OA Biology
Open Access

Review

Structural

Solvent accessible surface of globular proteins: how exposed and buried amino acid residues are divided

O Carugo*

Abstract

Introduction

Globular protein structures are often divided into two regions, the surface, which is in contact with the surrounding molecules, and the interior, which is not accessible to the external molecules. Since most of the surrounding molecules, both *in vitro* and *in vivo*, are water molecules, it is clear that protein surface is essentially polar in order to make the protein molecule soluble and stable in an aqueous environment, whereas the protein interior tends to be formed by apolar and hydrophobic amino acid residues. Despite this simple classification criterion, amino acids residues are usually identified, on the basis of the three-dimensional structures, according to a myriad of different criteria, some of which are briefly reviewed in the present article. A simple and ecumenical criterion can be based on the fact that the best separation of the two regions (surface and interior) must be reflected by the maximisation of the difference in hydrophobicity between the two regions. In this article, this simple criterion is examined on the basis of a statistical survey of an ensemble of carefully selected protein crystal structures.

Conclusion

In this article, it is shown that with the use of information published in scientific journals and disseminated in databases, it is easy to give a sound and quantitative framework to a fundamental concept like that of the amino acid polarity/apolarity antinomy.

Introduction

All textbooks on protein structure and chemistry state that globular proteins are stabilised, in their native and folded state, by the spatial separation of their apolar residues, which cluster in the protein interior, from their polar residues, which are disseminated at the protein surface and are then in contact with the aqueous solution where proteins exist and exert their function[1]. Although not all proteins are globular, for example, membrane or intrinsically disordered proteins, the segregation of polar and apolar residues is one of the major driving forces that allow life to exist. Thanks to that, globular proteins are in fact stable, soluble and functional in an aqueous environment. Their functions, in particular, are generally ensured by the residues that are at their surface and can be accessed not only by water molecules but also by a variety of other molecules, including substrates, inhibitors, cofactors and etc.

Not surprising, a myriad of studies were focused on protein surfaces, including procedures of computational docking[2], identification of functional protein–protein interfaces[3] and analysis of packing interactions in the solid state[4,5]. A common feature of all these studies is the need of a quantitative criterion to distinguish the two types of amino acid residues, those that are buried in the protein interior and those that are exposed to the solvent. Several pioneering studies were focused on techniques to measure numerically the area of the solvent accessible surface (*SASA*) of each atom and residue within a protein[6–9]. For this purpose, it is assumed, in general, that a water molecule occupies a spherical portion of the space of radius equal to 1.4 Å, although smaller radius values have been proposed[10].

Once it is possible to compute the area of the surface accessible to the solvent, the decision whether a residue is exposed to the solvent should be naïve: if the area is larger than 0, the residue is accessible while it is buried in the opposite case. However, the decision is somehow more complex. On the one hand, it is necessary to consider that different residues have different dimensions and so also different *SASA* areas. For example, triptophane is much larger and has a much larger surface than alanine. On the other hand, it is possible to select a threshold value different from 0. In fact, given that the area calculations are numerical, they cannot be systematically exact and, as a consequence, a larger threshold might be preferable to avoid false positives among the surface residues. The aim of this review is to discuss how exposed and buried amino acid residues are divided in the *SASA* of globular proteins.

Discussion

Identification of surface residues

For this reason, different criteria are used by different scientists. For example, according to Levy, an amino acid was considered to be at the protein surface if its relative solvent accessibility was higher than 25%[11]. In this case, the relative solvent accessibility of the *i*th residue, *SASA ref*(*i*), was defined as

* Corresponding author
Email: olicar04@unipv.it

Chemistry Department, Pavia University, Viale Taramelli 12, I-27100 Pavia, Italy

Competing interests: none declared. Conflict of interests: none declared.
All authors contributed to conception and design, manuscript preparation, read and approved the final manuscript.
All authors abide by the Association for Medical Ethics (AME) ethical rules of disclosure.

For citation purposes: *Carugo O.* Solvent accessible surface of globular proteins: how exposed and buried amino acid residues are divided. OA Biology 2013 Nov 01;1(1):1.

$$SASA\_rel(i) = \frac{SASA(i)}{REF}$$

where *SASA*(*i*) is the solvent accessible surface area of that residue and *REF* is a reference value necessary to standardise the *SASA* values by considering the different dimensions among the amino acid residues. Such a reference value is the *SASA* of the same type of residue measured in Gly-X-Gly tripeptide, taken from the work of Miller and colleagues[7]. While Levy computed the *SASA* values by using the AREAIMOL program of the CCP4 software suite[12], Zellner and co-workers computed them[13] with a different software library, BALL[14]. Apart from this difference, the *SASA_rel* values were computed exactly in the same way. However, Zellner and colleagues decided to consider accessible to the solvent the residues with a *SASA_rel* larger than 5%, whereas Levy preferred to use a threshold of 25%.

SASA_rel values were also used by Rost and Sander[15], though this time the *SASA* values were computed with DSSP[16]. However, Rost and Sander used different, arbitrary thresholds. Residues were classified in two categories: they were considered to be buried in the protein interior if their relative solvent accessibility was smaller than 16% and they were considered to be at the protein surface if their relative solvent accessibility was higher than (or equal to) 16%. Alternatively, a three-state model of accessibility was adopted: a residue was considered to be buried if its relative solvent accessibility was less than 9%, it was considered to be solvent accessible if its relative solvent accessibility was more than (or equal to) 36% and it was considered to be intermediate otherwise. Residues were classified with a binary scheme buried/exposed also by Pugalenthi and colleagues to predict residue accessible surface area from protein sequence information: for each residue, the buried and exposed states were computed using five, alternative thresholds (0%, 5%, 10%, 25% and 50%)[17].

Contrary to the previously mentioned studies, Duarte and colleagues did not use the relative *SASA* area value but considered that the solvent exposed the residues with at least 5 Å^2 of their surface exposed to the solvent (surface accessible solvent area values were computed with ASAPY[18])[3].

Obviously, a binary classification in buried/accessible residues might seem rather coarse, since the degree of solvent accessibility of a residue can be a variable, ranging from residues that have very little accessibility, perhaps with only one atom that can be truly solvated by water, to residues that are completely accessible to the solvent and can be surrounded by numerous water molecules. As a consequence, one might prefer a different approach, in which solvent accessibility can be handled as a real, continuous variable. For example, Scherrer and colleagues[19] published a model of sequence evolution that explicitly accounts for the solvent accessibility of each residue in a protein, where the evolutionary rate varies linearly with the solvent accessibility. Also to study the relationship between chemical shifts and atom solvent accessibility, it was necessary to use real *SASA* area values[20].

Improvement strategy

It is evident that some confusion and divergent criteria are routinely used to identify the residues that are at the surface of globular proteins and, in order to try to solve this controversy, it is better to go back to the basic ideas of protein chemistry: the folding of a long polypeptide is mainly dictated by the necessity to segregate most of the apolar residues in the protein interior while keeping most of the polar residue in contact with water. As a consequence, the best *SASA* or *SASA_rel* value is the value that can separate an apolar interior and a polar surface.

Certainly, it is preferable to use the relative *SASA* area values, since they take into account the diverse dimensions of different amino acids. Moreover, it is necessary to limit the attention to monomeric, globular proteins, since the surface of oligomeric proteins is partially buried into inter-molecular interfaces. Further, the structures containing too many (e.g. 5%) heteroatoms different from water molecules must be eliminated because they can seriously alter the polarity of both the interior and the surface of the proteins. It is also necessary to focus attention on high-quality crystal structures, by removing those with a resolution worse than 2.0 Å and with missing atoms or residues, and to avoid redundancy, for example, with a maximal pairwise percentage of sequence identity of 25%. With these criteria, it is possible to assemble a group of 501 protein structures from the Protein Data Bank[21,22] that contains 142,663 residues.

It is then mandatory to verify if the computations of *SASA* area are independent of the algorithm and of the computer programs that are used. For this purpose, three different computer programs were compared–AREAIMOL[12], DSSP[16] and Naccess[23]. On average, the *SASA* values of Naccess and AREAIMOL differ by 3.11(4) Å^2, those of Naccess and DSSP by 5.45(4) Å^2, and those of AREAIMO and DSSP by 5.60(4) Å^2. The *SASA_rel* values of Naccess and AREAIMOL differ by 2.99(1)–the comparison with DSSP is impossible since this program does not compute these values. Since the average agreement among these three programs is excellent, only one of them can be used. The rest of the data presented in this article were computed with Naccess.

Given a certain threshold value for *SASA_rel*, it is possible to classify all the amino acid residues into two groups, one containing the residues that are buried in the protein interior

Competing interests: none declared. Conflict of interests: none declared.
All authors contributed to conception and design, manuscript preparation, read and approved the final manuscript.
All authors abide by the Association for Medical Ethics (AME) ethical rules of disclosure.

and the other with the amino acids that are exposed to the solvent at the protein surface. It is then possible to compute the average hydrophobicity for each of the two groups and the mean difference (*DHy*) between the hydrophobicity of the buried residues and that of the surface residues. Among the numerous hydrophobicity scales that are available, a consensus scale was used for the data shown below[24].

Figure 1 shows the average hydrophobicity of the protein core and of the protein surface at *SASA_rel* threshold values ranging from 0 to 60 (only proteins containing 100–200 residues were considered to make the figure). If the threshold is very small, for example, 0, all the residues are considered exposed to the solvent with the exception of the few that are completely buried. As a consequence, the hydrophobicity of the core is extremely high and also the hydrophobicity of the surface is relatively big, since some residues, considerably apolar and nearly inaccessible to the solvent, are considered to be part of the surface. If the *SASA_rel* threshold increases, less and less amino acids are considered to be at the surface, whereas the number of core residues increases. Some apolar and nearly inaccessible residues, previously assigned to the protein surface, move into the group of the core residues. As a consequence, the hydrophobicity of the core gradually decreases like that of the protein surface. Similar trends can be observed also for proteins smaller or larger than those used to prepare (Figure 1).

The difference between the average hydrophobicity of the core and of the surface is depicted in Figure 2, again for proteins containing 100–200 amino acids. The *DHy* values increase slightly if the *SASA_rel* threshold goes from 0 to 3 and then gradually decrease. At large threshold values, the difference between the average hydrophobicity of the core and of the surface decreases more steeply.

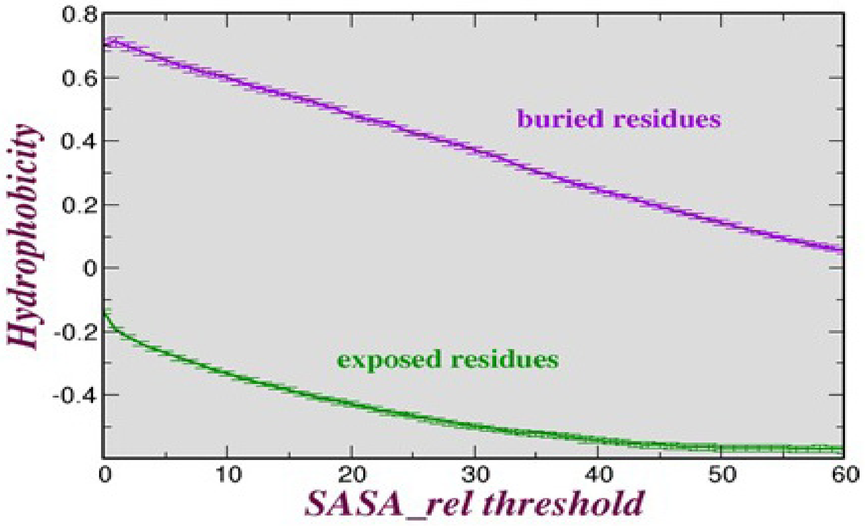

Figure 1: Dependence of the average hydrophobicity of the buried and exposed amino acids on the *SASA_rel* value used to differentiate protein interior and protein surface. The figure was prepared by considering proteins containing 100–200 amino acids and the standard errors are indicated by vertical bars.

Similar trends can be observed also for proteins of different dimensions.

The maximal value of *DHy* in Figure 2, 0.932(18), is observed at the *SASA_rel* threshold equal to 5. All the *DHy* values are computed for *SASA_rel* threshold values ranging from 1 to 27 and are statistically identical to

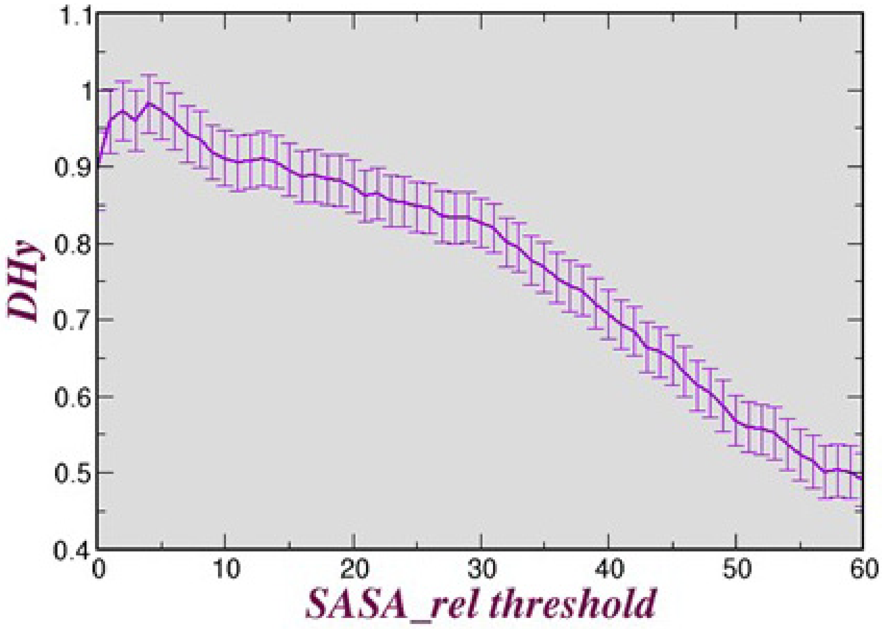

Figure 2: Relationship between the *SASA_rel* value used to differentiate protein interior and protein surface and the difference between the average hydrophobicities of the buried and exposed residues (Dhy). The figure was prepared by considering proteins containing 100–200 amino acids and the standard errors are indicated by vertical bars.

Competing interests: none declared. Conflict of interests: none declared.
All authors contributed to conception and design, manuscript preparation, read and approved the final manuscript.
All authors abide by the Association for Medical Ethics (AME) ethical rules of disclosure.

For citation purposes: *Carugo O.* Solvent accessible surface of globular proteins: how exposed and buried amino acid residues are divided. OA Biology 2013 Nov 01;1(1):1.

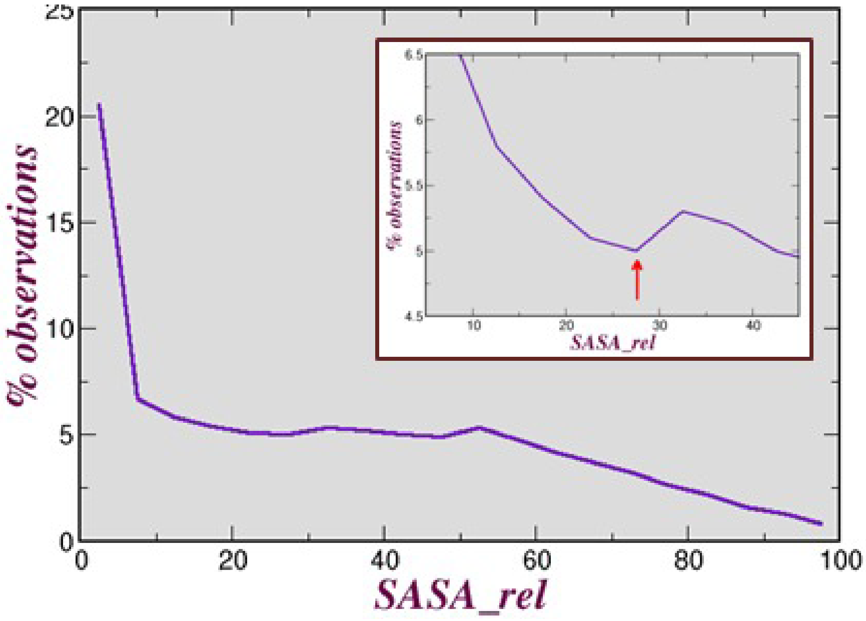

Figure 3: Distribution of the *SASA_rel* values for proteins containing 100–200 amino acids. The minimum at *SASA_rel* value around 25–30 is shown by an arrow in the inset.

the maximal value. It is then possible to conclude that any value of *SASA_rel* from 1 to 27 can be used as a threshold that allows one to maximise the difference of hydrophobicity of the protein surface and of the protein core. It is also possible to identify the same type of plateau of *DHy* values in protein containing less than 100 or more than 200 residues. These plateaus are shown in Table 1. They tend to be narrower for small proteins and to increase if the protein gets bigger. They are systematically quite large and extend from 0 to about 20 for small proteins and from 0 to more than 40 for larger proteins.

It may appear quite odd that the ranges of *SASA_rel* values that allow the maximisation of the difference of hydrophobicity between the protein core and surface are so large. However, by looking at the distribution of the *SASA_rel* values, it is possible to make an interesting observation. In Figure 3, where the distribution of the *SASA_rel* values for proteins containing 100–200 residues is depicted, one can note that the frequency of observations is higher at low *SASA_rel* values and tends to decrease at higher *SASA_rel* values. Along this trend, there is a discontinuity at SASA_rel values close to 25–30, where the frequency of observations does not decrease anymore and where, on the contrary, it tends to increase slightly,

Table 1 Ranges of *SASA_rel* that allow one to maximise the difference of hydrophoby between protein core and surface in proteins of diverse dimensions

Number of residues	*SASA_rel* values range of the plateau	Number of residues	*SASA_rel* values range of the plateaus
0–100	0–17	300–400	1–40
20–120	1–22	320–420	1–38
40–140	1–21	340–440	1–38
60–160	1–24	360–460	1–36
80–180	1–29	380–480	1–37
100–200	1–27	400–500	1–39
120–220	1–28	420–520	2–39
140–240	1–26	440–540	1–42
160–260	1–27	460–560	6–44
180–280	1–28	480–580	3–44
200–300	1–28	500–600	1–48
220–320	1–30	520–620	1–47
240–340	1–33	540–640	1–47
260–360	1–37	560–660	1–48
280–380	1–37	580–680	0–41

For citation purposes: *Carugo O.* Solvent accessible surface of globular proteins: how exposed and buried amino acid residues are divided. OA Biology 2013 Nov 01;1(1):1.

Competing interests: none declared. Conflict of interests: none declared.
All authors contributed to conception and design, manuscript preparation, read and approved the final manuscript.
All authors abide by the Association for Medical Ethics (AME) ethical rules of disclosure.

before to decrease again at higher *SASA_rel* values. The interruption in the monotonical decrease reflects the antinomy between polarity and apolarity, with apolar residues that tend to be buried in the protein interior and polar residues that tend to be at the protein surface. Of course, it is impossible to perfectly separate the two types of residues and, as a consequence, there are two (or more) distributions that are partially superposed, one at low *SASA_rel* values and the second at higher *SASA_rel* values. The minimum observed *SASA_rel* values close to 25–30 is due to the partial superposition of two distributions and corresponds to the best *SASA_rel* value that can separate apolar from polar residues.

Conclusion

Although *SASA*s are a useful concept to describe the thermodynamic bases of globular protein folding and stability, they are frequently handled in a rather qualitative and arbitrary manner. In this article, it is shown that with the use of information published in scientific journals and disseminated in databases, it is easy to give a sound and quantitative framework to a fundamental concept like that of the amino acid polarity/apolarity antinomy.

Abbreviation list

SASA, solvent accessible surface.

References

1. Lesk AM. (Introduction to protein science: architecture, function, and genomics. Oxford: Oxford University Press; 2010.
2. Poupon A, Janin J. Analysis and prediction of protein quaternary structure. Methods Mol Biol. 2010;609:349–64.
3. Duarte JM, Srebniak A, Schärer MA, Capitani G. Protein interface classification by evolutionary analysis. BMC Bioinformatics. 2012 Dec;13:334.
4. Carugo O, Argos P. Protein-protein crystal-packing contacts. Protein Sci. 1997 Oct;6(10):2261–3.
5. Janin J. Specific versus non-specific contacts in protein crystals. Nat Struct Biol. 1997 Dec;4(12):973–4.
6. Lee B, Richards FM. The interpretation of protein structures: estimation of static accessibility. J Mol Biol. 1971 Feb;55(3):379–400.
7. Miller S, Janin J, Lesk AM, Chothia C. Interior and surface of monomeric proteins. J Mol Biol. 1987 Aug;196(3):641–56.
8. Rose GD, Geselowitz AR, Lesser GJ, Lee RH, Zehfus MH. Hydrophobicity of amino acid residues in globular proteins. Science. 1985 Aug;229(4716):834–8.
9. Shrake A, Rupley JA. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. J Mol Biol. 1973 Sep;79(2):351–71.
10. Hubbard SJ, Argos P. Cavities and packing at protein interfaces. Protein Sci. 1994 Dec;3(12):2194–206.
11. Levy ED. A simple definition of structural regions in proteins and its use in analyzing interface evolution. J Mol Biol. 2010 Nov;403(4):660–70.
12. Winn MD, Ballard CC, Cowtan KD, Dodson EJ, Emsley P, Evans PR, et al. Overview of the CCP4 suite and current developments. Acta Crystallogr D Biol Crystallogr. 2011 Apr;67(Pt 4):235–42.
13. Zellner H, Staudigel M, Trenner T, Bittkowski M, Wolowski V, Icking C, et al. PresCont: predicting protein-protein interfaces utilizing four residue properties. Proteins. 2012 Jan;80(1):154–68.
14. Hildebrandt A, Dehof AK, Rurainski A, Bertsch A, Schumann M, Toussaint NC, et al. BALL—biochemical algorithms library 1.3. BMC Bioinformatics. 2010 Oct;11:531.
15. Rost B, Sander C. Conservation and prediction of solvent accessibility in protein families. Proteins. 1994 Nov;20(3):216–26.
16. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983 Dec;22(12):2577–637.
17. Pugalenthi G, Kandaswamy KK, Chou KC, Vivekanandan S, Kolatkar P. RSARF: prediction of residue solvent accessibility from protein sequence using random forest method. Protein Pept Lett. 2012 Jan;19(1):50–6.
18. Ho BK. Emporium of words, Peddling the frenzied recollections of an unrepentant bioinformatician & degenerate webmonkey, 2007 Nov. http://boscoh.com/protein/asapy.
19. Scherrer MP, Meyer AG, Wilke CO. Modeling coding-sequence evolution within the context of residue solvent accessibility. BMC Evol Biol. 2012 Sep;12:179.
20. Rieping W, Vranken WF. Validation of archived chemical shifts through atomic coordinates. Proteins. 2010 Aug;78(11):2482–9.
21. Berman HM , Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The protein data bank. Nucleic Acids Res. 2000 Jan;28(1):235–42.
22. Bernstein FC, Koetzle TF, Williams GJ, Meyer EF Jr, Brice MD, Rodgers JR, et al. The Protein Data Bank: a computer-based archival file for macromolecular structures. J Mol Biol. 1977 May;112(3):535–42.
23. Hubbard SJ, Campbell SF, Thornton JM. Molecular recognition. Conformational analysis of limited proteolytic sites and serine proteinase protein inhibitors. J Mol Biol. 1991 Jul;220(2):507–30.
24. Carugo O. Prediction of polypeptide fragments exposed to the solvent. In Silico Biol. 2003;3(4):417–28.

Competing interests: none declared. Conflict of interests: none declared.
All authors contributed to conception and design, manuscript preparation, read and approved the final manuscript.
All authors abide by the Association for Medical Ethics (AME) ethical rules of disclosure.