)/(r_{max} - r_{min})$ so that $f = 1.0$ for Ile and $f = 0.0$ for Lys. This hydrophobicity factor is given for the SAS and SES in Table 5, where the amino acids are arranged in ascending order of f_{SES}.

Table 5. SES and SAS Hydrophobicity Factors (f_{SxS}), Relative Contributions of SES, SAS, and WMS Areas (Δp_{xxS}), and Hydropathy Index of Amino Acid Residues

residue	f_{SES}	f_{SAS}	Δp_{SES}	Δp_{SAS}	Δp_{WMS}	hydropathy[a]
Lys	0.00	0.00	8.6	9.1	2.9	−3.9
Arg	0.08	0.09	3.1	3.1	1.8	−4.5
Glu	0.11	0.11	2.7	3.0	0.9	−3.5
Gln	0.11	0.11	1.4	1.7	0.3	−3.5
Asp	0.12	0.13	2.0	2.5	0.0	−3.5
Asn	0.15	0.16	0.9	1.3	−0.3	−3.5
His	0.16	0.18	−0.1	−0.2	0.3	−3.2
Met	0.18	0.18	−0.3	−0.4	0.3	1.9
Thr	0.23	0.24	−0.3	−0.4	−0.7	−0.7
Pro	0.23	0.23	0.9	0.9	0.0	1.6
Tyr	0.25	0.25	1.0	0.8	1.7	−1.3
Ser	0.25	0.26	−0.8	−0.4	−2.3	−0.8
Gly	0.35	0.34	−4.2	−4.2	−5.0	−0.4
Trp	0.39	0.41	−0.4	−0.6	0.8	−0.9
Cys	0.40	0.41	−1.1	−1.4	0.1	2.5
Phe	0.41	0.43	−1.3	−1.6	1.2	2.8
Ala	0.45	0.47	−2.5	−2.4	−2.4	1.8
Leu	0.61	0.61	−2.9	−3.3	0.8	3.8
Val	0.70	0.68	−3.7	−4.1	−0.7	4.2
Ile	1.00	1.00	−3.0	−3.3	0.2	4.5

[a] Reference 39.

Let us first analyze the percentages of relative contribution. For the van der Waals molecular surface, Δp_{WMS} may be viewed as an estimate of the relative steric contribution and displays two extreme values, −5.0 for glycine and +2.9 for lysine. Being as how glycine is the smallest amino acid as well as the more abundant in the database of proteins, the negative difference is the obvious result. As noticed in Table 5, lysine is the amino acid giving by far the largest contribution to the three surface areas. Among the amino acids with large side chains, Lys has the third largest WMS area after Arg and Tyr (see Table 4), while its occurrence is higher than these amino acids which explains its large Δp_{WMS}. Much more interesting is the information conveyed by Δp_{SAS} and Δp_{SES}. Notice the close resemblance and parallel ordering between both percentage differences. Charged amino acids display positive Δp values above 2.0, that is, their larger contribution to total surface areas as compared to their occurrence reveal their strong hydrophilic behavior. There follows a group of polar amino acids with Δp_{SAS} values between 1.7 and −1.4 and Δp_{SES} values between 1.4 and −1.1, including the aromatic Tyr and Trp. For tyrosine $\Delta p_{SAS} = 0.8$ and $\Delta p_{SES} = 1.0$, positive values

due to its hydroxyl group, while for tryptophan $\Delta p_{SAS} = -0.6$ and $\Delta p_{SES} = -0.4$, negative values suggesting a much lower polarity like that due to nitrogen in the indole ring. Nonpolar phenylalanine and residues with aliphatic groups display more negative Δp values indicative of their hydrophobic character. Amino acids with sulfur deserve particular comments. Methionine is essentially nonpolar so that one should expect more negative values for Δp. However, its small contributions to SAS and SES areas (about 1% in both cases) happens to compensate its low occurrence, 1.3%, the lowest percentage in the database. The slightly polar cysteine presents mostly in the form of disulfide bonds frequently buried in the interior of proteins and therefore withdrawn from the aqueous solvent, which explains its position near nonpolar amino acids in Table 5.

Despite the disparities in the definition of the SAS and SES which indeed lead to the rather distinct areas in Table 4, the consistent behavior of amino acids with regard to both solvent sensitive surfaces is especially highlighted by the hydrophobicity factor. The remarkable agreement between f_{SAS} and f_{SES} in Table 5 allows one to choose only *one* factor, say f_{SES}, to analyze the information given by *both* SAS and SES areas. There is a first group of charged and strongly polar amino acids with $f_{SES} < 0.20$ including Met (see the comments in the preceding paragraph). A group of moderately polar amino acids then follows with similar f_{SES} values between 0.23 and 0.25. After a gap of 0.10 units, a group formed by Gly, Cys, Ala, and aromatic Trp and Phe displays f_{SES} between 0.35 and 0.45, and finally the most hydrophobic amino acids with larger aliphatic side chains show the largest f_{SES} values. The behavior of Cys suggested by this factor (see also the comments in the preceding paragraph) agrees with the evidence on its hydrophobic nature recently reported.[17]

The well-known hydropathy index of Kyte and Doolittle[39] is also included in Table 5. This is a widespread used scale combining in a continuous manner hydrophilic (negative values) and hydrophobic (positive values) residues. This index is compared with our f_{SES} and Δp_{SES} values in Figure 6, where the horizontal lines split amino acids into three blocks as suggested by the values of the hydropathy index. If one looks at the relative differences in the values listed in Table 5, three gaps in the f_{SES} index are noticed at Thr (0.23), Gly (0.35), and Leu (0.61). Accordingly, the vertical lines in the upper plot in Figure 6 divide values of f_{SES} into four intervals, (0.0, 0.23), (0.23, 0.35), (0.35, 0.61), and (0.61, 1.0), while the vertical line in the lower plot just separates positive and negative relative contributions given by Δp_{SES}. Note how the hydropathy index divides the residues into three groups, a hydrophobic set with positive values, an intermediate group composed by Gly, Ser, Thr, Trp, and Tyr, with values between 0.0 and −2.0, and a hydrophilic set with negative values below −3.0. Our indexes f_{SES} and Δp_{SES} separate unambiguously the more hydrophobic Ala, Ile, Leu, Phe, Trp, and Val residues and the more hydrophilic Arg, Asn, Asp, Gln, Glu, His, and Lys residues. Pro, Ser, Thr, and Tyr residues are undoubtedly located in the intermediate group, but some considerations must be pointed out before definitely locating Cys, Gly, and Met.

The frequent presence of cysteine in the form of disulfide bonds withdrawn from the aqueous environment is the reason of its very small SAS and SES areas in Table 4. It seems

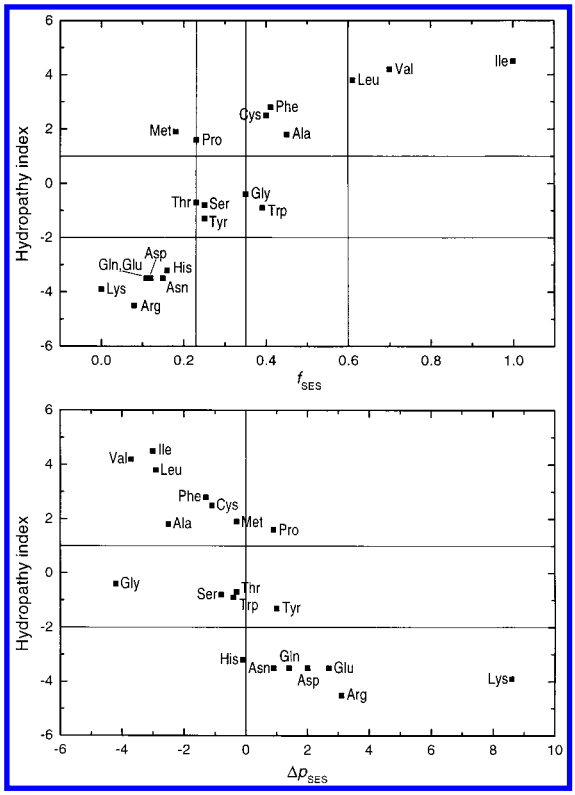

Figure 6. Hydropathy index of Kyte and Doolittle[39] vs SES hydrophobicity factor f_{SES} (upper panel) and relative contribution index Δp_{SES} (lower panel).

therefore reasonable including Cys within the hydrophobic group. The great occurrence of glycine in the database of proteins (actually it has the largest percentage in Table 4) is the reason of the large negative values of Δp_{SAS} or Δp_{SES}. However, Gly displays moderate contributions to SAS and SES areas so that its presence within the intermediate group as suggested by its f_{SES} factor seems most appropriate. Methionine displays a real intermediate behavior: if one looks at the percentages in Table 4, the relative contribution of this amino acid in occurrence (1.3%), van der Waals surface (1.6%), and both solvent surfaces (0.9% for SAS, 1.0% for SES) is very similar, so that its role is neither hydrophilic nor hydrophobic.

After these considerations, the analysis of the information provided by the SAS and SES areas leads to the division of the amino acid residues into three groups:

(a) hydrophilic (charged and very polar): Arg, Asn, Asp, Gln, Glu, His, and Lys;

(b) intermediate (moderately polar): Gly, Met, Pro, Ser, Thr, and Tyr;

(c) hydrophobic (nonpolar): Ala, Cys, Ile, Leu, Phe, Trp, and Val.

The f_{SES} factor in Table 5 sets the relative position of every amino acid within these groups. This classification agrees with that proposed by Rose et al.[20] after comparing their SAS A^0 values with the average area buried upon folding $A^0 - \langle A \rangle$. The only differences are their location of Pro in group (a) and His in group (b).

Finally, we explore the relationship of our surface area results with recent experimental free energies $\Delta G°$ measured by Wimley and White[15] for studying the energetics of protein-bilayer interactions. Two scales of $\Delta G°$ were determined, one for the transfer of protein chains from water to

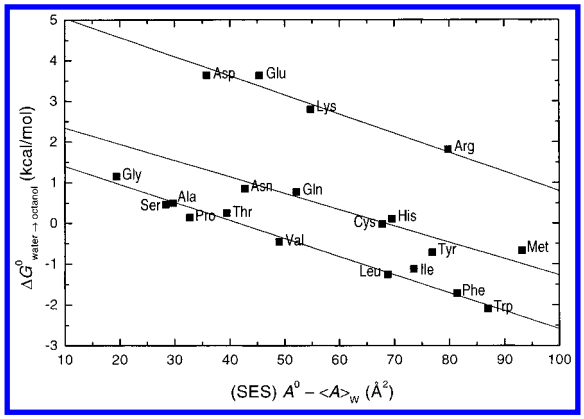

Figure 7. Whole-residue free energies $\Delta G°$ of transfer from water to *n*-octanol[15,16] vs weighted average SES areas buried upon folding $A^0 - <A>_W$ calculated in this work.

the bilayer interface and one for the transfer from the aqueous solvent to the bilayer interior evaluated as the transfer from water to *n*-octanol. Both sets of energies include peptide bonds since whole residues, not just side chains, partition into membranes.[16] As far as our surface areas refer to whole amino acid residues, these transfer free energies are thus appropriate for reference purposes. Figure 7 plots the whole-residue free energy of transfer from water to *n*-octanol[15,16] against weighted average SES areas buried upon folding, i.e., $A^0 - <A>_W$ values. This plot divides amino acids into three groups: a group of charged residues Asp, Glu, Lys, and Arg (correlation coefficient 0.9744), a group of polar and moderately polar residues Asn, Gln, His, Tyr, and Met, including Cys (correlation coefficient 0.9370), and a group of hydrophobic amino acids including Ser and Thr (correlation coefficient 0.9884). Since this scale of $\Delta G°$ intends to estimate the relative stability between the solvent environment and the apolar bilayer interior, charged residues are clearly separated from the rest displaying the most unfavorable free energies with a rate of change in Figure 7 determined by their relative size. Amino acids with aromatic and large aliphatic side chains give place to favorable $\Delta G°$ values in the bilayer interior while moderately polar amino acids with small side chains such as Ser and Thr show less favorable interactions and are near the smallest residues (Ala, Pro and Val). Residues containing sulfur as well as more polar amino acids display intermediate free energies between +1 and −1 kcal/mol. The SES manages to correctly measure the hydrophilicity behavior, while it is better suited than the SAS to describe the bulk size of the molecule (see Figure 1), which makes the comparison in Figure 7 particularly sensitive.

Figure 8 plots the two whole-residue scales of Wimley and White against the SES hydrophobicity factor f_{SES}. This factor rationalizes the behavior of amino acids by splitting them into separate groups for both $\Delta G°$ sets, that of transfer from water to the bilayer interface (upper plot) and that of transfer from water to *n*-octanol (lower plot). The hydrophilic residues in group (a) are located at the upper left box characterized in both plots by positive values of $\Delta G°$ and f_{SES} values between 0.0 and 0.2. Methionine appears as an intermediate residue heading the group (b) above (see Table 5) being now the only residue in the lower left box corresponding to the same f_{SES} interval yet with a very little negative free energy. The upper two middle boxes contain

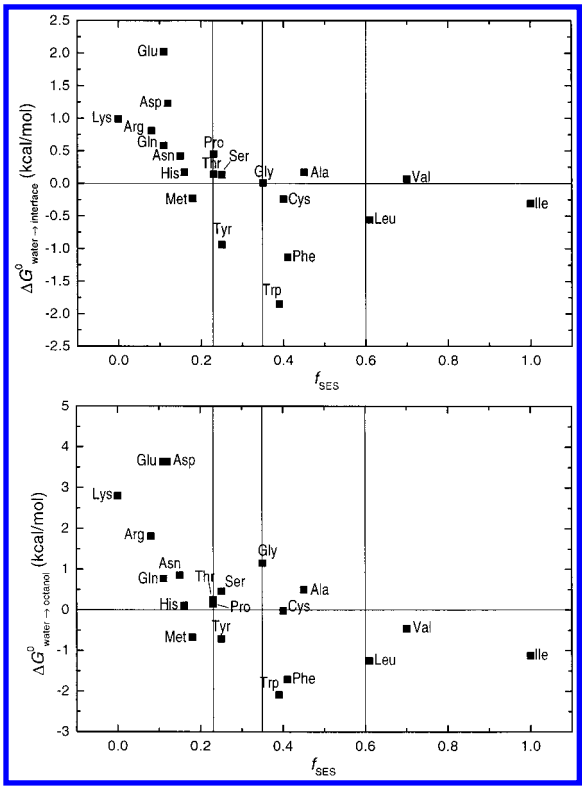

Figure 8. Whole-residue free energies $\Delta G°$ of transfer from water to bilayer interface (upper panel) and $\Delta G°$ of transfer from water to *n*-octanol[15,16] (lower panel) vs SES hydrophobicity factor f_{SES}.

the moderately polar residues in group (b) except the aromatic Tyr which is located at the lower box between $f_{SES} = 0.23$ and 0.35 with obvious favorable free energy for the transfer to the bilayer interface and the bilayer interior. Finally, the two boxes on the right collect the hydrophobic residues forming the group (c) with alanine displaying $\Delta G°$ values near 0.0 kcal/mol and the most hydrophobic Phe (aromatic) and Ile and Leu (large aliphatic side chains) showing favorable free energies in both transfers from the aqueous solvent. Cysteine appears in Figure 8 at a location heading the hydrophobic set in accordance with the considerations regarding this amino acid presented before.

5. CONCLUSIONS

The complex conformational behavior of an amino acid residue with regard to the aqueous solvent in a protein is usually rationalized in terms of the hydrophobicity concept. As far as many intermolecular effects are implied in that behavior, no single universal hydrophobicity scale exists for the naturally occurring amino acids, although it is obvious that the hydrophobicity can be related with the contribution of a residue to the protein surface area. This relationship has been explored in this paper by studying different ways to represent the molecular surface and deriving essential common information from the distinct approaches to this representation.

Surface areas for the three definitions presented in section 2 have been computed at the highest level of precision allowed by GEPOL93, a numerical algorithm whose reliability compare well with precise reference analytical calculations. The dependence of the computed surface areas on the atomic radii used as well as the parameters of the

calculation, especially critical in the obtention of the SES, has been studied to render reliable surface areas. This optimized methodology has been applied to the reference states of amino acids as well as to the database of 23 proteins presented in section 3.

The analysis of the two solvent sensitive molecular surfaces (SAS and SES) has demonstrated that despite the great differences in their definitions that in turn lead to disparate surface areas for a molecule given, it is possible to find a consistent common information conveyed by both surfaces with regard to the hydrophobicity behavior. Two indexes have been devised from the calculated SES areas to extract this information.

A classification of the amino acid residues in terms of their hydrophobic/hydrophilic character has been obtained by using exclusively the information provided by these indexes. Three hydrophobicity groups arise from this analysis:

(a) hydrophilic group consisting of charged and strongly polar amino acids;

(b) intermediate group consisting of Tyr and weakly polar amino acids;

(c) hydrophobic group consisting of aromatic Phe and Trp as well as residues with aliphatic side chains.

The ordering within each group is established by the indexes derived from the SES areas which are in complete agreement with those of SAS areas. The results presented in section 4 agree well with existing information like the hydropathy index and whole-residues free energies of transfer recently measured to study the changes of a protein from water to the membrane bilayer.

ACKNOWLEDGMENT

The author gratefully acknowledges financial support from the Dirección General de Enseñanza Superior e Investigación Científica, Project No PB97-0268.

REFERENCES AND NOTES

(1) Kyte, J. *Structure in Protein Chemistry*; Garland Publishing: New York, 1995; Chapters 5 and 6.
(2) Cornette, J.; Cease, K. B.; Margalit, H.; Spouge, J. L.; Berzofsky, J. A.; DeLisi, C. Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *J. Mol. Biol.* **1987**, *195*, 659−685.
(3) Naray-Szabo, G.; Balogh, T. The average molecular electrostatic field as a QSAR descriptor. Hydrophobicity scales for amino acid residues. *J. Mol. Struct. (THEOCHEM)* **1993**, *284*, 243−248.
(4) Charton, M.; Charton, B. I. The structural dependence of amino acid hydrophobicity parameters. *J. Theor. Biol.* **1982**, *99*, 629−644.
(5) Clothia, C. Principles that determine the structure of proteins. *Annu. Rev. Biochem.* **1984**, *53*, 537−572.
(6) Bernstein, F. C.; Koetzle, T. F.; Williams, J. B.; Meyer, E. F.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. The Protein Data Bank. A computer-based archival file for macromolecular structures. *J. Mol. Biol.* **1977**, *112*, 535−542.
(7) Grand, S. L.; Merz, K. M. Rapid approximation to molecular surface area via the use of boolean logic and look-up tables. *J. Comput. Chem.* **1993**, *14*, 349−352.
(8) Pascual-Ahuir, J. L.; Silla, E.; Tuñon, I. GEPOL: An improved description of molecular surfaces. III. A new algorithm for the computation of a solvent-excluding surface. *J. Comput. Chem.* **1994**, *15*, 1127−1138.
(9) Pacios, L. F. Variation of surface areas and volumes in distinct molecular surfaces of biomolecules. *J. Mol. Model.* **1995**, *1*, 46−53.
(10) Sanner, M. F.; Olson, A. J.; Spehner, J. C. Fast and robust computation of molecular surfaces. *ACM 11th. Symp. Comput. Geom.* **1995**, *C6*−*C7*, Vancouver, Canada.
(11) Sanner, M. F.; Olson, A. J.; Spehner, J. C. Reduced surface: an efficient way to compute molecular surfaces. *Biopolymers* **1996**, *38*, 305−320.
(12) Gerstein, M.; Clothia, C. Packing at the protein-water interface. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 10167−10172.
(13) Eisenhaber, F.; Argos, P. Hydrophobic regions on protein surfaces: definition based on hydration shell structure and a quick method for their computation. *Prot. Eng.* **1996**, *9*, 1121−1133.
(14) Liang, J.; Edelsbrunner, H.; Fu, P.; Sudhakar, P. V.; Subramanian, S. Analytical shape computation of macromolecules: I. Molecular areas and volume through alpha shape. *Proteins: Struct. Funct. Genet.* **1998**, *33*, 1−17.
(15) Wimley, W. C.; White, S. H. Experimentally determined hydrophobicity scale for proteins at membrane interfaces. *Nature Struct. Biol.* **1996**, *3*, 842−848.
(16) White, S. H.; Wimley, W. C. Membrane protein-folding and stability: physical principles.*Annu. Rev. Biophys. Biomolec. Struct.* **1999**, *28*, 319−365.
(17) Nagano, N.; Ota, M.; Nishikawa, K. Strong hydrophobic nature of cysteine residues in proteins. *FEBS Lett.* **1999**, *458*, 69−71.
(18) Smith, B. J. Solvation parameters for amino acids. *J. Comput. Chem.* **1999**, *20*, 428−442.
(19) Dima, R. I.; Settanni, G.; Micheletti, C.; Banavar, J. R.; Maritan, A. Extraction of interaction potentials between amino acids from native protein structures. *J. Chem. Phys.* **2000**, *112*, 9151−9166.
(20) Rose, G. D.; Geselowitz, A. R.; Lesser, G. J.; Lee, R. H.; Zehfus, M. H. Hydrophobicity of amino acid residues in globular proteins. *Science* **1985**, *229*, 834−838.
(21) Lee, B. K.; Richards, F. M. The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **1971**, *55*, 379−400.
(22) Richards, F. M. Areas, volumes, packing, and protein structures. *Annu. Rev. Biophys. Bioeng.* **1977**, *6*, 151−176.
(23) Connolly, M. L. Solvent-accessible surfaces of proteins and nucleic acids. *Science* **1983**, *221*, 709−713.
(24) Connolly, M. L. Computation of molecular volume. *J. Am. Chem. Soc.* **1985**, *107*, 1118−1124.
(25) Pascual-Ahuir, J. L.; Silla, E. An improved description of molecular surfaces. I. Building the spherical surface set. *J. Comput. Chem.* **1990**, *11*, 1047−1060.
(26) Silla, E.; Tuñon, I.; Pascual-Ahuir, J. L. An improved description of molecular surfaces. II. Computing the molecular area and volume. *J. Comput. Chem.* **1991**, *12*, 1077−1088.
(27) Sanner, M. F. Molecular Surfaces Computation. MSMS program at the Web site http://www.scripps.edu/pub/olson-web/people/sanner/html/msms_home.html.
(28) Connolly, M. L. Molecular surfaces: a review; Network-Science, 1996. (Available at the Web site http://www.netsci.org/Science/Compchem/feature14.html).
(29) Pacios, L. F. ARVOMOL/CONTOUR: Molecular surface areas and volumes on personal computers. *Comput. Chem.* **1994**, *18*, 377−385.
(30) Pacios, L. F. ARVOMOL-3: Surface AReas and VOLumes of MOLecules. Quantum Chemistry Program Exchange: Program QCMP132. *QCPE Bull.* **1998**, *18*, 4−5.
(31) Pacios, L. F. A numerical study on the effective dimension of protein surfaces. *Chem. Phys. Lett.* **1995**, *242*, 325−332.
(32) Gavezzotti, A. The calculation of molecular volumes and the use of volume analysis in the investigation of structured media and of solid-state organic reactivity. *J. Am. Chem. Soc.* **1983**, *105*, 5220−5225.
(33) Clothia, C. Structural invariants in protein folding. *Nature* **1975**, *254*, 304−308.
(34) Richards, F. M. Calculation of molecular volumes and areas for structures of known geometries. *Methods Enzymol.* **1985**, *115*, 440−464.
(35) Gerstein, M.; Tsai, J.; Levitt, M. The volume of atoms on the protein surface: calculated from simulation using Voronoi polihedra. *J. Mol. Biol.* **1995**, *249*, 955−966.
(36) Weiner, S. J.; Kollman, P. A.; Nguyen, D. T.; Case, D. An all atom force field for simulations of proteins and nucleic acids. *J. Comput. Chem.* **1986**, *7*, 230−252.
(37) *SYBYL*; Tripos Inc.: 1699 South Hanley Rd., St. Louis, MO, U.S.A.
(38) *PC SPARTAN Plus*, *V.1.4*; Wave function Inc.: 18401 Von Karman Ave., Irvine, CA, U.S.A.
(39) Kyte, J.; Doolittle, R. F. A simple method for displaying the hydropathic character of a protein, *J. Mol. Biol.* **1982**, *157*, 105−132.

CI010369N