Modeling Unfolded States of Proteins and Peptides. II. Backbone Solvent Accessibility

Trevor P. Creamer,‡ Rajgopal Srinivasan, and George D. Rose*

Department of Biophysics and Biophysical Chemistry, The Johns Hopkins University School of Medicine, 725 N. Wolfe St., Baltimore, Maryland 21205

Received November 14, 1996; Revised Manuscript Received January 2, 1997[⊗]

ABSTRACT: Buried surface area is often used as a measure of the contribution to protein folding from the hydrophobic effect. Quantitatively, the surface buried upon folding is reckoned as the difference in area between the native and unfolded states. This calculation is well defined for a known structure but model-dependent for the unfolded state. In a previous paper [Creamer, T. P., Srinivasan, R., & Rose, G. D. (1995) *Biochemistry 34*, 16245−16250], we developed two models that bracket the surface area of the unfolded state between limiting extremes. Using these extrema, it was shown that earlier models, such as an extended tripeptide, overestimate the surface area of side chains in the unfolded state. In this sequel to our previous paper, we focus on backbone surface in the unfolded state, again adopting the strategy of trapping the area between limiting extrema. A principal conclusion of this present study is that most backbone surface in proteins is buried within local structure.

Protein folding is defined as the reversible transition from a disordered ensemble, called the unfolded state, to a uniquely folded structure (Anfinsen, 1973). Unfolded states have resisted precise characterization, with most folding studies focusing exclusively on the native structure. However, to understand the thermodynamics of folding, both sides of the folding transition must be considered (Dill and Shortle, 1991).

Surface area is a quantity of particular interest in folding studies. The loss of accessible surface upon folding has been employed to good effect in numerous studies of hydration thermodynamics (Eisenberg & McLachlan, 1986; Ooi *et al.*, 1987; Wesson & Eisenberg, 1992; Murphy & Freire, 1992; Privalov & Makhatadze, 1992; and Spolar *et al.*, 1992). Numerical algorithms to calculate the solvent accessible surface area (ASA) were developed by Lee and Richards (1971) and Shrake and Rupley (1973), and analytic procedures soon followed (Connolly, 1983; Richmond, 1984).

The surface area lost upon folding is taken as the difference in ASA between native and unfolded states. The calculation is straightforward for a defined native structure but problematic for the unfolded state. Recently, we described two models of unfolded peptides and proteins that bracket side chain ASA between limiting extremes (Creamer *et al.*, 1995). A bound on maximal ASA was obtained using Monte Carlo computer simulations of peptides modeled with a hard sphere potential. A bound on minimal ASA was obtained by excising fragments from proteins of known structure. From these models, it was argued that most earlier procedures result in a substantial overestimate of the ASA lost upon folding.

Our previous work (Creamer *et al.*, 1995) focused on side chain accessibility; we turn now to backbone accessibility. In the following, the backbone accessibility of each residue type in unfolded proteins and peptides is calculated by trapping its area between limiting extrema. As previously (Creamer *et al.*, 1995), the upper bound overestimates exposure, while the lower bound overestimates burial. Using the lower bound in particular, it is shown that most backbone surface is buried within local structure.

Accessible surface area can be utilized to clarify a controversy involving the energy hydrogen bonds contribute to protein stability. Theoretical work (Honig & Yang, 1995) indicates that protein hydrogen bonds may be destabilizing. In sharp contrast, interpretations of experimental data indicate that hydrogen bonds stabilize proteins by approximately 2 kcal mol^{-1} (Myers & Pace, 1996). Myers and Pace suggest one reason for this discrepancy may be that polar groups, both side chain and backbone, are less accessible to solvent in the unfolded state than taken into account by theory. In agreement with this suggestion, Creamer *et al.* (1995) found that earlier models [*e.g.*, Lesser and Rose (1990)] overestimate side chain exposure in the unfolded state. We now report a similar finding for the peptide backbone.

In this paper, we depart from our earlier, computer-intensive model for ASA at the upper bound (Creamer *et al.*, 1995) and adopt instead the straightforward approach of Spolar *et al.* (1992), which yields similar values. Using this new upper bound and our earlier lower bound, accessibility data are presented for both backbones and side chains of all 20 residues.

METHODS

Lower bounds were obtained from chain segments of length $N = 3, 5, 7, ...,$ and 45 excised in their native conformation from a dataset of 43 protein chains (Creamer *et al.*, 1995). These highly self-associated segments are expected to be more compact, and therefore less solvent exposed, than unfolded peptides. The accessibility of the

* Author to whom correspondence should be addressed. Phone: (410) 614-3970. Facsimile: (410) 614-3971. Email: rose@grserv.med.jhmi.edu.

‡ Current address: Center for Structural Biology, Department of Biochemistry, University of Kentucky, 800 Rose Street, Lexington, KY 40536.

⊗ Abstract published in *Advance ACS Abstracts,* February 15, 1997.

[1] Abbreviation: ASA, solvent accessible surface area.

S0006-2960(96)02819-X CCC: $14.00

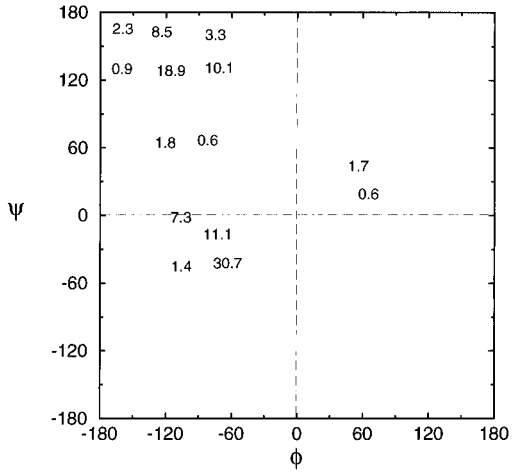

FIGURE 1: Distribution of backbone dihedral (ϕ, ψ) angles for all nonprolyl, nonglycyl residues in the dataset of 43 protein chains. Each number indicates the percentage of residues found in the region immediately surrounding the digits.

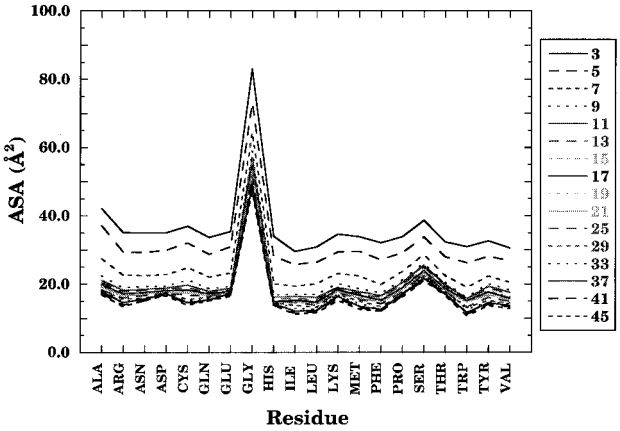

FIGURE 2: Plot of mean residue backbone ASA from the lower bound as a function of fragment length. Backbone accessible surface areas (Å^2) were calculated for the central residue in protein fragments of length $N = 3$, 5, 7, ..., 21, 25, 29, ..., and 45, excised from the dataset of 43 protein chains. The raw areas were then averaged over all occurrences by residue type.

central residue was determined in each segment and averaged by residue type over every occurrence. ASA was calculated using the method of Richmond (1984) with Richards' radii (1977), united atoms, and a solvent probe radius of 1.4Å.

Upper bounds (Spolar *et al.*, 1992) were obtained from the same dataset of 43 protein chains. In each, the protein was modeled in an extended conformation, with backbone dihedrals set to $\phi = -120°$ and $\psi = +120°$ and side chain torsions set to 180°. The choice of backbone dihedrals for the extended state was chosen to reflect the distribution of residues in proteins of known structure. Figure 1 is a Ramachandran plot of this distribution for all nonprolyl, nonglycyl residues in the database of 43 protein chains used in this study [see also Karplus (1996)]. A maximum in the extended region is evident near $\phi = -120°$ and $\psi = +120°$. This upper bound model is more solvent exposed than expected for an unfolded peptide. Using the model, solvent accessible surface areas were calculated for each residue in every chain and averaged over all occurrences by residue type. To avoid end effects, values of the first and last three residues of each chain were excluded from these averages.

The 43 well-refined, high-resolution (*R*-factor of 20% or better, resolution of 2.0 Å or better) protein chains taken from the Brookhaven protein data bank (PDB) (Bernstein *et al.*, 1977) are 1bp2, 1crn, 1ecd, 1gcr, 1gd1(O), 1gp1(A), 1hmq(A), 1hoe, 1lz1, 1mbo, 1ppt, 1rdg, 1sn3, 1snc, 1tpp, 1ubq, 2act, 2aza(B), 2ca2, 2cdv, 2cts, 2lhb, 2ovo, 2pcy, 2rhe, 2wrp(R), 351c, 3app, 3grs, 3ins(C), 3ins(D), 3lzm, 3rnt, 3tln, 4dfr(B), 4fxn, 4pep, 5cha(A), 5cpa, 5cyt(R), 5pti, 7rsa, and 9pap. In multimeric structures, the subunits used in the analyses are shown in parentheses.

RESULTS

Backbone ASA. Figure 2 shows the average backbone ASA by residue type in protein fragments, for fragment lengths of $N = 3$, 5, 7, ..., 21, 25, 29, ..., and 45. The backbone is defined as the peptide unit including the α-carbon (N−C_α−C═O). Values of backbone ASA converge rapidly, approaching a plateau at $N = 11$. To remain consistent with previous work (Creamer *et al.*, 1995), $N = 17$ was chosen to represent backbone ASA at the lower bound; the choice is insensitive beyond $N = 11$. Table 1 lists the average backbone accessibility of excised fragments of length $N = 17$ (lower bound) and fully extended chains (upper bound).

Table 1. Backbone and Side Chain Average ASA Values for Both Lower and Upper Bound Models[a]

	average			
	backbone ASA		side chain ASA	
residue	lower	upper	lower	upper
Ala	19.8	35.9	46.6	63.6
Arg	17.1	33.0	156.9	185.3
Asn	17.6	32.7	84.5	95.6
Asp	18.1	33.9	79.2	94.8
Cys	18.2	34.5	62.9	83.0
Gln	17.2	33.4	105.0	128.7
Glu	17.9	33.5	102.8	123.9
Gly	54.6	75.7	0.0	0.0
His	14.9	33.4	103.9	119.1
Ile	15.2	24.7	100.1	134.1
Leu	14.7	30.7	101.4	117.7
Lys	18.3	33.8	142.5	158.8
Met	16.7	33.8	105.3	139.5
Phe	15.3	33.3	118.7	139.8
Pro	18.9	26.1	83.5	90.5
Ser	23.8	35.0	59.7	73.3
Thr	18.6	29.5	77.3	91.2
Trp	15.1	32.0	154.7	158.4
Tyr	17.7	33.5	131.0	152.3
Val	15.9	24.9	81.8	110.9

[a] Lower bound values are from fragments of length $N = 17$. Upper bound values are from the extended chain.

In Figure 3, the mean backbone ASA of excised fragments is plotted against fragment length for Ala, Ser, Cys, Val, Phe, and Lys. The rapid decrease in ASA from $N = 3$ to about $N = 11$ is conspicuous, with little additional change observed between $N = 11$ and $N = 45$. All residues display similar behavior, including the 14 not shown. It is apparent that backbone surface is buried predominantly within local structure because, by definition, there is none other in excised fragments at $N = 11$. As seen, the mean backbone ASA is small though nonzero for all residues; any further burial would arise from longer range interactions ($N > 45$).

Side Chain ASA. Table 1 also lists mean side chain accessibility. Lower bound values were taken from excised fragments of length $N = 17$ (Creamer *et al.*, 1995), and upper

Table 2: Lower Bound, Upper Bound, and Tripeptide Models of ASA in Unfolded Proteins

protein	PDB code[a]	number of residues	ASA estimate of folded protein (Å^2)	ASA estimate for unfolded protein (Å^2)[b]		
				lower boundary	upper boundary	tripeptide
Barnase	1rnb	110	5959	12 122	15 535	19 112
Staphylococcal nuclease	2sns	149	8020	16 910	21 890	26 451
T4 phage lysozyme	2lzm	164	8572	18 447	23 782	29 155
Triosephosphate isomerase	3tim (A)	250	11306	26 508	34 646	42 214
Subtilisin	2sbt	275	10320	26 631	35 590	42 794

[a] PDB four-character identifier. Chain A of 3tim used in analysis. [b] Upper and lower bounds computed from data in Table 1. Tripeptide sums computed from data in Lesser and Rose (1990).

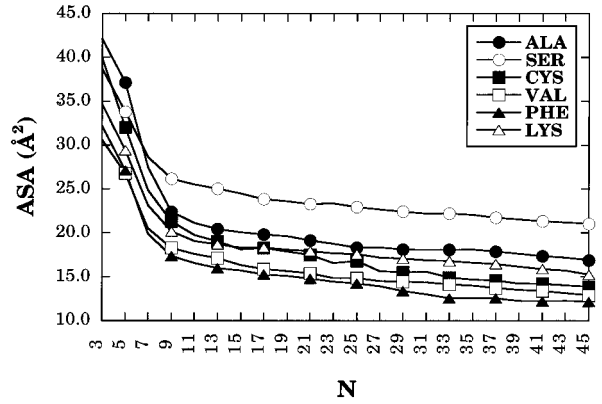

FIGURE 3: Backbone accessible surface areas (Å^2) for Ala, Ser, Cys, Val, Phe, and Lys plotted against fragment length (N = 3, 5, 7, ..., and 45).

bound values were averaged over all protein chains in the dataset, modeled in extended conformations. For most residues, the two boundaries yield significantly different values. Exceptions include Trp, Asn, and, to a lesser extent, Pro.

Values at the upper bound are similar in magnitude to those obtained previously using simulated peptides (Creamer *et al.*, 1995). For example, side chains of Ala and Val in the extended chain model are 64 and 111 Å^2, respectively. In comparison, Ala and Val side chains have values of 61 and 112 Å^2, respectively, in simulated peptides of length N = 15.

In this study, the maximal ASA of unfolded peptides is substantially less than that reported in most earlier work. For example, Lesser and Rose (1990), who used a tripeptide standard state, obtained side chain accessibilities that are 24% larger, on average, than the upper bounds (Table 1).

Protein ASA. For an arbitrary protein sequence, ASA in the unfolded state is calculated by summing over the appropriate values from Table 1. Table 2 lists these sums at both upper and lower bounds for 5 protein chains, together with the ASA of the folded molecule. For comparison, Table 2 also includes sums reckoned from the tripeptide standard state of Lesser and Rose (1990).

These data can be used to estimate the average increase in ASA upon unfolding. Compared to the folded protein, the ASA of the unfolded ensemble is larger by a factor of approximately $2^1/_4$ at the lower bound, 3 at the upper bound, and $3^1/_2$ in the tripeptide model.

DISCUSSION

The most notable finding of this study is that burial of the peptide backbone is primarily a consequence of forming local structure. As seen in Figures 2 and 3, little additional backbone area loss occurs beyond fragment lengths of N = 11. This finding is consistent with earlier conclusions of Chothia (1975) and with Stickle *et al.* (1992), who observed that most backbone−backbone hydrogen bonds are within local structure (*e.g.*, helices and turns), with concomitant loss of accessible surface. It should also be noted that the intramolecular interactions which give rise to local structure are not necessarily local.

On average, the accessibility of nonglycyl, nonprolyl backbone ranges between 17 and 32 Å^2 in the unfolded state, less than the ASA of a side chain methyl group (see Ala in Table 1). In comparision, the average backbone accessibility is 39 Å^2 from stochastic tripeptides (Lesser & Rose, 1990). Consequently, any analysis based on a tripeptide model will overestimate the degree to which backbone−solvent hydrogen bonds stabilize the unfolded state. While our data cannot provide a definitive resolution to the controversy about hydrogen bond stability (Myers & Pace, 1996), they do raise a caution that theoretical analyses often exaggerate backbone ASA in the unfolded state.

Figure 2 contains useful information about the backbone microenvironment. For example, the backbone of Cys is comparatively exposed at shorter fragment lengths, consistent with having a side chain that is isosteric to the Ser side chain. However, at longer fragment lengths, the Cys backbone profile changes from exposed to buried ($N \geq 25$ in Figures 2 and 3) as disulfide bonds, which are well buried (Behe *et al.*, 1991), increase in likelihood (Harrison & Sternberg, 1994).

Side chain ASAs presented here are considerably smaller than corresponding tripeptide values. Still, side chains retain substantial exposed surface (Table 1), even at the lower extreme, which is derived from collapsed, highly self-associated chain fragments. Burial of apolar atoms from this exposed surface would still contribute significantly to the energetics of protein stability, as proposed by Kauzmann (1959). Indeed, in our analysis, the area buried upon folding (74−82%) is contributed largely by side chains.

The two models presented here establish limits on the accessible surface area of unfolded peptides and proteins, and Table 1 can be used to estimate these limiting values. This procedure has been illustrated for five proteins in Table 2. It can be concluded from these estimates that unfolded proteins average between two and three times the area of their folded counterparts. As pointed out by Richards (1977), only a modest increase in volume is needed for solvent penetration sufficient to effect this large increase in area. This conceptual picture is consistent with recent experimental evidence that even highly denatured proteins remain compact (Calmettes *et al.*, 1994; Lattman, 1994).

UNFOLDED PROTEIN ASA ESTIMATES ON THE WWW

Using data from Table 1, the ASA of an arbitrary unfolded protein chain can be obtained by accessing the WWW page at http://cherubino.med.jhmi.edu/~folded. ASA values of a user-provided sequence are computed and returned in tabular format.

ACKNOWLEDGMENT

We thank Rajeev Aurora and C. Nick Pace for useful suggestions, Andy Karplus for sending a manucript prior to publication, and the NIH for support.

REFERENCES

Anfinsen, C. B. (1973) *Science 181*, 223−230.
Behe, M. J., Lattman, E. E., & Rose, G. D. (1991) *Proc. Natl. Acad. Sci. U.S.A. 88*, 4195−4199.
Bernstein, F. C., Koetzle, T. G., Williams, G. J. B., Meyer, E. F., Jr., Brice, M. D., Rogers, J. R., Kennard, O., Shimanouchi, T., & Tasumi, M. (1977) *J. Mol. Biol. 112*, 535−542.
Calmettes, P., Durand, D., Desmadril, M., Minard, P., Receveur, V., & Smith, J.C. (1994) *Biophys. Chem. 53*, 105−114.
Chothia, C. (1975) *Nature (London) 254*, 304−308.
Connolly, M. L. (1983) *Science 221*, 709−713.
Creamer, T. P., Srinivasan, R., & Rose, G. D. (1995) *Biochemistry 34*, 16245−16250.
Dill, K. A., & Shortle, D. (1991) *Annu. Rev. Biochem. 60*, 795−825.
Eisenberg, D., & McLachlan, A. D. (1986) *Nature (London) 319*, 199−203.
Harrison, P. M., & Sternberg, M. J. E. (1994) *J. Mol. Biol. 244*, 448−463.
Honig, B., & Yang, A.-S. (1995) *Adv. Prot. Chem. 46*, 27−58.
Karplus, P. A. (1996) *Protein Sci. 5*, 1406−1420.
Kauzmann, W. (1959) *Adv. Prot. Chem. 14*, 1−64.
Lattman, E. E. (1994) *Curr. Opin. Struct. Biol. 4*, 87−92.
Lee, B., & Richards, F. M. (1971) *J. Mol. Biol. 55*, 379−400.
Lesser, G. J., & Rose, G. D. (1990) *Proteins 8*, 6−13.
Murphy, K. P., & Freire, E. (1992) *Adv. Protein Chem. 43*, 313−361.
Myers, J. K., & Pace, C. N. (1996) *Biophys. J. 71*, 2033−2039.
Ooi, T., Oobatake, M., Némethy, G., & Scheraga, H. A. (1987) *Proc. Natl. Acad. Sci., U.S.A. 84*, 3086−3090.
Privalov, P. L., & Makhatadze, G. I. (1992) *J. Mol. Biol. 224*, 715−723.
Richards, F. M. (1977) *Annu. Rev. Biophys. Bioeng. 6*, 151−176.
Richmond, T. J. (1984) *J. Mol. Biol. 178*, 63−89.
Shrake, A., & Rupley, J. A. (1973) *J. Mol. Biol. 79*, 351−371.
Spolar, R. S., Livingstone, J. R., & Record, M. T., Jr. (1992) *Biochemistry 31*, 3947−3955.
Stickle, D. F., Presta, L. G., Dill, K. A., & Rose, G. D. (1992) *J. Mol. Biol. 226*, 1143−1159.
Wesson, L., & Eisenberg, D. (1992) *Protein. Sci. 1*, 227−235.

BI962819O