Sequence-Specific Solvent Accessibilities of Protein Residues in Unfolded Protein Ensembles

Pau Bernadó,*† Martin Blackledge,* and Javier Sancho‡

*Institut de Biologie Structurale Jean-Pierre Ebel CNRS-CEA.-UJF, Grenoble, France; †Institut de Recerca Biomèdica, Parc Científic de Barcelona, Barcelona, Spain; and ‡Departamento de Bioquímica y Biología Molecular y Celular, Facultad de Ciencias, and Biocomputation and Complex Systems Physics Institute-BIFI, University of Zaragoza, Zaragoza, Spain

ABSTRACT Protein stability cannot be understood without the correct description of the unfolded state. We present here an efficient method for accurate calculation of atomic solvent exposures for denatured protein ensembles. The method used to generate the ensembles has been shown to reproduce diverse biophysical experimental data corresponding to natively and chemically unfolded proteins. Using a data set of 19 nonhomologous proteins containing from 98 to 579 residues, we report average accessibilities for all residue types. These averaged accessibilities are considerably lower than those previously reported for tripeptides and close to the lower limit reported by Creamer and co-workers. Of importance, we observe remarkable sequence dependence for the exposure to solvent of all residue types, which indicates that average residue solvent exposures can be inappropriate to interpret mutational studies. In addition, we observe smaller influences of both protein size and protein amino acid composition in the averaged residue solvent exposures for individual proteins. Calculating residue-specific solvent accessibilities within the context of real sequences is thus necessary and feasible. The approach presented here may allow a more precise parameterization of protein energetics as a function of polar- and apolar-area burial and opens new ways to investigate the energetics of the unfolded state of proteins.

INTRODUCTION

The unfolded state of proteins is as important as the native state in determining protein stability and the mechanism of the protein folding reaction (1–2). However, most folding studies are focused on the native state due to the increased facility of studying folded proteins.

The thermodynamic stability of proteins is dependent on a delicate balance between different interactions involving protein and solvent atoms in both the native and unfolded states (2–3). However, there is still no general agreement on the net contribution of the different fundamental interactions, such as hydrogen bonds, van der Waals, etc. (3–7). On the other hand, the hydrophobic effect, which describes the observed tendency of apolar compounds to minimize their exposure to water, is widely acknowledged as stabilizing the native state (8–11). The experimental quantification of the contribution of the hydrophobic effect to protein stability is not trivial, one reason being that, unlike in the native state, the side-chain solvent exposures in the unfolded state are hard to estimate due to the large number of discrete conformations available to the main chain. So far, side-chain accessibilities have been approximated by calculations performed on different models of the unfolded state, including Gly-X-Gly extended tripeptides (12), Gly-X-Gly peptides with dihedral angles characteristic of protein structures (13,14), and Ala-X-Ala simulated ensembles (15), or, more recently, by averages of peptide-fragment collections extracted from native structures (16,17). Large differences in solvent exposures have been reported depending on the model. Until now, reported exposures have tended to be residue-type averages, rather than residue-specific exposures calculated in specific sequences. The solvent accessibility in the unfolded state is intimately linked to the conformational sampling occurring on the backbone, and it has been shown that neighboring residues limit the conformational space sampled by certain amino acids (18,19). In addition, large amino acids may bury the chain from the solvent more effectively. Therefore, it is clear that both the interpretation of stability mutational studies (20) and the parameterization of protein stability calculations (21) could benefit from using sequence-specific solvent exposure data calculated from accurate models of unfolded protein ensembles.

The consensus view of the unfolded state is that of a large ensemble of more or less randomized conformations in fast equilibrium, although certain bias toward the native conformation (22,23) or toward certain types of secondary structure, especially the polyproline II, has been reported in some cases (24,25). At present, diverse structural techniques can provide valuable structural information about highly disordered proteins. Nuclear magnetic resonance, by measuring residual dipolar couplings in partially aligned proteins (26), has provided insight into the conformational sampling observed in intrinsically and chemically unfolded proteins (22,23,27–32). Paramagnetic relaxation enhancement experiments measured in spin-labeled mutants of several proteins

Submitted April 21, 2006, and accepted for publication August 21, 2006.

Address reprint requests to Pau Bernadó, Institut de Recerca Biomèdica, Parc Científic de Barcelona, Josep Samitier 1-5, 08028 Barcelona, Spain. E-mail: pbernado@pcb.ub.es; or to Javier Sancho, Departamento de Bioquímica y Biología Molecular y Celular, Facultad de Ciencias, Universidad de Zaragoza, 50009-Zaragoza, Spain. E-mail: jsancho@unizar.es.

0006-3495/06/12/4536/08 $2.00

doi: 10.1529/biophysj.106.087528

have provided information about the presence of long-range contacts in unstructured chains (33–36). Small-angle x-ray scattering experiments have become an important tool for the study of size characteristics of the unfolded state (37–39). Recently, we developed an algorithm, Flexible-Meccano, which generates ensembles of realistic atomic models that are compatible with biophysical data measured using NMR and small-angle x-ray scattering (31).

We present here a fast method that provides sequence-specific solvent exposures for any residue in a given unfolded ensemble based on the flexible-Meccano algorithm (31). Large differences in solvent exposures are observed for residue types, depending essentially on the different primary sequence context and, to a lesser extent, on the length of the protein and its global amino acid composition.

METHODS

Generation of the conformational ensemble

The flexible-Meccano algorithm samples efficiently the conformational space representing the unfolded state using a Monte Carlo technique based on residue-specific propensity and specific side-chain volume (31). Consecutive peptide planes and tetrahedral junctions are constructed from the primary sequence starting from the C-terminus. The position of the peptide plane (i) is defined in terms of the C^{α} and C′ atoms of plane ($i + 1$), the selected ϕ/ψ combination and the tetrahedral angle (set to 109°). Amino-acid-specific ϕ/ψ combinations used to create the main chain are randomly extracted from a database built from 500 high-resolution x-ray structures with resolutions of <1.8 Å and B factors <30 Å^2, from which all residues in α-helices and β-sheets were removed (40). Residues preceding a proline are considered additional residue types because of their restricted conformational sampling (41). Moreover, the available conformational space for Gly derived from the database was made symmetric. A residue-specific exclusion volume is also introduced, placing at each C^{β} (or C^{α} for Gly) atom spheres of volumes derived from the Levitt's simplified force field (42). In the case of steric clash with another residue of the chain, the flexible-Meccano algorithm rejects a given ϕ/ψ pair and another set of ϕ/ψ dihedral angles is selected, until no overlap is found or 500 ϕ/ψ combinations have been tested; otherwise, a completely new structure is calculated from the last residue. In a second step (Fig. 1), side chains are incorporated to the ensemble using the program Sccomp, which places and optimizes side-chain conformations in a fixed protein backbone (43). Although Sccomp has been developed and tested for folded proteins, it has been demonstrated to be especially accurate for partially exposed side chains, due to the inclusion of a solvent-accessible term that accounts for the solvation free energy. Additionally, a different version of the flexible-Meccano algorithm, which builds the chain from the N-terminus, has been used to test whether the calculated solvent exposures are influenced by the directionality of chain growth.

Calculation of the solvent exposures

The program Naccess has been used for the calculation of solvent exposures for each individual conformation using a probe of radius 1.4 Å (44). The solvent exposure for each amino acid of each protein was obtained by averaging over the 2000 conformations generated that represent the unfolded ensemble of the protein. This averaging was not applied to the first and last four residues of the chain because their accessibilities essentially reflect their terminal location.

All calculations, for both the generation of the unfolded ensembles and quantification of solvent accessibilities, have been done on the computation center at the Biocomputation and Complex Systems Physics Institute in Zaragoza.

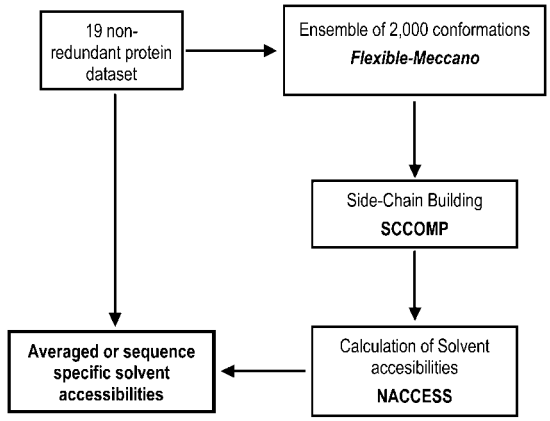

FIGURE 1 Flow-chart of the programs used for generation of denatured-state ensembles (flexible-Meccano and Sccomp) and for the calculation of the atom- and residue-specific solvent exposures (Naccess).

Protein sequences used

A set of 19 proteins corresponding to Set3 from Eyal et al. (43) was used for the calculation of solvent exposures. The PDB codes and residue lengths of the proteins are shown in Table 3. This set was originally collected based on the structural characteristics of the proteins. Of importance for our study, the proteins included in the set share $<20\%$ sequence identity. This implies a variety of amino acid contexts that should provide enough cases to derive reliable solvent-exposure statistics, as well as to reveal sequence-specific solvation characteristics. The total number of amino acid residues in the database was 4346. Notice that cysteine residues are simulated in their reduced form. For each one of these 19 amino acid sequences, an ensemble of 2000 conformations was generated using the flexible-Meccano algorithm.

In parallel, polyalanine chains from 51 to 601 residues, built following the procedure explained above, were simulated to test the effect of protein length on solvent exposure.

RESULTS AND DISCUSSION

Tests for the robustness of the modeling and solvent-exposure calculation protocol

Fig. 2 shows the individual solvent accessibilities calculated in 1000 conformations for two different residues, Lys-65 and Ala-179, of protein 1FCQ. Large variations in solvent exposure were found among the conformations. For the extraction of reliable averaged solvent exposures, the ensembles calculated have to be large enough to guarantee convergence. Two tests were performed to confirm the robustness of the calculated solvent accessibilities with respect to the number of conformations in the ensemble. First, a calculation was performed on 1FCQ, a protein comprising 350 amino acid residues, using 4000 conformations in the ensemble instead of the usual 2000. Residue-specific solvent exposures calculated from the 2000 conformations were equivalent, within 0.5%, to those calculated using 4000 conformations. In a second test, the whole calculation and averaging over all conformations of each of the 19 proteins was repeated from independently generated ensembles. The resulting residue-type averaged solvent exposures from the two calculations were

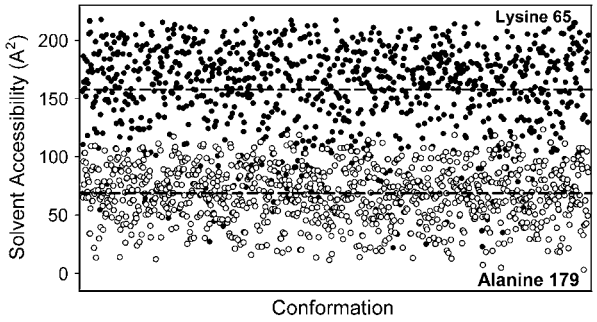

FIGURE 2 Individual solvent exposures of residues Lys-65 and Ala-179 in 1FCQ, calculated in 1000 conformations representing the denatured state of the protein. Dashed lines represent the averaged surface accessibilities of the two residues: 157.57 and 68.72 Å^2, respectively.

equivalent within 0.2%. These results indicate that ensembles of 2000 conformations are large enough to ensure the convergence of the solvent exposures. On the other hand, it was possible that building conformations in one direction could bias the resulting ensemble because in the growing chain, each residue added could have, on average, a significantly smaller volume available to occupy. To test whether there is a directionality effect in the solvent exposures calculated for specific residues in the unfolded ensemble we have developed a new version of the flexible-Meccano algorithm that builds the chain from the N-terminus. The accessibilities calculated for the different residues in a given protein whose unfolded ensemble is built from the C-terminus correlate with those derived from the unfolded ensemble built from the N-terminus with $r > 0.99$ and slopes of 1.0, and there is no systematic deviation in the accessibilities toward either the N- or the C-terminus (see Supplementary Material).

Average residue solvent exposures in denatured ensembles

The calculated solvent exposures of the 20 proteinogenic residues, shown in Table 1, represent the average over all the residues of the same type found in the 19 proteins of the set. Averaged solvent exposures for each residue type were calculated from the individual exposures of at least 51 residues for the rare cysteine, and up to 408 residues for leucine. As expected, average solvent exposures scale with residue size, so that arginines and tryptophans display the largest solvent-exposed areas and glycine the smallest. Compared to earlier reports, the exposures calculated in the denatured ensembles are much lower. Early models of the unfolded state based on

TABLE 1 Solvent accessibilities (Å^2) of amino acid residues in protein unfolded ensembles

	This work					Creamer et al. (17)*		Zielenkiewicz &		
Residue	N¶	Average**	Minimum††	Maximum‡‡	% difference§§	Minimum	Maximum	Saenger (15)†	Miller et al. (12)‡	Rose et al. (14)§
Ala	350	73.1	58.1	83.6	30.5	66.4	99.5	111.6	113	118.1
Arg	234	178.6	154.8	193.7	20.1	174.0	218.3	231.4	241	256.0
Asn	199	109.1	91.6	123.4	25.8	102.1	128.3	151.2	158	165.5
Asp	255	102.0	83.0	117.6	29.4	97.3	128.7	154.7	151	158.7
Cys	51	88.3	76.0	97.7	22.2	81.1	117.5	136.9	140	146.1
Glu	292	125.9	108.4	145.5	25.5	120.7	157.4	179.9	183	186.2
Gln	171	125.6	107.1	140.4	23.7	122.2	162.1	183.2	189	193.2
Gly	312	54.2	36.2	65.5	44.7	54.6	75.7	75.6	85	88.1
His	115	129.3	109.3	140.2	22.0	118.8	152.5	187.2	194	202.5
Ile	230	122.2	106.4	136.2	21.8	115.3	158.8	188.4	182	181.0
Leu	408	131.5	108.8	146.1	25.5	116.1	148.4	192.2	180	193.1
Lys	249	149.8	130.9	167.3	21.8	160.8	192.6	209.9	211	225.8
Met	103	133.6	121.5	148.6	18.3	122.0	173.3	196.6	204	203.4
Phe	175	146.1	131.3	160.9	18.4	134.0	173.1	210.6	218	222.8
Pro	217	100.0	81.0	121.9	33.6	102.4	116.6	146.2	143	146.8
Ser	199	75.8	59.2	89.9	34.2	83.5	108.3	123.2	122	129.8
Thr	246	93.2	78.1	109.2	28.5	95.9	120.7	145.8	146	152.5
Trp	71	173.0	160.4	185.5	13.5	169.8	190.4	242.1	259	266.3
Tyr	149	156.8	141.5	172.2	17.8	148.7	185.8	218.0	229	236.8
Val	320	102.0	84.0	116.1	27.6	97.7	135.8	164.8	160	164.5
Mean		118	101	133	25	114	147	172	175	182

*Calculated using atomic radii from Richards (48).
†Calculated using atomic radii from Schrake and Rupley (49).
‡Calculated using atomic radii from Chothia (50).
§Calculated using atomic radii from Lee and Richards (51).
¶Total number of residues of that kind found in the 19 proteins simulated.
**Residue-specific solvent exposure averages.
††Minimum solvent exposure found in one of the 19 denatured ensembles.
‡‡Maximum solvent exposure found in one of the 19 denatured ensembles.
§§Percentage difference found between the maximum and minimum solvent exposures for one residue type: (max − min)/max × 100.

Gly-X-Gly tripeptides (12–14) yielded a mean solvent exposure (averaged over the 20 residue types) of 175–182 Å^2 (Table 1). These values were later lowered to around 172 Å^2 by including Ala-X-Ala tripeptides (15). A further and more substantial reduction took place when extended conformations of natural sequences corresponding to a data set of 43 proteins were used (17); which set the mean exposure to 147 Å^2. As explained by Creamer and co-workers (17), those polypeptide models were expected to display lower solvent exposures than the ones calculated from the tripeptide models, and were considered as upper bounds for residue-specific solvent exposures. Lower bounds were calculated by the same authors from 3- to 45-residue-long chain segments excised from native protein conformations, which gave rise to a mean exposure of 114 Å^2 (Table 1). Intermediate solvent exposures in the unfolded state between these upper and lower bounds have been recently described (19). According to our data on unfolded ensembles, the mean exposure is ~118 Å^2/residue, lying between the two bounds proposed by Creamer et al. (17), but much closer to the lower than to the upper bound. In fact, there is a good correlation between Creamer's lower-bound exposures of each of the 20 residues (Cr) and the corresponding averaged exposures calculated in the unfolded ensembles reported here (Ue): (Ue = 3.0 + 1.011 × Cr; r = 0.98, not shown. Although most residue types are, on average, more exposed to solvent in the unfolded ensembles than in Creamer's lower bound, there are noticeable exceptions, such as Ser and Lys. It is worth noting that Creamer's lower bounds are probably sensitive to the natural propensity of the different residues to appear in secondary structural elements. Therefore, in Creamer's lower bounds, those residues with a higher probability of appearing in loops could be biased to display larger solvent exposures. In fact, those residue types that have a lower solvent accessibility when using our methodology than Creamer's lower-bound one—Gly, Lys, Pro, Ser, and Thr—are either disruptors of secondary-structure elements or have a moderately low tendency to be in them according to the Chou-Fasman classification (45).

Table 2 displays the backbone and side-chain contributions to the averaged residue solvent exposures. The contribution of backbone atoms is very similar for all residues except for glycine, as expected. The polar and apolar exposed surfaces for the different residues are also shown in Table 2. All residue types expose significant amounts of both apolar and polar area to solvent (at least 30 and 13 Å^2, respectively) in the denatured ensemble. Atom-specific averaged solvent exposures for each residue type are provided in Supporting Material.

TABLE 2 Contributions to residue solvent accessibilities (Å^2) in the unfolded ensembles averaged over the19 proteins

Residue	Overall	Side chain*	Backbone*	Nonpolar†	Polar†
Ala	73.1	45.6	27.5	55.0	18.1
Arg	178.6	153.8	24.8	64.4	114.2
Asn	109.1	84.9	24.1	30.0	79.0
Asp	102.0	78.4	23.5	35.0	66.7
Cys	88.3	64.3	23.9	72.0	16.3
Glu	125.9	101.2	24.7	49.9	76.0
Gln	125.6	103.2	22.5	46.5	79.2
Gly	54.2	0.0	54.2	32.1	22.1
His	129.3	106.7	22.6	73.8	55.5
Ile	122.2	102.0	20.2	107.0	15.2
Leu	131.5	110.6	20.9	114.3	17.2
Lys	149.8	125.2	24.6	91.7	58.1
Met	133.6	111.6	22.1	116.6	17.1
Phe	146.1	124.6	21.5	129.8	16.3
Pro	100.0	77.4	22.6	86.6	13.4
Ser	75.8	49.9	25.9	41.9	33.9
Thr	93.2	70.7	22.5	59.6	33.6
Trp	173.0	152.0	21.0	136.7	36.3
Tyr	156.8	135.4	21.4	106.8	50.1
Val	102.0	81.9	20.1	87.1	14.9

*Definition of backbone includes the C_α atom.
†Definition of polarity is according to Naccess.

Large sequence dependence of solvent exposures

Unlike previous calculations of solvent exposures in the unfolded state, the ensembles described here allow a practical way to calculate sequence-specific solvent exposures for every residue of any particular protein. This is interesting, because it allows one to evaluate the extent to which different combinations of neighboring residues influence solvent exposures. The analysis of the unfolded ensembles corresponding to the 19 proteins modeled (Table 1) indicates that the solvent exposure of any residue type is strongly dependent on the sequence context. On average, there is a 26% difference in exposure for a given residue depending on the sequence. However, not all amino acid types show the same variability. The two extreme cases are Trp, with the smallest variation, 14%, and Gly, with the largest, 45%. This large neighboring effect clearly shows that any interpretation of the energetics of mutational experiments in terms of solvent-exposed area will benefit from knowledge of the exposures of the specific mutated residues in the denatured state, within their sequence contexts.

The sequences flanking the least and most exposed residue of each type are shown in Fig. 3 *a*. A statistical analysis has been performed to compare the enrichment of sequences in certain residues with respect to their overall abundance in the proteins studied (Fig. 3 *b*). A general prevalence of Pro immediately after poorly exposed residues was found. This is due to the proximity of the Pro Cδ atom, which also imposes the special conformational restriction to X-Pro residues (41). In addition to Pro, the sequences flanking the least exposed residues are rich in the three aromatic residues, Trp, Tyr and Phe, which, due to their size, can easily screen neighboring residues from solvent. The least exposed sequences are also moderately enriched in Gly. A possible explanation for this counterintuitive result is that the large conformational freedom of Gly would facilitate the peptide chain folding

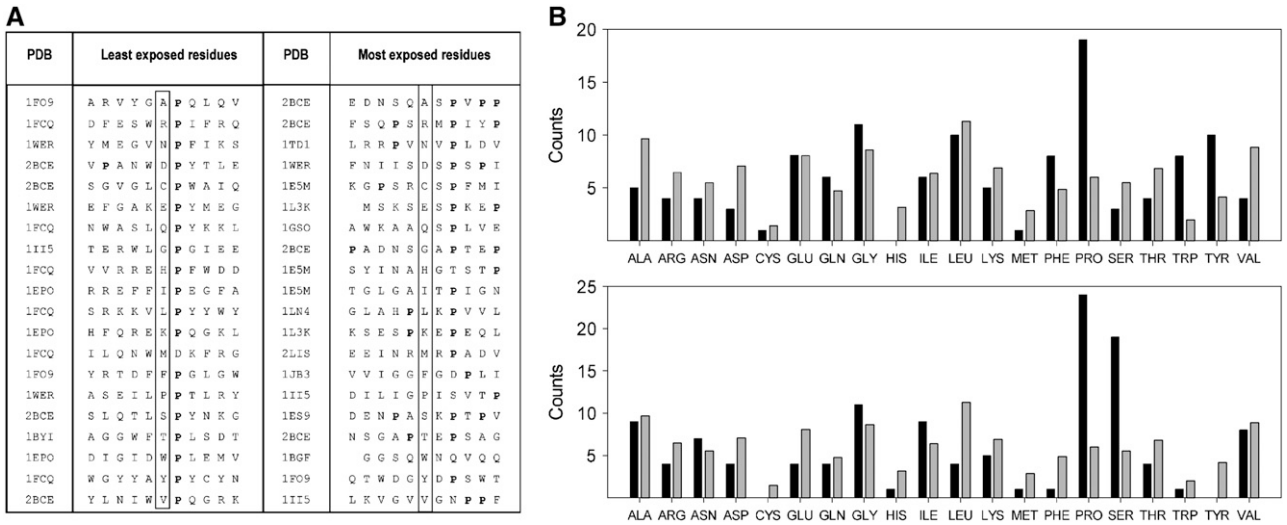

FIGURE 3 Influence of sequence context on solvent accessibilities. (*A*) The most and least accessible residues of each type found among the 19 proteins analyzed, shown in their sequence context and enclosed in black squares. Proline residues appear in bold to highlight the relevant role they play in determining high and low accessibilities. (*B*) Amino acid population in the least (*top*) and most (*bottom*) accessible residue sequences. Black bars represent the times a kind of residue is found in the three residues flanking the most and least accessible residues. Gray bars represent the times a kind of residue should be found at random, assuming the population statistics derived from the set of proteins studied.

into itself, thus enhancing solvent screening. On the other hand, most exposed sequences are rich in Pro as well. However, these Pro residues appear located at position $i + 2$ and $i + 3$, probably forming rigid elbows that could direct the following chain away from the exposed residue. Most exposed sequences are rich in small residues such as Ser and Gly, especially in the closest positions, and poor in large residues such as Tyr, Phe, Leu, and Arg.

Protein-size and protein-composition dependence of solvent exposures

One potential influence in the specific solvent exposure of any protein residue is the protein size. It is not clear how often remote residues in the sequence are brought into proximity in the unfolded ensemble. We have analyzed this issue in the following way. The global average accessibilities for the unfolded ensembles of each of the 19 proteins in the data set have been compared with their corresponding predicted accessibilities, as calculated from their amino acid composition and the averaged solvent exposures of each residue type in the data set (Table 1). Notice that the five terminal residues were not used in this calculation because their high solvent accessibilities essentially reflect their terminal position. The differences between averaged and estimated global accessibilities for the 19 studied proteins are plotted as percentages in Fig. 4 *a*, highlighting those proteins that are more (positive) or less (negative) solvent-accessible than expected based on their amino acid composition. A weak correlation is observed with protein size, suggesting that residues in larger proteins may be slightly less exposed than in smaller ones, but the scatter of the data is too large to be explained by the protein-size effect alone. To clarify whether there is a specific influence of sequence length on solvent exposures, we have generated polyalanine chains of different lengths and calculated the solvent accessibility of the 11-residue-long central fragment (Fig. 4 *b*). A systematic decrease is observed when chain length increases, indicating an effect of remote residues on residue solvent exposures.

Although the latter observation can explain the general trend observed in Fig. 4 *a*, additional factors must play a significant role in the global solvent accessibility in the denatured state. In that respect, we explored whether there is a role of the amino acid composition in the overall solvent exposure of protein ensembles. To that end, we compared the amino acid composition of 1LN4 and 1TD1, two proteins of 98 and 100 residues, respectively, that display completely different behavior regarding their solvent exposure (Fig. 4 *a*). On one hand, 1LN4 is the protein that presents the highest percentage of increased accessibility, 2.92, whereas 1TD1 is a relatively solvent-screened protein with a negative percentage, −0.17. As shown in Fig. 4 *c*, 1TD1 contains a high number of bulky amino acids, such as Phe, Tyr, and Arg, which have been found often in low accessible sequences (Fig. 3 *a*). In addition, 1TD1 shows a low occurrence of small residues such as Ala and Thr. On the other hand, 1LN4 lacks both Trp and Phe residues, and it contains only one Tyr. Therefore, it seems that an abundance of bulky or small residues may influence the overall solvent exposure of a given protein in the unfolded state. As a whole, the analysis of the 19 unfolded ensembles indicates that both protein size and amino acid composition exert some influence on the

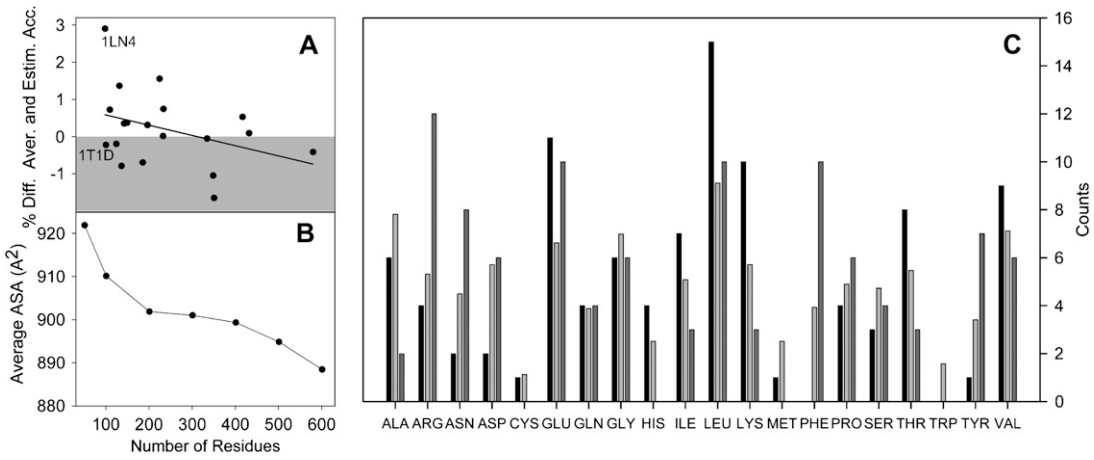

FIGURE 4 Influence of size and amino acid composition on the total solvent accessibility of the set of 19 proteins. (*A*) Percentage difference between the solvent exposures averaged over the 2000 conformations of the atomistic models of the denatured ensembles, and the ones obtained using protein composition and the residue averaged values shown in Table 1. Positive values (*white area*) indicate that the actual protein is more accessible than expected, whereas negative values (*gray area*) indicate less solvent accessibility than expected from the contribution of individual residues. The solid line represents the linear regression slope. (*B*) Solvent accessibility of the 11-amino-acid-long central fragment of polyalanine chains of different lengths. (*C*) Amino acid composition of proteins 1LN4 (*left bars*) and 1TD1 (*right bars*) bearing 98 and 100 residues, respectively. Middle bars correspond to the number of residues expected for a protein of that size that followed the residue statistics of the 19 proteins in the data set.

solvent exposure of denatured proteins (even in the absence of unpredictable clustering interactions), and therefore reinforces the importance of calculating protein-specific solvent exposures, where those effects are specifically accounted for.

Global solvent accessibilities

Different thermodynamic terms that characterize the free energy of unfolding, ΔH, ΔS, and ΔC_p, have been parameterized in terms of the change in the total (ΔA_{tot}), the apolar (ΔA_{ap}), and the polar (ΔA_{pol}) surface area exposed upon unfolding. A detailed description of these relationships can be found in Robertson and Murphy (46). Whereas calculating the surface accessibility of the folded state is straightforward if the three-dimensional structure of the protein is available, performing the same calculation for the denatured state is much more difficult, and approximations based on tripeptides or small protein fragments (see Introduction) are normally used. Using more realistic polar and apolar solvent exposures may help to improve the parameterizations. Total, as well as apolar and polar, accessibilities in the unfolded states of the 19 proteins studied are shown in Table 3. Notice that in this case, all residues in the sequence were used for the calculation. The accessibilities obtained using the atomic model of the denatured ensembles are in good agreement with those estimated using the amino acid compositions and the average residue accessibilities. The correlations of the calculated and estimated global accessibilities, shown in Fig. S1 of Supplementary Material, present slopes close to 1.0, and the root-mean-square deviations found are 324, 186, and 141 Å^2 for the total, apolar, and polar solvent accessibilities, respectively. These results indicate that good approximations for the global, as well as the apolar and polar, accessibilities of the denatured state of proteins can be estimated from the residue averaged solvent exposures derived in this study. Notice that in the reported residue averaged solvent exposures (Table 1), the excess of exposed surface typically present in the chain termini is not accounted for. Therefore, the accessibilities derived from the detailed simulations of the denatured state ensembles are more accurate.

It is interesting that all three solvent exposures also present good correlations with the number of residues in the protein, with regression coefficients >0.98. Total, apolar, and polar solvent exposures follow the relationships

$$A_{tot}\,(\text{Å}^2) = 1740\,(\pm 424) + 107.8\,(\pm .6) \times N, r^2 = 0.996,$$

$$A_{ap}\,(\text{Å}^2) = 715.2\,(\pm 245.1) + 70.6\,(\pm 0.9) \times N, r^2 = 0.997,$$

and

$$A_{pol}\,(\text{Å}^2) = 1025\,(\pm 292) + 37.2\,(\pm 1.1) \times N r^2 = 0.986,$$

where N is the number of residues of the protein. This correlation indicates that the global, as well as the apolar and polar, accessibilities of the denatured state of proteins can in principle be derived from the number of residues. However, the root-mean-square deviations from those fittings (848, 491, and 585 Å^2, respectively) are notably larger than those calculated from the averaged residue-specific accessibilities.

CONCLUSIONS

In this work, a methodology has been presented for the calculation of protein solvent exposures based on a detailed description of the denatured state that is consistent with

TABLE 3 Calculated solvent accessibilities for the 19 proteins in the folded state and in the unfolded ensemble

Protein PDB code	Number of residues	Access. folded (Å^2)*	Access. unfolded simul. (Å^2)†	Access. unfolded estim. (Å^2)‡	% Difference access. §	Total access. (Å^2) ¶	Total apolar access. (Å^2) §	Total polar access. (Å^2) §
1LN4	98	5821.4	10497.0	10200.5	2.91	11825.7	7657.5	4168.2
1T1D	100	5913.9	11136.0	11160.9	−0.22	12544.1	7498.6	5045.5
1BKR	109	6061.9	11666.1	11582.3	0.72	13193.2	8557.6	4635.6
1BGF	124	7876.8	13807.9	13835.2	−0.20	15045.6	9371.1	5674.4
1JB3	131	7633.7	14261.4	14069.0	1.37	15749.6	9833.6	5916.1
2LIS	136	8658.0	15302.0	15423.8	−0.79	16891.6	10996.5	5895.1
1QGV	142	7120.8	15632.2	15577.2	0.35	17217.5	11119.0	6098.4
1EY4	149	7550.9	16429.7	16368.8	0.37	17580.1	10961.7	6618.4
1EPO	185	9640.5	20528.6	20672.5	−0.70	21983.5	13742.6	8240.9
1L3K	196	9317.0	21435.8	21369.1	0.31	22859.8	13828.0	9031.8
1BYI	224	10779.9	23392.9	23034.1	1.56	24902.5	16478.2	8424.2
1ES9	232	9778.3	25249.9	25245.9	0.02	26485.5	16765.2	9720.4
1II5	233	10175.6	25257.8	25070.8	0.75	26463.3	17119.2	9344.1
1WER	334	16682.6	37845.4	37865.6	−0.06	39325.3	25428.3	13897.0
1FO9	348	14893.8	39491.8	39910.2	−1.05	40895.2	26150.2	14744.9
1FCQ	350	14623.7	39876.8	40546.4	−1.65	41097.8	25503.0	15594.9
1E5M	416	15655.5	43429.1	43200.2	0.53	44935.2	29141.7	15793.6
1GSO	431	18884.8	45628.2	45586.4	0.09	47154.8	30721.9	16432.9
2BCE	579	20718.8	62630.6	62891.7	−0.42	63887.8	41742.2	22145.6

Access., accessibility; simul., simulation; estim., estimation.

*Solvent accessible area calculated with Naccess for the native structures.

†Solvent accessibility averaged over 2000 unfolded structures. Calculated without taking into account the first and the last five residues of each protein.

‡Estimated from the amino acid composition and the averaged values of each residue type, as found in Table 1. Calculated without taking into account the first and the last five amino acids.

§Percentage difference between the solvent exposures calculated and those estimated from residue-type average values. Negative (positive) values indicate less (more) solvent exposure than that estimated from the amino acid composition.

¶Solvent accessibility averaged over 2000 unfolded structures. Calculated using the whole sequence.

diverse biophysical data measured in solution for natively and chemically denatured proteins (31). The improved description of the unfolded state provides ensembles of conformations that can be confidently used to describe biophysical parameters that may be difficult to predict using simplified models or molecular dynamics simulations (10,47).

The unfolded states of 19 proteins with low sequence and structural homology have been simulated. They represent a database of residues large enough to allow deriving statistically robust averaged atom- and residue-specific solvent accessibilities that can in turn be used for the parameterization of the different contributions involved in protein stability. It is important that, despite the usefulness of those averaged solvent exposures, the sequence-specific context of the different residues of any particular protein exerts a strong influence on the solvent exposures, and thus sequence-specific solvent exposures of the residue of interest should be used for the interpretation of mutational studies. On the other hand, we anticipate that the simulated unfolded ensembles could be useful to investigate the elusive balance of interactions occurring in the native and denatured states that is so important for understanding protein stability.

We thank the BIFI computer center staff for support.

J.S. acknowledges financial support from Institute of Biocomputation and Physics of Complex Systems (BIFI) grant BFU2004-01411. M.B. acknowledges UMR 5075, Centre National de la Recherche Scientifique, Commissariat a l'Energie Atomique, and Universite Joseph Fourier Grenoble, and ANR NT05-4_42781 for financial support. P.B. acknowledges the European Molecular Biology Organization for a long-term fellowship and funds from the Ramon y Cajal program (Spain).

REFERENCES

1. Shortle, D. 1996. The denatured state (the other half of the folding equation) and its role in protein stability. *FASEB J.* 10:27–34.
2. Funahashi, J., Y. Sugita, A. Kitao, and K. Yutani. 2003. How can free energy component analysis explain the difference in protein stability caused by amino acid substitutions? Effect of three hydrophobic mutations at the 56th residue on the stability of human lysozyme. *Protein Eng. Des. Sel.* 16:665–671.
3. Lazaridis, T., and M. Karplus. 2003. Thermodynamics of protein folding: a microscopic view. *Biophys. Chem.* 100:367–395.
4. Myers, J. K., and C. N. Pace. 1996. Hydrogen bonding stabilizes globular proteins. *Biophys. J.* 71:2033–2039.
5. Campos, L. A., S. Cuesta-Lopez, J. Lopez-Llano, F. Falo, and J. Sancho. 2005. A double-deletion method to quantifying incremental binding energies in proteins from experiment: example of a destabilizing hydrogen bonding pair. *Biophys. J.* 88:1311–1321.
6. Kono, H., M. Saito, and A. Sarai. 2000. Stability analysis for the cavity-filling mutations of the Myb DNA- binding domain utilizing free-energy calculations. *Proteins.* 38:197–209.
7. Gatchell, D. W., S. Dennis, and S. Vajda. 2000. Discrimination of near-native protein structures from misfolded models by empirical free energy functions. *Proteins.* 41:518–534.
8. Baldwin, R. L. 1986. Temperature dependence of the hydrophobic interaction in protein folding. *Proc. Natl. Acad. Sci. USA.* 83:8069–8072.

9. Kauzmann, W. 1959. Some factors in the interpretation of protein denaturation. *Adv. Protein Chem.* 14:1–63.
10. Privalov, P. L. 1979. Stability of proteins: small globular proteins. *Adv. Protein Chem.* 33:167–241.
11. Dill, K. A. 1990. Dominant forces in protein folding. *Biochemistry*. 29: 7133–7155.
12. Miller, S., J. Janin, A. M. Lesk, and C. Chothia. 1987. Interior and surface of monomeric proteins. *J. Mol. Biol.*196:641–656.
13. Shrake, A., and J. A. Rupley. 1973. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *J. Mol. Biol.* 79: 351–371.
14. Rose, G. D., A. R. Geselowitz, G. J. Lesser, R. H. Lee, and M. H. Zehfus. 1985. Hydrophobicity of amino acid residues in globular proteins. *Science*. 229:834–838.
15. Zielenkiewicz, P., and W. Saenger. 1992. Residue solvent accessibilities in the unfolded polypeptide chain. *Biophys. J.* 63:1483–1486.
16. Creamer, T. P., R. Srinivasan, and G. D. Rose. 1995. Modeling unfolded states of peptides and proteins. *Biochemistry*. 34:16245–16250.
17. Creamer, T. P., R. Srinivasan, and G. D. Rose. 1997. Modeling unfolded states of proteins and peptides. II. Backbone solvent accessibility. *Biochemistry*. 36:2832–2835.
18. Pappu, R. V., R. Srinivasan, and G. D. Rose. 2000. The Flory isolated-pair hypothesis is not valid for polypeptide chains: implications for protein folding. *Proc. Natl. Acad. Sci. USA.* 101:12565–12570.
19. Goldenberg, D. P. 2003. Computational simulation of the statistical properties of unfolded proteins. *J. Mol. Biol.* 326:1615–1633.
20. Fersht, A. R., A. Matouschek, and L. Serrano. 1992. The folding of an enzyme. I. Theory of protein engineering analysis of stability and pathway of protein folding. *J. Mol. Biol.* 224:771–782.
21. Freire, E. 1993. Structural thermodynamics: prediction of protein stability and protein binding affinities. *Arch. Biochem. Biophys.* 303: 181–184.
22. Shortle, D., and M. S. Ackerman. 2001. Persistence of native-like topology in a denatured protein in 8 M urea. *Science*. 293:487–489.
23. Ohnishi, S., A. L. Lee, M. H. Edgell, and D. Shortle. 2004. Direct demonstration of structural similarity between native and denatured eglin C. *Biochemistry*. 43:4064–4070.
24. Tran, H. T., X. Wang, and R. V. Pappu. 2005. Reconciling observations of sequence-specific conformational propensities with the generic polymeric behavior of denatured proteins. *Biochemistry*. 44:11369–11380.
25. Shi, Z., K. Chen, Z. Liu, and N. R. Kallenbach. 2006. Conformation of the backbone in unfolded proteins. *Chem. Rev.* 106:1877–1897.
26. Tjandra, N., and A. Bax. 1997. Direct measurement of distances and angles in biomolecules by NMR in a dilute liquid crystalline medium. *Science*. 278:1111–1114.
27. Louhivouri, M., K. Pääkkönen, K. Fredriksson, P. Permi, J. Lounila, and A. Annila. 2003. On the origin of residual dipolar couplings from denatured proteins. *J. Am. Chem. Soc.* 125:15647–15650.
28. Mohana-Borges, R., N. K. Goto, G. J. A. Kroon, H. J. Dyson, and P. E. Wright. 2004. Structural characterization of unfolded states of apomyoglobin using residual dipolar couplings. *J. Mol. Biol.* 34: 1131–1142.
29. Meier, S., S. Güthe, T. Kiefhaber, and S. Grzesiek. 2004. Foldon, the natural trimerization domain of T4 fibritin, dissociates into a monomeric A-state form containing a stable β-hairpin: atomic details of trimer dissociation and local β-hairpin stability from residual dipolar couplings. *J. Mol. Biol.* 344:1051–1069.
30. Jha, A. K., A. Colubri, K. F. Freed, and T. R. Sosnick. 2005. Statistical coil model of the unfolded state: resolving the reconciliation problem. *Proc. Natl. Acad. Sci. USA.* 102:13099–13104.
31. Bernadó, P., L. Blanchard, P. Timmins, D. Marion, R. W. H. Ruigrok, and M. Blackledge. 2005. A structural model for unfolded proteins from residual dipolar couplings and small-angle x-ray scattering. *Proc. Natl. Acad. Sci. USA.* 102:17002–17007.
32. Bernadó, P., C. W. Bertoncini, C. Griesinger, M. Zweckstetter, and M. Blackledge. 2005. Defining long-range oand local disorder in native α-synuclein using residual dipolar couplings. *J. Am. Chem. Soc.* 127: 17968–17969.
33. Gillespie, J. R., and D. Shortle. 1997. Characterization of long-range structure in the denatured state of staphylococcal nuclease. I. Paramagnetic relaxation enhancement by nitroxide spin labels. *J. Mol. Biol.* 268:158–169.
34. Lindorff-Larsen, K., S. Kristjansdottir, K. Teilum, W. Fieber, C. M. Dobson, F. M. Poulsen, and M. Vendruscolo. 2004. Determination of an ensemble of structures representing the denatured state of the of the bovine acyl-coenzyme A binding protein. *J. Am. Chem. Soc.* 126: 3291–3299.
35. Bertoncini, C. W., Y.-S. Jung, C. O. Fernandez, W. Hoyer, C. Griesinger, T. M. Jovin, and M. Zweckstetter. 2005. Release of long-range tertiary interactions potentiates aggregation of natively unstructured α-synuclein. *Proc. Natl. Acad. Sci. USA.* 102:1430–1435.
36. Dedmon, M. W., K. Lindorff-Larsen, J. Christodoulou, M. Vendruscolo, and C. M. Dobson. 2005. Mapping long-range interactions in α-synuclein using spin-label NMR and ensemble molecular dynamics simulations. *J. Am. Chem. Soc.* 127:476–477.
37. Doniach, S. 2001. Changes in biomolecular conformation seen by small angle x-ray scattering. *Chem. Rev.* 101:1763–1798.
38. Kohn, J. E., I. S. Millett, J. Jacob, B. Zagrovic, T. M. Dillon, N. Cingel, R. S. Dothager, S. Seifert, P. Thiyagarajan, T. R. Sosnick, M. Z. Hasan, V. S. Pande, I. Ruczinski, S. Doniach, and K. W. Plaxco. 2004. Random-coil behaviour and dimensions of chemically unfolded proteins. *Proc. Natl. Acad. Sci. USA.* 101:12491–12496.
39. Fitzkee, N. C., and G. D. Rose. 2004. Reassessing random-coil statistics in unfolded proteins. *Proc. Natl. Acad. Sci. USA.* 101:12497–12502.
40. Lovell, S. C., I. W. Davis, W. B. Arendall III, P. I. W. de Bakker, J. M. Word, M. G. Prisant, J. S. Richardson, and D. C. Richardson. 2003. Structure validation by Cα geometry: ϕ,ψ and Cβ deviation. *Proteins*. 50:437–450.
41. MacArthur, M. W., and J. M. Thornton. 1991. Influence of proline residues on protein conformation. *J. Mol. Biol.* 218:397–412.
42. Levitt, M. 1976. A simplified representation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol.* 104:59–107.
43. Eyal, E., R. Najmanovich, B. J. McConkey, M. Edelman, and V. Sobolev. 2004. Importance of solvent accessibility and contact surfaces in modeling side-chains conformations in proteins. *J. Comp. Chem.* 25:712–724.
44. Hubbart, S. J., and J. M. Thornton. 1993. NACCESS Computer Program. Department of Biochemistry and Molecular Biology, University College London, London, UK.
45. Chou, P. Y., and G. D. Fasman. 1974. Conformational parameters for amino acids in helical, β-sheet, and random coil regions calculated from proteins. *Biochemistry*. 13:211–222.
46. Robertson, A. D., and K. P. Murphy. 1997. Protein structure and the energetics of protein stability. *Chem. Rev.* 97:1251–1267.
47. Lazaridis, T., and M. Karplus. 1999. Discrimination of the native from misfolded protein models with an energy function including implicit solvation. *J. Mol. Biol.* 288:477–487.
48. Richards, F. M. 1977. Areas, volumes, packing and protein structure. *Annu. Rev. Biophys. Bioeng.* 6:151–176.
49. Schrake, A., and J. A. Rupley. 1973. Environment and exposure to solvent of protein atoms. *J. Mol. Biol.* 79:351–371.
50. Chothia, C. 1975. Structural invariants in protein folding. *Nature*. 254: 304–308.
51. Lee, B., and Richards, F. M. 1971. The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* 55:379–400.