Maximum Allowed Solvent Accessibilites of Residues in Proteins

Matthew Z. Tien1, Austin G. Meyer2,3, Dariya K. Sydykova4,5, Stephanie J. Spielman4,5, Claus O. Wilke3,4,5

1 Department of Biochemistry & Molecular Biology, The University of Chicago, Chicago, Illinois, United States of America, 2 School of Medicine, Texas Tech University Health Sciences Center, Lubbock, Texas, United States of America, 3 Institute for Cellular and Molecular Biology, The University of Texas at Austin, Austin, Texas, United States of America, 4 Center for Computational Biology and Bioinformatics, The University of Texas at Austin, Austin, Texas, United States of America, 5 Department of Integrative Biology, The University of Texas at Austin, Austin, Texas, United States of America

Abstract

The relative solvent accessibility (RSA) of a residue in a protein measures the extent of burial or exposure of that residue in the 3D structure. RSA is frequently used to describe a protein’s biophysical or evolutionary properties. To calculate RSA, a residue’s solvent accessibility (ASA) needs to be normalized by a suitable reference value for the given amino acid; several normalization scales have previously been proposed. However, these scales do not provide tight upper bounds on ASA values frequently observed in empirical crystal structures. Instead, they underestimate the largest allowed ASA values, by up to 20%. As a result, many empirical crystal structures contain residues that seem to have RSA values in excess of one. Here, we derive a new normalization scale that does provide a tight upper bound on observed ASA values. We pursue two complementary strategies, one based on extensive analysis of empirical structures and one based on systematic enumeration of biophysically allowed tripeptides. Both approaches yield congruent results that consistently exceed published values. We conclude that previously published ASA normalization values were too small, primarily because the conformations that maximize ASA had not been correctly identified. As an application of our results, we show that empirically derived hydrophobicity scales are sensitive to accurate RSA calculation, and we derive new hydrophobicity scales that show increased correlation with experimentally measured scales.

Introduction

Relative solvent accessibility (RSA) has emerged as a commonly used metric describing protein structure in computational molecular biology, with the particular application of identifying buried or exposed residues. It is defined as a residue’s solvent accessibility (ASA) normalized by a suitable maximum value for that residue. RSA was first introduced in the context of hydrophobicity scales derived by computational means from protein crystal structures [1–5]. More recently, RSA has been shown to correlate with protein evolutionary rates and has been incorporated as a parameter into models which determine these rates [6–13]. As RSA straightforwardly characterizes the local environment of residues in protein structures, many studies have developed computational methods to predict RSA from protein primary and/or secondary structure [14–20]. Further applications of RSA include identification of surface, interior, and interface regions in proteins [21], protein-domain prediction [22], and prediction of deleterious mutations [23].

To derive a residue’s RSA from its surface area, an ASA normalization factor is needed for each amino acid. By convention, these normalization values have been derived by evaluating the surface area around a residue of interest X when placed between two glycines, to form a Gly-X-Gly tripeptide. Most commonly, the normalization values utilized are those previously calculated by either Rose et al. [2] or Miller et al. [3]. The primary distinction between these two sets of normalization values lies in the different θ and ψ dihedral backbone angles chosen when evaluating Gly-X-Gly tripeptide conformations. Rose et al. [2] considered tripeptides with backbone angles representing an average of observed θ and ψ angles, whereas Miller et al. [3] considered tripeptides in the extended conformation (θ = −120°, ψ = 140°).

As the number of empirically determined 3D protein crystal structures has grown over the years, it has become apparent that neither the Rose [2] nor the Miller [3] scale accurately identifies the true upper bound for a residue’s ASA. In fact, virtually all amino acids display, on occasion, ASA values in excess of the normalization ASA values provided by either scale. Some do so quite frequently (e.g. R, D, G, K, P), reaching RSA values of up to 1.2. This discrepancy, which leads to RSA values > 1, is generally known in the field though rarely acknowledged in print. One exception is a recent study that carried out an extensive empirical survey of ASA values in PDB structures [20]. That study found that the most accessible conformations are generally found in loops and turns, not in the extended conformation, and it suggested to
use conformation-dependent maximum ASA values for normalization [20].

Here, we derive a new set of ASA normalization values that provide a tight upper bound on ASA values observed in biophysically realistic tripeptide conformations. To calculate these normalization values, we pursue two complementary strategies—one empirical and one theoretical. For the empirical approach, we mined thousands of 3D crystal structures and recorded the maximum ASA values we found for each amino acid across all structures. For the theoretical approach, we computationally built Gly-X-Gly tripeptides and systematically evaluated all biophysically allowed conformations to determine a maximum theoretical ASA value. These two strategies yield congruent results and ultimately produce comparable normalization scales that tightly bound ASA for all 20 amino acids. We then return to the historic motivation for RSA and investigate the implications of our results for hydrophobicity scales. We find that ASA normalization affects the performance of empirically derived hydrophobicity scales, and we propose new scales that show improved correlation with experimentally measured scales.

Results

Published ASA normalization values are too small

We initially assessed the accuracy of Rose’s [2] and Miller’s [3] ASA normalization scales through an exhaustive survey of the ASA values found in experimentally determined protein structures. We obtained a list of 3197 high-quality PDB structures from the PISCES server [24]. We then calculated ASA for each residue in all 3197 structures, excluding any chain-terminating residues. ASA values were subsequently normalized using the scales of either Rose et al. [2] or Miller et al. [3] to obtain RSA. For either scale and each amino acid, we found that residues with RSA > 1 were not uncommon (Figure 0); RSA values exceeded unity by up to 20%. The amino acids that most commonly displayed RSA > 1 were R, D, G, K, P. For those amino acids, RSA values > 1 occurred at frequencies of 1% to 3% of all residues, depending on the normalization scale used (Figure 1).

To determine the underlying factors leading to RSA > 1, we examined the association between RSA and the following factors: residue neighbors, secondary structure, bond lengths, bond angles, and dihedral angles. For most of these quantities, we found no strong association with RSA. We did, however, find a clear association with residues’ \( \phi \) and \( \psi \) backbone angles. For example, consider the Ramachandran plot of alanine (Figure 2). A noticeable cluster of high-RSA residues falls into the \( \alpha \)-helix region of \( \phi \approx -50, \psi \approx -45 \). We therefore propose that this region contains the maximally exposed conformation of alanine and should be used for calculating maximum ASA.

![Ramachandran plot for alanine residues in our empirical data set.](image)

Figure 1. Frequency of residues with RSA > 1 in empirical protein structures. Nearly all amino acids, and notably R, D, K, G, and P, show RSA > 1 when RSA is calculated using the normalization values of either Rose et al. [2] or Miller et al. [3].

doi:10.1371/journal.pone.0080635.g001

Figure 2. Ramachandran plot for alanine residues in our empirical data set. Coordinates which correspond to RSA values > 1 are shown in red and are clearly concentrated around coordinates (−50, −45). We therefore propose that this region contains the maximally exposed conformation of alanine and should be used for calculating maximum ASA.

doi:10.1371/journal.pone.0080635.g002
To derive maximum ASA values for each amino acid X, we computationally constructed Gly-X-Gly tripeptides and systematically rotated them through all biophysically allowed conformations (see Methods and Text in File S1 for details.) When constructing the tripeptides, we set bond lengths and angles (excluding $\omega$, $\theta$, $\psi$, and $\chi$ angles) for each amino acid equal to the average values observed for that amino acid in our reference set of 3197 PDB structures. We set $\omega = 180^\circ$. We then rotated the $\theta$ and $\psi$ around the X residue in discrete 1° steps, exhaustively enumerating all conformations. Additionally, we iterated through all rotamer angles $\chi$ that were sterically possible with each $(\theta, \psi)$ combination. For those amino acids with more than 10 possible distinct rotamer conformations, as determined by the Dunbruck database [24], we evaluated ten randomly chosen rotamer conformations. We recorded the maximum ASA in each bin.

Next, we compared the resulting theoretical maximum ASA values to the empirically observed maximum ASA values. We binned both the theoretical and the empirical values into discrete $5 \times 5$ bins of $(\theta, \psi)$ and recorded the maximum ASA in each bin. To eliminate nonexistent or rare conformations, we defined four Ramachandran regions for each amino acid: CORE, containing at least 80% of the empirical observations; ALLOWED, containing at least 97% of the empirical observations; GENEROUS, extending the core region by 20° in all directions; and ALL, containing all non-empty bins. The definitions of the CORE, ALLOWED, and GENEROUS regions are consistent with the definitions used in Ref. [25]. For each region, we displayed the maximum ASA value in each bin in side-by-side Ramachandran plots (Figure 3 and Figures S1–S3 in File S1) and generally found good congruence between the theoretical and the empirical values for all amino acids. Regions that had the highest maximum ASA in the theoretical data set also had the highest maximum ASA in the empirical data set. The highest ASA values were generally observed in the $\alpha$-helix region of the Ramachandran plot (Figure 3). Based on these results, we propose new maximum ASA values (Table 1 and Table S1 in File S1) and maximally exposed geometries for each amino acid (Table S2 in File S1).

We further evaluated our model’s performance by directly comparing theoretical and empirical maximum ASA values in each $(\theta, \psi)$ bin. We calculated the difference between these two values for each $5 \times 5$ bin (now including all bins with at least one observation in the empirical data set). We then plotted this difference against the number of empirical observations obtained for each bin (Figure 4). We found that with increasing amounts of empirical data, this difference approached zero; the maximum ASA values from both approaches converged as more data was available. Moreover, even for sparsely populated bins, at least some bins showed a difference near zero, regardless of the number of observations in each bin. Therefore, while our results did improve with increasing amounts of data, they were also largely robust to smaller data sets.

As Table S1 shows, the maximum ASA values observed in the empirical data set were nearly identical for different Ramachandran regions. Scales for the ALLOWED, GENEROUS, and ALL regions were identical, with the exception of a $1 \AA^2$ difference for Val between ALLOWED and GENEROUS/ALL. The scale for the CORE region was nearly identical as well, with most
differences on the order of 1–2 Å². The only larger difference (15 Å²) arose for Cys, the rarest amino acid in our data set. For the theoretical scales, we similarly found that differences between the CORE and ALLOWED regions were minor, typically on the order of 2–5 Å². The biggest difference again arose for Cys. Theoretical maximum values in the GENEROUS and ALL regions were up to 10–15 Å² larger than in the ALLOWED region, and generally substantially larger than the largest ASA values observed in the entire empirical data set. We conclude from this finding that the GENEROUS and ALL regions are too permissive of unphysical and/or rare backbone conformations, and we recommend that the maximum ASA values of the ALLOWED region be used in actual applications. Table 1 summarizes these values and compares them to the previously published scales by Miller et al. [3] and Rose et al. [2]. All results in the remainder of this work were derived using the scales obtained for the ALLOWED region.

**Figure 3. Ramachandran plots for empirical and theoretical maximum ASA values of alanine.** (A) Empirical maximum ASA values for each 5° by 5° bin. All bins in the ALLOWED region are shown. (B) Theoretical maximum ASA values, as determined by computational modeling, shown for non-empty bins in (A). Both the empirical and the theoretical approach find the largest ASA values in the α-helix region around (–50, –45°). By contrast, the extended conformation (–120°,140°) leads to relatively low maximum ASA.

doip://10.1371/journal.pone.0080635.g003

**Figure 4. Difference between theoretically and empirically determined maximum ASA values for alanine, across 5° by 5° bins.** As the amount of data per bin increases, the difference between theoretical and empirical maximum ASA approaches zero, demonstrating that our two methods converged with increasing amounts of data. Furthermore, the difference between values is frequently close to zero, indicating that our two methods converged with increasing amounts of data. This observation indicates that our theoretically derived maximum ASA values provide a tight bound on the empirically observed ones.

doip://10.1371/journal.pone.0080635.g004

**Relation to Empirically Derived Hydrophobicity Scales**

The solvent exposure of an amino acid, averaged over many occurrences of that amino acid in many different protein structures, should correlate with the amino acid’s hydrophobicity. Therefore, solvent exposure has long been used as a means to empirically derive hydrophobicity scales from protein crystal structures [1,2]. In particular, Rose et al. [2] derived a hydrophobicity scale by calculating the mean ASA for each amino acid across a set of reference crystal structures, using the ASA normalization values derived in the same work [2]. Since those normalization values are inaccurate, as shown above, we assessed how using our normalization values would alter the Rose hydrophobicity scale.

We first compared the Rose scale to a number of experimentally derived scales (Table 2, [26–32]). We included in the list of experimental scales the scale by Kyte & Doolittle [27], which is a hybrid scale partially based on solvent-accessibility data from protein structures, and the scale by Mac Callum et al. [30], which is based on molecular-dynamics simulations. A brief description of each scale is given in the legend to Table 2. The Rose scale correlated reasonably well (50%–70% of variance explained) with most experimental scales. It correlated the highest with the scale of Fauchere & Pliska [28] (82% of variance explained) and it did not correlate significantly with the scales of Wimley et al. [32] and of Mac Callum et al. [30] (Table 2).

We next derived two scales based on mean ASA, calculated using either our theoretical or our empirical ASA normalization values (Table S3 in File S1). Both of our mean ASA scales correlated well with the Rose scale (r = 0.96 and r = 0.97, respectively, with P < 10⁻¹⁰ in both cases) but were not identical to it. The biggest difference arose for histidine, which is ranked as the 8th-most hydrophobic amino acid according to the Rose scale but as the 10th- or 13th-most hydrophobic amino acid, respectively, according to our scales. Our scales correlated more strongly than the Rose scale with all experimental scales except the Mac Callum scale, which did not correlate significantly with either our or the Rose scale (Table 2). For the majority of experimental scales, the percent variance explained increased by approximately 10 percentage points using our normalization over the Rose normalization. We can conclude from these results that mean ASA is a useful measure of amino acid hydrophobicity and that correct ASA normalization is required to assign appropriate hydrophobicity scores to all amino acids.
Table 2. Absolute value of correlation coefficients $r$ between empirically derived and experimentally derived hydrophobicity scales.

| Experimental scale | Mean RSA (Rose)$^a$ | Mean RSA (theor)$^b$ | Mean RSA (emp)$^c$ | $100\%$ buried$^d$ | $95\%$ buried$^e$ |
|--------------------|---------------------|----------------------|-------------------|-------------------|-------------------|
| Wolfenden et al.$^f$ | 0.614              | 0.681               | 0.681             | 0.827             | 0.774             |
| Kyte & Doolittle$^g$ | 0.841              | 0.879               | 0.881             | 0.953             | 0.948             |
| Radzicka & Wolfenden$^h$ | 0.852          | 0.855               | 0.851             | 0.844             | 0.888             |
| Moon & Fleming$^i$ | 0.704              | 0.748               | 0.752             | 0.678             | 0.764             |
| Fauchere & Pliksa$^j$ | 0.904              | 0.906               | 0.910             | 0.734             | 0.878             |
| Wimley et al.$^m$ | 0.463$^y$          | 0.464               | 0.473             | 0.323$^y$         | 0.417$^y$         |
| MacCallum et al.$^k$ | 0.27$^y$          | 0.265$^y$           | 0.285$^y$         | 0.116$^y$         | 0.227$^y$         |

The largest significant correlation in each row is highlighted in bold.

$^a$Mean RSA of residues in protein structures, as calculated by Rose et al. [2].
$^b$Mean RSA of residues in protein structures, as given in column 2 of Table S3 in File S1.
$^c$Mean RSA of residues in protein structures, as given in column 3 of Table S3 in File S1.
$^d$Fraction of $100\%$ buried residues, as given in column 4 of Table S3 in File S1.
$^e$Fraction of $95\%$ buried residues, as given in column 5 of Table S3 in File S1.
$^f$Transfer energy of pentapeptides between octanol and water [32].
$^g$Transfer energy between octanol and water [28].
$^h$Transfer energy calculated from molecular-dynamic simulations of side-chain analogs within a bilayer [30].
$^i$Transfer energy from cyclohexane to water [29].
$^j$Transfer energy from vapor to water [26].
$^k$Hybrid scale based on transfer energy from vapor to water and on the percentages of $95\%$ and $100\%$ buried residues in protein structures [27].
$^l$Correlation not statistically significant; all other correlations are significant at $\alpha = 0.05$. doi:10.1371/journal.pone.0080635.0002

Discussion

We have derived significantly improved ASA normalization values. Our normalization values provide a tight upper bound to the largest observed ASA values in empirical structures. By contrast, previously published ASA normalization values were too small, by up to 20%, and frequently led to RSA values $>1$. We estimated the maximum allowed ASA for each amino acid by computationally modeling Gly-X-Gly tripeptides, where X is the amino acid of interest, and exhaustively surveying ASA over all biophysically feasible conformations. We found that maximally exposed conformations tend to fall into the $\alpha$-helix region of Ramachandran plots, and that extended conformations display some side-chain burial. The results of our modeling approach were consistent with maximum ASA values found by surveying over 3000 empirical protein crystal structures. We also revisited the problem of deriving empirical hydrophobicity scales from protein structures. We found that improved ASA normalization values lead to improved empirical hydrophobicity scales. Further, scales based on both mean RSA and on the fraction of buried residues correlated well with experimentally measured scales. Overall, the fraction of $95\%$ buried residues seems to be the best-performing empirical hydrophobicity scale, but mean RSA correlates well with an experimental scale based on side-chain transfer between octanol and water.

Our method of obtaining ASA normalization values was similar to the methods employed by Rose et al. [2] and by Miller et al. [3]. Rose et al. [2] calculated their ASA normalization values by computing the ASA of residue X in Gly-X-Gly tripeptides whose conformations were chosen based on the average dihedral angles from available empirical data at the time. Miller et al. [3], on the other hand, calculated their ASA normalization values by computing the ASA of an extended trimer structure with $\beta = -120^\circ$, $\phi = 140^\circ$ and with side-chain conformations that were frequently observed in the empirical data. The key distinction between these previous approaches and ours lies in our exhaustive sampling of tripeptide conformations. By modeling all biophysically feasible discrete combinations of $\beta$ and $\phi$ angles and varying rotamers, we identified the ideal conformations which yield maximum allowed ASA. To pursue our modeling strategy, we developed a program that allowed us to easily construct peptide chains from scratch in arbitrary conformations (see Text in File S1 for details).

Our results are broadly consistent with a recent paper by Singh and Ahmad [20]. These authors did an extensive empirical survey...
of ASA values in tripeptides from PDB structures. They found that
the highest observed ASA values were found in loops and turns,
not in the extended conformation used by Miller et al. Their
highest ASA values are generally consistent with ours. Further,
Singh and Ahmad found that the highest observed ASA values
were dependent on the neighboring residues around the focal
residue. Finally, Singh and Ahmad showed that for RSA
prediction from primary sequence, prediction accuracy could be
improved by approximately 10% if ASA values were normalized
by (neighbor-dependent) highest observed ASA values rather than
by ASA values observed in the extended conformation [20]. Our
work serves as a useful complement to their work, by (i) providing,
through molecular modeling, highest possible ASA values rather
than just highest observed ASA values, by (ii) providing highest
observed and highest possible ASA values as a function of
backbone dihedral angle, and by (iii) demonstrating that improved
RSA normalization yields empirical hydrophobicity scales that are
more similar to experimentally measured ones.

In our modeling approach, we calculated ASA values for Gly-X-
Gly tripeptides. Other authors have considered normalizations
based on Ala-X-Ala tripeptides [18,33] or even neighbor-specific
normalizations (i.e., a different normalization for each specific
tripeptide [20]). We chose Gly-X-Gly tripeptides because we
wanted to calculate the highest possible ASA values of tripeptides,
and glycines will generally occlude less solvent than alanines. From
a practical perspective, we prefer a simple normalization scheme,
and hence highest possible ASA values are attractive to us. However,
for certain applications, it may be the case that neighbor-specific or
backbone-specific normalizations are preferable. Singh and Ahmad
[20] provided neighbor-specific normalization values, but didn’t
control for backbone angles. We have shown here that maximum
ASA values depend substantially on backbone angles (e.g., Fig. 3),
and we provide both highest observed and highest possible ASA
values as a function of backbone angles (see “Data and code
availability” in Methods). It is not known at this time whether
neighbor-dependent or backbone-dependent normalization is
preferable, and the answer may depend on the specific application.
In principle, one could also normalize by both neighboring amino
acids and backbone dihedral angles. A modeling approach such as
ours could be employed to calculate the highest possible ASA values
for any tripeptide in any conformation. The computational
resources required would be substantial, however, since we would
have to model 400 times more tripeptides than we did for the
present work.

Our theoretical modeling approach to exhaustively survey
tripedptides has two potential shortcomings. First, for bond lengths
and angles (except major dihedral angles), we used mean values
observed in a large number of protein crystal structures. This
approach neglects the variation around the mean, and there could
be rare cases where unusually large bond lengths or unusual bond
angles might cause ASA to become larger than estimated here.
Such scenarios would have to be exceedingly rare, however, since
we did not find a single case in which the largest empirically
derived maximum ASA value exceeded the largest theoretically
derived maximum ASA value (Table 1). Second, for amino acids
with more than 10 distinct rotamer conformations, we did not
exhaustively enumerate all possible conformations but only
sampled 10 conformations at random. Thus, in principle it is
possible that we missed a particular rotamer conformation that
would have corresponded to a larger ASA value than the
maximum we observed. Two arguments suggest that this issue is
not likely a major source of error. First, again, we did not find a
single case in which the empirical maximum ASA was larger than
the theoretical maximum ASA. Second, maximum ASA varied
slowly with θ and ψ, and by exhaustively enumerating conforma-
tions in 1' steps, in effect we sampled the most exposed
conformations multiple times, thus reducing the chance of missing
a rare, large-ASA conformation.

As our RSA calculations are based on ASAs of tripeptides, we
excluded all chain terminating residues from both the empirical
and the theoretical analysis. Even with our improved ASA normaliza-
tion values, then, chain-terminating residues may still display
RSA > 1. We therefore recommend that future analyses making use
of RSA similarly exclude any chain-terminating residues, as their
RSA estimates will not be precise. Suitable normalization values for
chain-terminating residues are not available at present.

The normalization values we have derived here are, strictly
speaking, only valid for solvent-accessible surface areas calculated
with the DSSP program [34]. However, more generally, we expect
them to be correct as long as solvent accessibility is calculated
according to the definition of Lee and Richards [35], which
assumes that a sphere of radius 1.4A is rolled over the surface of
the molecule. For cases in which solvent accessibility is calculated
differently, our results suggest that one can follow an empirical
approach to normalization. In other words, one need not
exhaustively evaluate tripeptides, as we have done here. Instead,
one can obtain a representative sample of structures from the
protein data bank, exclude all terminal residues and residues in
unusual conformations, and then find for each amino acid the
maximum solvent accessibility within that data set, according to
one’s chosen definition of solvent accessibility. As Table 1 shows,
this empirical approach should generally yield results that are quite
similar to the theoretical normalization values.

In many applications, specifically in the context of sequence
evolution, RSA is treated as a site-specific property that is
invariant under mutation. While RSA values of homologous
structures tend to be strongly correlated [9,14], individual sites,
in particular exposed ones, can show substantial RSA variability
[14]. In this context, we would like to emphasize that one potential
source of RSA variability in previous studies was RSA normal-
ization. For example, Ref. [14] used the Rose scale, which differs
quite substantially from the scale we propose here. In particular,
the corrections we propose to the Rose scale range from 4% (for
Leu) to 18% (for Asp), and are approximately uniformly
distributed in that range over the 20 amino acids. Thus, one can envision scenarios under which a substitution that might not
change RSA under our scale might change it by over 10% under
the Rose scale. At the same time, we have to realize that RSA can
show variability even in the absence of mutation, in particular for
exposed residues. A residue in a surface loop will undergo
thermodynamic fluctuations, and its solvent exposure state will
vary over time as neighboring residues move closer in or further
out. By contrast, a residue in the core will likely remain solvent-
occluded at all times. To obtain a reliable RSA value for a surface
residue, one would thus ideally calculate an average over a
thermodynamic ensemble of structures. A detailed analysis of RSA
variability under thermodynamic fluctuations and among homol-
ogous structures is beyond the scope of this work but should be
undertaken in the future.

The comparison between experimentally and empirically derived
hydrophobicity scales has been a persistent topic in biochemistry. As
of this writing, the AAindex database [36] contains over 40 scales
related to amino acid hydrophobicity or polarity. While these scales
tend to cluster [37,38], there are substantial dissimilarities among
hydrophobicity scales, and any two scales within the hydrophobicity
cluster may not correlate that well. Any further insight into the
mechanisms that cause differences among scales derived under
different conditions or using different methodologies would improve
our understanding of protein biochemistry. In particular, resolving discrepancies between empirically-derived data and experimentally derived thermodynamics of hydrophobicity could provide crucial insight into algorithms of protein-structure prediction and de-novo protein folding.

Wollfenden et al. [26] were the first to propose an approach for reconciling the empirical and the experimental approach, by correlating the distribution of amino acid exposure with their experimental behaviors in water/vapor solutions. More recently, Moebert et al. [4] attempted to reconcile these disparities by correlating hydrophobic states with surface-exposure patterns of protein structures. Additionally, Shaytan et al. [5] assessed the distribution of amino acid exposure in proteins to discern apparent free energies of transfer between protein interior and surface states, and found that free energy is highly correlated with experimental hydrophobicity scales [5]. Each of these approaches used the ASA normalization values from either Rose et al. [2] or Miller et al. [3]. Since the normalization ASA values developed here are more accurate, we believe that our findings are valuable for determining exposure states. Using the Rose hydrophobicity scale as an example, we have shown here that improved ASA normalization values consistently yield improved correlations with experimental scales, irrespective of the exact type of experimental scale considered. Of all empirical scales we analyzed, however, the fraction of 95% buried residues was most consistently strongly correlated with different experimental scales and thus could be considered the overall best-performing empirical scale.

Further, in agreement with Shaytan et al. [5], we found that different experimental scales corresponded to different empirical scales. For example, transfer energies from water to vapor correlated the strongest with the fraction of 100% buried residues, while transfer energies from water to cyclohexane correlated the strongest with the fraction of 95% buried residues, and transfer energies from water to octanol correlated the strongest with mean RSA. Since mean RSA puts more weight on exposed residues than does the fraction of either 100% buried or 95% buried residues, this finding agrees with the three distinct types of scales found by Shaytan et al. [5]. The pentapeptide scale by Wimley et al. [32], however, did not correlate well with either of the empirical scales we considered. Wimley et al. performed a partitioning experiment between water and 1-octanol using pentapeptide species, Ace-WLXLL, with X being one of the naturally occurring 20 amino acids. Otherwise, their set up was similar to the one of Fauchere et al. [28], with X being one of the naturally occurring 20 amino acids. By using pentapeptides rather than individual amino acids, the Wimley et al. hydrophobicity scale does not seem to accurately reflect the hydrophobic character of individual amino acids but rather that of the pentapeptides.

In summary, we have presented significantly improved ASA normalization values. We recommend that our theoretical normalization values for the ALLOWED region (column 1 of Table 1) be used to normalize ASA. The optimal hydrophobicity scale will depend on the specific application, but the fraction of 95% buried residues seems to be the best general-purpose empirical scale.

Materials and Methods

Empirical maximum ASA values

We obtained a set of 3197 high-quality protein crystal structures using the PISCES server [24]. We imposed the following requirements: resolution of 1.8 A or less, an R-free value <0.25, and a pairwise mutual sequence identity of at most 20%. For each amino-acid residue in all 3197 structures, we retrieved bond lengths, bond angles, dihedral angles, peptide bond lengths, and nearest neighbors. Chain-terminating residues, defined as those residues whose peptide bond lengths with any neighboring residue was greater than six standard deviations from the protein’s mean peptide bond length, were excluded from all subsequent analyses. We further identified all residues in the data set that had either missing atoms or atoms with ambiguous occupancy data (PDB occupancy column contained a number <1.0 for at least one atom in the residue). We eliminated these residues and their immediate neighbors from all subsequent analyses as well.

We used the program DSSP (2011 version) [34] to calculate solvent accessibility (ASA) and to identify the secondary structure of each residue across all proteins. Because of the quality control we imposed on residues (see preceding paragraph), our final ASA data set only contained residues that were complete and unambiguous and whose neighbors were complete and unambiguous as well.

We next filtered by allowed Ramachandran angles. For each amino acid, we binned all observed θ,ψ combinations into 5° × 5° squares, and assigned each square to one or more of the following regions: The CORE region was defined to contain at least 80% of the observed Ramachandran angles. The ALLOWED region was defined to contain at least 97% of the observed Ramachandran angles. For both the CORE and the ALLOWED regions, we identified, for each amino acid, the number of observations per 5° × 5° bin required for that bin to be part of the respective region. Table S4 in File S1 lists these bin cutoffs. The GENEROUS region was defined to extend the ALLOWED region by 20° in all directions, regardless of whether the particular Ramachandran angles have been observed. Finally, the ALL region was defined to contain all observed Ramachandran angles. The definitions of the CORE, ALLOWED, and GENEROUS regions are consistent with current biochemical convention [25,39]. For all four regions, we identified the maximum ASA observed.

We calculated RSA as RSA = ASA/MaximumASA, where “Maximum ASA” corresponds to the maximum ASA value, as determined by the normalization scale used, for the focal amino acid.

Theoretical maximum ASA values

To find the theoretical maximum solvent accessibility (ASA) for each amino acid X, we computationally constructed Gly-X-Gly tripeptides. Each tripeptide was modeled by specifying coordinates of each constituent atom, using bond lengths and angles from our empirically mined protein structures. Briefly, we first constructed peptides in a defined conformation by placing each atom at the correct position in 3D space. We then adjusted θ, ψ, and χ angles to obtain the desired conformation. This method is described in more detail in Text (File S1), and the computer code to carry out tripeptide construction has been published as a stand-alone library [40].

Once constructed, we exhaustively rotated θ and ψ dihedral backbone angles in discrete 1° increments, holding χ constant at 180°. For each (θ,ψ) combination, we additionally rotated through all possible χ rotamer angles, as found in the Dunbruck Rotamer Database [24]. Rotamer angles were grouped into three 120° sectors (60°–60°, 60°–180°) and averaged within each sector. For amino acids where the side chain could assume more than ten distinct rotamer conformations (e.g. for L, I, M, K, N), we selected ten rotamer conformations at random instead of exhaustively enumerating all rotamer conformations. A different set of randomly chosen rotamer conformations was generated for each combination of (θ,ψ) angles.

For each tripeptide conformation examined, a corresponding PDB file was created and inputted into the program DSSP [34] to
compute the ASA of amino acid X. For each amino acid and \((\phi, \psi)\) combination, we recorded the largest ASA value from all rotamer variations examined. To determine the theoretical maximum ASA value for each amino acid, we identified the largest ASA value observed for any \((\phi, \psi)\) combination within one of the four Ramachandran regions defined above (CORE, ALLOWED, GENEROUS, ALL).

**Hydrophobicity scales**

We calculated empirical hydrophobicity scales on the same set of 3197 crystal structures. Mean RSA of each amino acid was calculated as the RSA averaged over all occurrences of that amino acid in the data set. The corresponding hydrophobicity scale was defined as \(1 - q meanRSA\), where RSA was calculated using the theoretical normalization values of Table 1 (ALLOWED region).

\[
\text{ASA} = 1 - q meanRSA
\]

**References**

1. Chothia C (1976) The nature of the accessible and buried surfaces in proteins. J Mol Biol 1: 1–14.
2. Rose GD, Geselowitz AR, Lesser GJ, Lee RH, Zehfus MH (1983) Hydrophobicity of amino acid residues in globular proteins. Science 229: 834–838.
3. Miller S, Janin J, Lesk AM, Chothia C (1987) Interior and surface of monomeric proteins. J Mol Biol 196: 641–656.
4. Moellbert S, Emberly E, Tang C (2004) Correlation between sequence hydrophobicity and surface-exposure pattern of database proteins. Prot Sci: 732–742.
5. Shatyan AK, Shaitan KV, Khalidov AR (2009) Solvent accessible surface area of amino acid residues in globular proteins: Correlations of apparent transfer free energies with experimental hydrophobicity scales. Biomacromolecules 10: 1224–1237.
6. Goldman N, Thorne JL, Jones DT (1999) Assessing the impact of secondary structure and solvent accessibility on protein evolution. Genetics 149: 445–450.
7. Bloom JD, Drummond DA, Arnold FH, Wilke CO (2006) Structural determinants of the rate of protein evolution in yeast. Mol Biol Evol 23: 1751–1761.
8. Franzosa EA, Xia Y (2009) Structural determinants of protein evolution are context-sensitive at the residue level. Mol Biol Evol 26: 2397–2398.
9. Zhou T, Weems M, Wilke CO (2009) Translationally optimal codons associate with structurally sensitive sites in proteins. Mol Biol Evol 26: 1571–1580.
10. Franzosa EA, Xia Y (2012) Independent effects of protein core size and expression on structure-evolution relationships at the residue level. PLoS One 7: e46692.
11. Scherrr MP, Meyer AG, Wilke CO (2013) Modeling coding-sequence evolution within the context of residue solvent accessibility. BMC Evol Biol 12: 179.
12. Meyer AG, Wilke CO (2012) Integrating sequence variation and protein structure to identify sites under selection. Mol Biol Evol.
13. Contant GC, Stadler PF (2009) Solvent exposure imparts selective pressures across a range of yeast proteins. Mol Biol Evol 26: 1155–1161.
14. Rost B, Sander C (1993) Conservation and prediction of solvent accessibility in protein families. Proteins 20: 216–226.
15. Pollastri G, Baldi P, Fariselli P, Casadio R (2002) Prediction of coordination Function, and Bioinformatics 56: 753–767.
16. Adamczak R, Porollo A, Miller J (2004) Accurate prediction of solvent accessibility using neural networks-based regression. Proteins: Structure, Function, and Bioinformatics 56: 753–767.
17. Nguyen MN, Rajapakse JC (2003) Prediction of protein relative solvent accessibility with a two-stage SVM approach. Proteins 59: 30–37.
18. Petersen T, Nordløkken R, Hansen JO, Nielsen M, Lundegaard C (2009) A generic method for assignment of reliability scores applied to solvent accessibility predictions. BMC Struct Biol 9: 51.
19. Singh H, Ahmad S (2009) Context dependent reference states of solvent accessibility derived from native protein structures and assessed by predictability analysis. BMC Struct Biol 9: 25.
20. Levy ED (2010) A simple definition of structural region in proteins and its use in analyzing interface evolution. J Mol Biol 403: 660–670.
21. Cheng J, Svendsen MJ, Baldis P (2006) DOMpro: protein domain prediction using profiles, secondary structure, relative solvent accessibility, and recursive neural networks. Data Mining and Knowledge Discovery 13: 1–10.
22. Ahmed S, Xia Y (2009) Context-dependent reference states of solvent accessibility derived from native protein structures and assessed by predictability analysis. BMC Struct Biol 9: 25.
23. Levy ED (2010) A simple definition of structural region in proteins and its use in analyzing interface evolution. J Mol Biol 403: 660–670.
24. Wang G, Dunbrack RL (2003) PISCES: a protein sequence culling server. Bioinformatics 19: 1509–1519.
25. Morris AL, MacArthur MW, Hutchinson EG, Thornton JM (1992) Stereochemical quality of protein structure coordinates. Proteins 12: 345–364.
26. Wolfenden R, Anderson L, Cullis P, Southgate CB (1981) Affinities of amino acid side chains for solvent water. Biochemistry: 849–853.
27. Kyte J, Doolittle RF (1982) A simple method for displaying the hydrophatic character of a protein. J Mol Biol 157: 105–132.
28. Fauchere JL, Pliska VE (1983) Hydrophobic parameters of amino-acid side chains from the partitioning of N-acetyl-amino-acid amides. Eur J Mol Chem 18: 369–373.
29. Radzicka A, Wolfenden R (1988) Comparing the polarities of the amino acids: side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution. Biochem 27: 1664–1670.
30. MacCallum JL, Bennett WD, Tieleman DP (2007) Partitioning of amino acid side chains into lipid bilayers: results from computer simulations and comparison to experiment. J Gen Physiol 129: 371–377.
31. Moon CP, Fleming KG (2011) Side chain hydrophobicity scales derived from transmembrane protein folding into lipid bilayers. Proc Natl Acad Sci USA 1081: 10174–10177.
32. Winley WC, Creamer TP, White SH (1996) Solvation energies of amino acid side chains and backbone in a family of host-guest pentapeptides. Biochem 35: 5109–5124.
33. Ahmad S, Gromiha MM, Sarai A (2003) Real-value prediction of solvent accessibility from amino acid sequence. Proteins 50: 629–633.
34. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22: 2577–2637.
35. Lee B, Richards FM (1971) The interpretation of protein structures: estimation of static accessibility. J Mol Biol 55: 379–400.
36. Kawashima S, Kanehisa M (2000) Amino acid index database. Nucleic Acids Res 28: 374.
37. Tomii K, Kanehisa M (1996) Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. Protein Eng 9: 27–36.
38. Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, et al. (2000) Amino acid index database, progress report 2000. Nucleic Acids Res 28: 275–278.
39. Laskowski RA, MacArthur MW, Moss DS, Thornton JM (1993) PROCHECK – a program to check the stereochemical quality of protein structures. J App Cryst 26: 283–291.
40. Tien MZ, Sydykova DK, Meyer AG, Wilke CO (2013) PeptideBuilder: a simple Python library to generate model peptides. PeerJ 1: e80.

**Data and code availability**

All results and all computer code used to generate these results have been deposited to GitHub (https://github.com/miten/RSA-normalization-values). This includes maximum observed ASA values (both empirical and theoretical) as a function of backbone dihedral angles.

**Supporting Information**

File S1 Combined pdf of Figures S1–S4, Tables S1–S4, and Supporting Text. (PDF)

**Acknowledgments**

We thank Jeff Gray for insightful discussions on this work.

**Author Contributions**

Conceived and designed the experiments: MZT AGM COW. Performed the experiments: MZT. Analyzed the data: MZT DKS COW. Contributed reagents/materials/analysis tools: MZT AGM DKS. Wrote the paper: MZT AGM DKS SJS COW.