Protein folding: could hydrophobic collapse be coupled with hydrogen-bond formation?

Ariel Fernández[a,b,*], József Kardos[b,c], Yuji Goto[b]

[a] *Institute for Biophysical Dynamics, The University of Chicago, Chicago, IL 60637, USA*
[b] *Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka 565, Japan*
[c] *Eötvös University, Department of Biochemistry, Budapest H-1117, Hungary*

Received 27 November 2002; revised 31 December 2002; accepted 13 January 2003

First published online 23 January 2003

Edited by Robert B. Russell

Abstract A judicious examination of an exhaustive PDB sample of soluble globular proteins of moderate size ($N<102$) reveals a commensurable relationship between hydrophobic surface burial and number of backbone hydrogen bonds. An analysis of 50 000 conformations along the longest all-atom MD trajectory allows us to infer that not only the hydrophobic collapse is concurrent with the formation of backbone amide-carbonyl hydrogen bonds, they are also dynamically coupled processes. In statistical terms, hydrophobic clustering of the side chains is inevitably conducive to backbone burial and the latter process becomes thermodynamically too costly and kinetically unfeasible without amide-carbonyl hydrogen-bond formation. Furthermore, the desolvation of most hydrogen bonds is exhaustive along the pathway, implying that such bonds guide the collapse process.

Key words: Protein folding; Hydrophobic collapse; Hydrogen bonds; Statistical mechanics

1. Introduction

The intensely debated issue of whether hydrophobic collapse occurs concomitantly or is even coupled with secondary structure formation is central to understand the factors that efficiently guide the protein folding process [1–6]. Recent experimental evidence suggests that backbone H-bond formation is commensurate with surface burial at the transition state [6], while compaction and extent of secondary structure for a few conformations appear to be linearly correlated [2,3]. Furthermore, we report here a statistically significant correlation between total surface burial and number of desolvated backbone H-bonds for single-domain PDB proteins. This evidence implies commensurability between structure formation and hydrophobic clustering, but are both processes dynamically coupled? Thermodynamic analysis [7] suggests that the protein cannot desolvate the backbone without making H-bonds: the free-energy cost of burying an unbound backbone amide and carbonyl is $\sim$5.5 kcal/mol higher than the cost of burying them as a H-bonded pair. To answer the question, we examined $\sim$50 000 *villin headpiece* conformations from the longest (1 μs) all-atom simulation [8]. Our findings reveal a dynamic coupling: (a) hydrophobic surface burial remains statistically proportional to backbone surface burial along the entire trajectory; (b) to lower the free-energy cost, desolvated amide-carbonyl H-bonds become the only plausible mode of backbone burial. Only backbone hydrogen bonds will be dealt with here, as they are the primary determinants of the basic structural motifs, and involve polar groups, the amides and carbonyls, whose self energies are directly affected by the hydrophobic collapse [7,9–13].

Two generic scenarios of protein folding have been proposed and remain conflictive to this day: (I) a secondary structure framework is formed first and induces the subsequent chain compaction by hydrophobic clustering [2,9]; (II) hydrophobic collapse occurs first and induces the subsequent formation of secondary and tertiary structure [10,11]. The aim of this work is to show that not only both processes are concerted [3]: they are also coupled and most backbone H-bonds occur only if thoroughly surrounded by hydrophobic groups.

First, we observed a statistically significant linear correlation in single-domain globular proteins of moderate size ($N<102$) between desolvated backbone H-bonds and total surface burial (Fig. 1A). Desolvated (dehydrated) H-bonds are important structural determinants because of their high stability which stems from the high free-energy cost associated with the exposure of the unbound amide and carbonyl groups to a water-deprived environment [5]. An exhaustive PDB sample of 2801 monomeric soluble proteins free from sequence redundancies was examined (Section 2). To perform the statistical analysis we defined an H-bond desolvation domain as made up of two intersecting spheres of radius 7 Å centered at the α-carbons of paired residues [5], and established that 16 ± 6 carbonaceous groups (CH_i, $i=1, 2, 3$) of side chains are contained within the desolvation domains of 92% of the H-bonds examined across the large PDB sample. Thus, a desolvated H-bond is statistically defined as being surrounded by at least 10 hydrophobic groups. The inferences made are robust within the range $R=7.0\pm0.2$ Å. An H-bond is operationally defined by an N–O distance within the range 2.6–3.44 Å and a 60° range in the N–H–O angle. The cut-off distances have not been chosen arbitrarily, rather they reflect the maximum ranges of significant pairwise interactions, that is, of those interactions which are at least of the order of the thermal fluctuation parameter RT (R = gas constant, T = absolute temperature) [12,13,15–18].

Thus, a clearcut property of soluble single-domain proteins emerges from Fig. 1A: The number of desolvated backbone

*Corresponding author. Fax: (1)-773-702 0439.
E-mail address: ariel@uchicago.edu (A. Fernández).

0014-5793/03/$22.00 © 2003 Published by Elsevier Science B.V. on behalf of the Federation of European Biochemical Societies.
doi:10.1016/S0014-5793(03)00056-5

hydrogen bonds is statistically commensurate with the total surface area burial, yielding a slope of 7 ± 0.4 desolvated H-bonds/1000 Å^2 for the best linear statistical fits within the confidence band. This observation only identifies a trend comparing the final results of folding processes: the native structures. It hints but does not necessarily imply that secondary structure formation must be coupled with hydrophobic collapse: a dynamic study is required.

2. Methods

To properly examine the 50 000 conformations of the Duan–Kollman trajectory [8] with our representational tools [5], we removed from it the time-evolving solvent coordinates. The sequential estimation of solvent-accessible surface areas for the conformations considered required a faster and necessarily coarser algorithm than GetArea1.1 [14], the algorithm used to generate Fig. 1A. Thus, the apolar, backbone and side-chain contributions for each residue averaged over 10^4 random-coil (rc) conformations are first determined. This ensemble is constructed by random attribution of (Φ, Ψ)-intrinsic coordinates to each residue governed by a distribution derived from the Boltzmann weighting of conformations in the Ramachandran plots [18]. Since the Ramachandran plots are built based solely on local side-chain–backbone interactions, no structural bias is incorporated in the rc ensemble.

To estimate the contribution per residue for a particular chain conformation, the rc-contribution is multiplied by a coefficient, γ, which describes the interactive context of the residue in the specified chain conformation. The coefficient γ is obtained by defining an interactive sphere of radius 7 Å centered at the α-carbon of the residue [16,18] and computing the fraction f of volume of that sphere occupied by other regions of the chain. Thus, we get $\gamma = 1 - f$ (so, for $f = 0$, i.e. no interactions, we get the random coil result). We tested this approximation on 100 moderately small proteins from the PDB, like the ones reported in Fig. 1A. Our coarse but fast estimation of the solvent-exposed areas for PDB native structures differs invariably by less than 6% from the GetArea1.1 values. For larger proteins ($N > 70$) the deviation becomes important and appears to scale with $N^{1/2}$, as befits a surface relation.

To determine a consistent criterion enabling us to classify a H-bond as buried, an exhaustive PDB sample of 2801 monomeric soluble proteins ($N < 102$) free from sequence redundancies was examined. If we define an H-bond desolvation domain as made up of two intersecting desolvation spheres of radius 6.3 Å centered at the α-carbons of the paired residues, we find that 16 ± 6 nonpolar carbonaceous groups (CH_i, $i = 1, 2, 3$) of the side chains are contained within the desolvation domains of 92% of the H-bonds examined. Thus, an under-desolvated H-bond is statistically defined as being surrounded by nine or less hydrophobic groups. The inferences made are robust within the range $R = 7.0 \pm 0.2$ Å. An H-bond is here operationally defined by an N–O distance within the range 2.6–3.44 Å and a 60° range in the N–H–O angle.

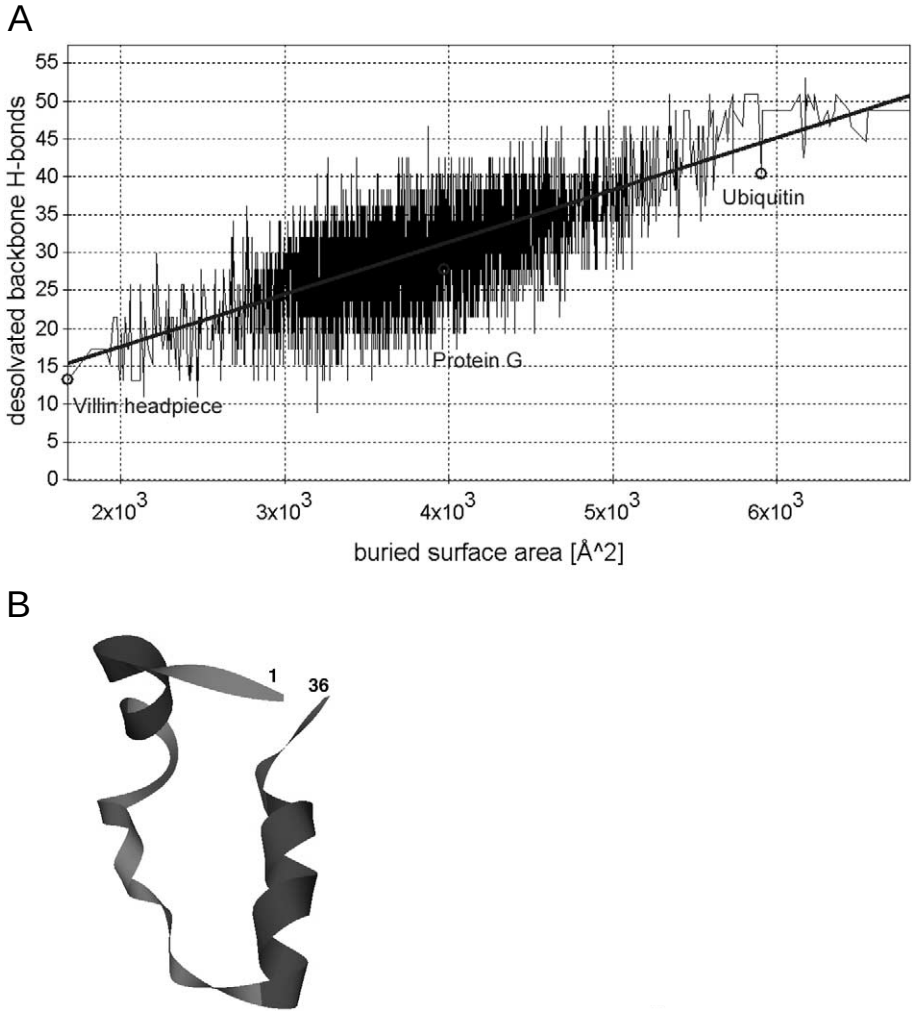

Fig. 1. A: Desolvated backbone H-bonds versus total surface area buried for single-domain soluble globular proteins from the exhaustive PDB sample ($N < 102$) specified in Section 2. A statistically significant linear correlation is drawn for the best fit within the confidence band: $7(\pm 0.4)$ H-bonds/1000 Å^2. B: Ribbon representation of native structure for *villin headpiece* (pdb.1vii).

3. Results

3.1. Dynamic analysis of backbone desolvation

The extent to which a protein desolvates its backbone as it buries hydrophobic groups results from the geometric constraints to which the backbone is subject [5,12]. The examination of the folding trajectory [8] for *villin headpiece* (Fig. 1B) using novel representational tools (Section 2) unambiguously reveals a proportionality between total hydrophobic surface burial and backbone polar surface burial (Fig. 2A), implying that hydrophobic side chains cannot be clustered together without concurrently desolvating the amides and carbonyls of the backbone [4,5,12].

On the other hand, thermodynamic-cycle considerations [7] reveal that the free-energy cost associated with transferring the unbound backbone amide and carbonyl from water to an organic phase is at least ~5.5 kcal/mol higher than that associated with transferring them as a H-bonded pair (+6.12 kcal/mol vs. +0.62 kcal/mol). Thus, the chain backbone is preferably desolvated only when it can form H-bonds, an inference robust to new and consistent calorimetric parameter revisions [7,13–15]. These facts alone imply that backbone H-bonds cannot be adventitious byproducts of chain compaction and that, in contrast with earlier treatments of the problem [1,10], they must be taken into account.

From the previous considerations and Figs. 1A and 2A we

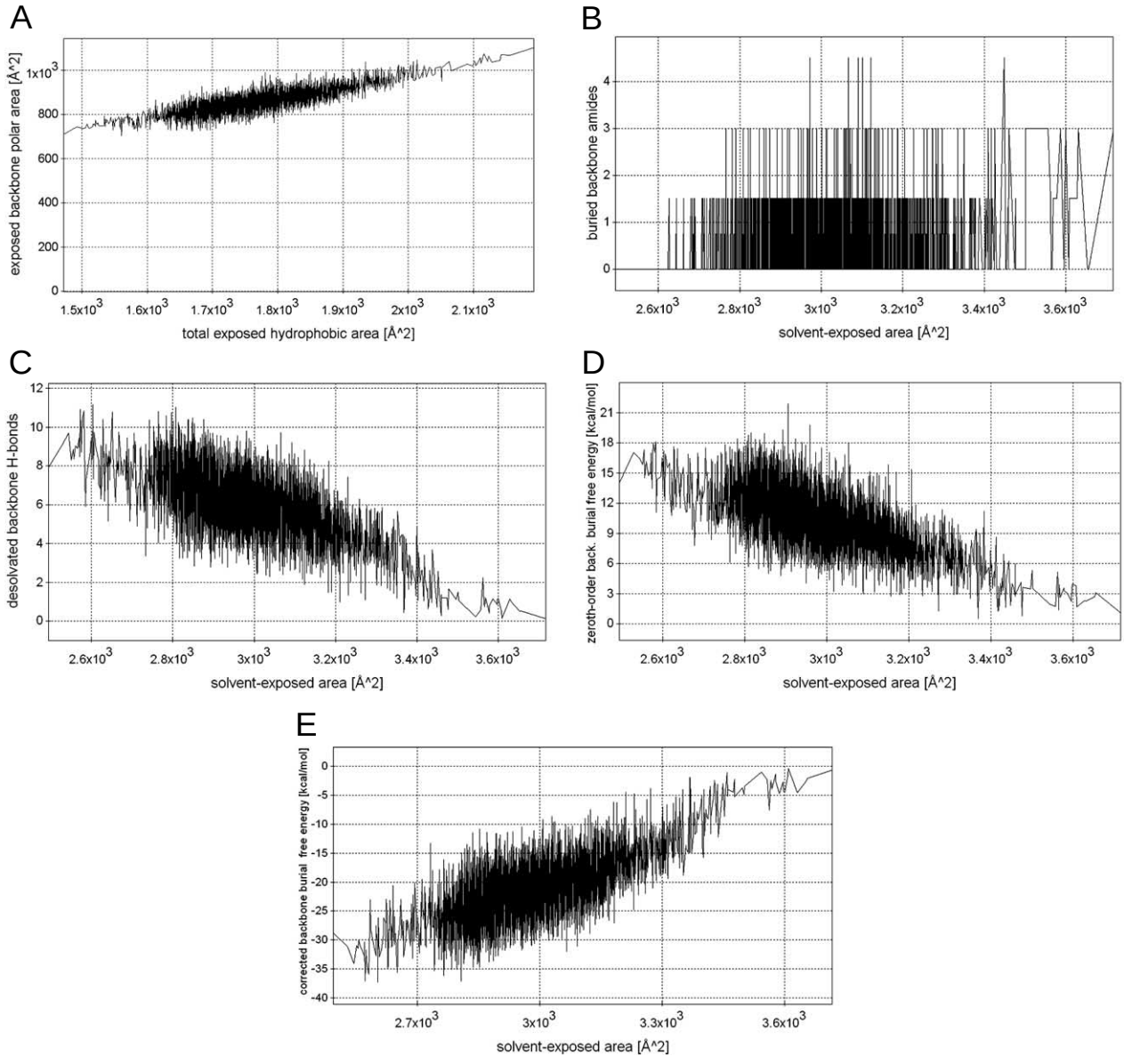

Fig. 2. Plots from the Duan–Kollman trajectory. Each point in the plots represents an average over 50 conformations. A: Solvent-exposed backbone area versus solvent-exposed area from hydrophobic regions of the side chains; B: number of non-bonded desolvated backbone amides versus total solvent-exposed area; C: number of desolvated backbone H-bonds versus total solvent-exposed area; D: zeroth-order free energy of backbone hydration versus total solvent-exposed area. The free energy increases as the backbone becomes buried and its reference value (0 kcal/mol) is adopted for the fully hydrated backbone. E: Hydration free energy of the backbone corrected for amide-carbonyl H-bond formation versus total solvent-exposed area. The net free-energy benefit is actually lower, because we must subtract from the ~51 kcal/mol the cost of forming the hydrogen bonds in bulk water (~27.5 kcal/mol if we are to follow Roseman's figures [7], for alternative estimations see main text).

may suspect that hydrophobic collapse must be coupled with backbone H-bond formation. This should be especially the case if the penalty for increasing the solvation free energy of amides and carbonyls combined with the conformational entropy cost associated with the hydrophobic clustering cannot be compensated by the free-energy gain resulting from the hydrophobic association. This is indeed the case, as direct inspection of Fig. 2B,C reveals (each point represents an average over 50 consecutive conformations): While the number of unbound yet buried amide groups of the backbone remains low (<5) and uncorrelated with the extent of surface burial (Fig. 2B), the number of backbone H-bonds is strongly correlated and even statistically proportional to surface burial within the confidence bands determined by the structural fluctuations (Fig. 2C).

A backbone amide is classified as buried when the number of side-chain nonpolar carbonaceous groups (CH_i, $i=1, 2, 3$) within a desolvation sphere of radius 6.3 Å centered at the N-atom is higher than 8 (cf. [5,16]). This criterion is robust to moderate alterations in the desolvation radius (± 0.3 Å) and is based on the fact that the number of side-chain hydrophobic groups desolvating a backbone amide across a rc ensemble consisting of 10^4 conformations is 4 ± 3.

The conclusion that transpires from Fig. 2B,C is that the intramolecular hydrophobic clustering occurs concurrently with the formation of backbone H-bonds and in this way, the protein dramatically lowers the thermodynamic cost associated with the inevitable and concomitant burial of the backbone. This conclusion is corroborated by a comparison between the zeroth-order or uncorrected free energy of backbone hydration (Fig. 2D), obtained using a standard parameter compilation of thermodynamic coefficients [15,17] for solvent exposure of different atoms, with the plot corrected to include the lower cost of burying amides and carbonyls as H-bonded pairs [7] (Fig. 2E). Each correction for a backbone H-bond is taken to be proportional to its extent of burial, λ, with $0<\lambda<1$ [7,18]. Thus, for each backbone H-bond formed, the value $-\lambda \times 5.5$ kcal/mol (cf. ref. [7]) is algebraically added to the total backbone hydration free energy. The burial coefficient is defined as $\lambda = v/V$, where v = volume of van der Waals spheres of hydrophobic groups contained in the desolvation domain of the H-bond and V = total volume of the H-bond desolvation domain. The cost of making the amide-carbonyl hydrogen bonds in bulk water is a separate and destabilizing contribution ($\Delta G° = +3.1$ kcal/mol per hydrogen bond, according to Roseman [7]) which does not offset the thermodynamic benefit of burying the backbone amides and carbonyls as hydrogen bonded pairs. Fig. 2D,E shows that a free-energy decrease of $\sim 51.0 \pm 1.5$ kcal/mol in the cost of burying the backbone results from the formation of

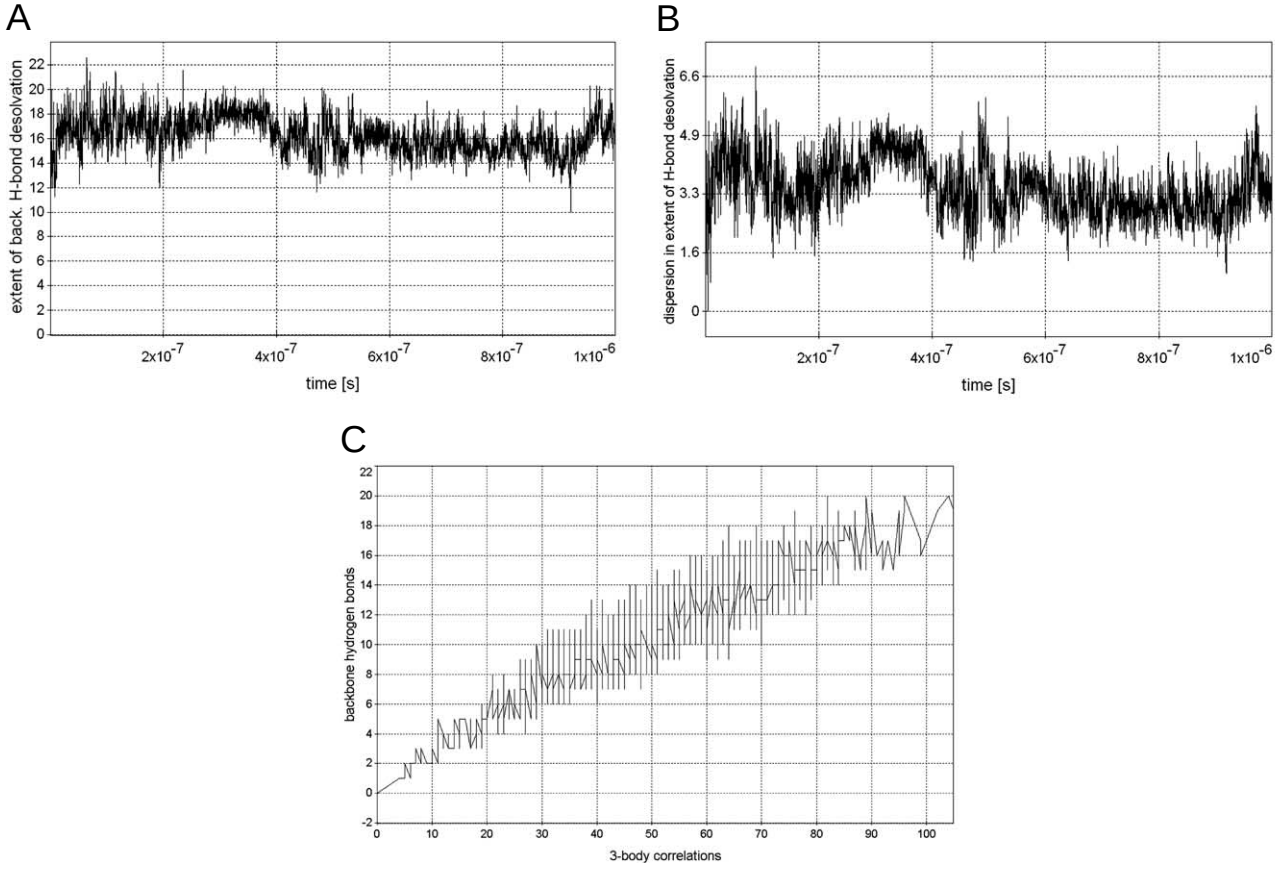

Fig. 3. A: Time-dependent average extent of desolvation of backbone H-bonds measured as number of carbonaceous groups in their desolvation spheres. The average is computed across all H-bonds formed at a given time. B: Time-dependent average dispersion in the extent of H-bond desolvation across all H-bonds formed at a given time. C: Number of backbone hydrogen bonds as a function of the number of three-body correlations. The latter are of the form: (surrounding residue with desolvating group)−(hydrogen bonded pair). Each such correlation signals the presence of a residue with hydrophobic carbonaceous groups contained in the desolvation domain of a hydrogen bond. We can gauge that five residues are typically clustered together to sustain a backbone hydrogen bond.

well-desolvated amide-carbonyl H-bonds. The net free-energy benefit is actually lower, because we must subtract from the over-all 51 kcal/mol the cost of forming the hydrogen bonds in bulk water (~23.5 kcal/mol with our consistent scaling and if we are to follow again Roseman's figures [7], for alternative estimations see below).

The net effect (of the order of −27.5 kcal/mol) reverses the sign of the free-energy change associated with backbone burial and has not been hitherto dissected, rather, it has been subsumed in the H-bonding contribution to the protein over-all stability [19]. Indeed, the nearly complete burial of backbone engaged in secondary structure is thermodynamically very favorable, in good agreement with recent results by Pace [20], while a mere hydrophobic clustering with structureless loops would be highly disfavored (Fig. 2D) and plainly incompatible with the geometric constraints of the backbone (cf. Fig. 2A).

To determine whether backbone H-bonds are thermodynamically significant [21] and provide a guidance to the folding process, our previous conclusions should be complemented by an analysis of the time evolution of desolvation patterns for the backbone H-bonds. Such plots are shown in Fig. 3A–C. The mean extent of desolvation, as measured by the number of vicinal hydrophobic groups per H-bond is nearly constant (~16.0) for the entire trajectory and the typical dispersion across all backbone H-bonds formed at each time is ~3.3. This implies that most backbone H-bonds are highly stabilized [19] by hydrophobic clustering around them. The stability is essentially due to the high cost (~6.2 kcal/mol) of exposing the amides and carbonyls to a highly desolvated environment [7]. Furthermore, the coupling between the hydrophobic clustering around hydrogen bonds and the number of hydrogen bonds formed at any given time is reflected in the linear correlation displayed in Fig. 3C, which introduces a sharp building constraint obeyed statistically across the entire folding timespan.

The coupling between backbone H-bonding and surface burial verified along the entire folding trajectory yields the linear coefficient, 8.3 ± 0.6 dehydrated H-bonds/1000 Å^2 buried, taken to be the slope of the best statistical fit within the confidence bands for the plot given in Fig. 2C. It should be emphasized that this value, stemming from mere counting, arises from purely geometrical analysis and is thus independent of energetic parameterizations. The native-structure parameters, 13 dehydrated backbone H-bonds (Fig. 1C) and 2200 Å exposed surface area, are consistent with the linear correlation and dispersion described above. Furthermore, the confidence bands defined by the structural fluctuations and indicated as boundary curves in Fig. 4, encapsulate the two alternative folding scenarios: H-bonding precedes hydrophobic collapse (I) or H-bonding follows after hydrophobic collapse (II). The actual scenario, as we have seen, can be best fit statistically by the linear correlation.

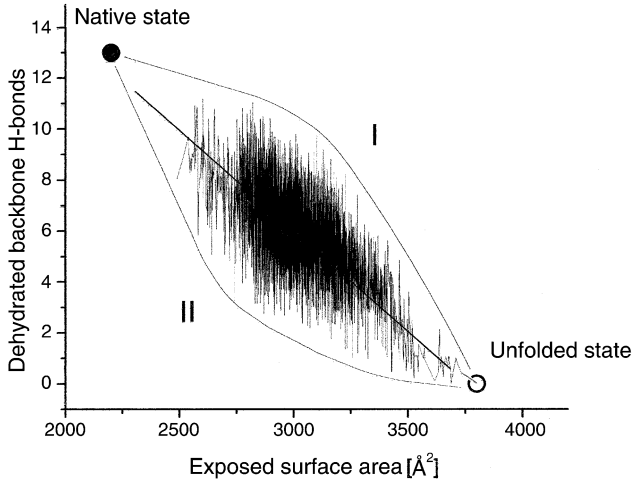

Fig. 4. Schematic view of the extrapolation of the linear correlation between surface burial and number of dehydrated H-bonds all the way up to the native structure (not reached by the Duan–Kollman trajectory). The two confidence bands defined by the boundaries of the structural fluctuations represent the two alternative folding scenarios. These are I: H-bond formation precedes hydrophobic collapse and II: hydrophobic collapse precedes H-bond formation. The real scenario is represented statistically by the linear correlation, and the fluctuations are invariably contained within the two limiting possibilities.

4. Discussion

There are alternative estimations to the one used here of the thermodynamic benefit entailed in burying a backbone amide group when engaged in a H-bond versus burying it unbound. Thus, Ben-Tal et al. [22] reports a lower benefit (~3.6 kcal/mol), in keeping with a more recent one [23]. If we adopt the alternative parameterization of Ben-Tal et al., we would find a free-energy benefit of -34.1 ± 1.2 kcal/mol, instead of our reported value -51.0 ± 1.5 kcal/mol associated with burying the protein backbone with a concurrent formation of amide-carbonyl hydrogen bonds. These numerical discrepancies do not qualitatively alter the statement made in this work that the thermodynamic cost of burying the backbone is to a considerable extent defrayed by forming amide-carbonyl hydrogen bonds.

Furthermore, our conclusions signaling a building constraint in the way backbone hydrogen bonds are made along a folding trajectory stem from purely geometrical considerations, i.e. counting hydrophobes in desolvation domains and computing exposed surface areas. Thus, the coupling of hydrophobic collapse and hydrogen-bond formation can be established independently of energetic parameterizations in the analysis of the simulations or raw PDB data.

Recent work on the energetics of helix formation [23] reveal that peptide H-bonds can prevail in water without the concomitant hydrophobic collapse. Such results apply to polyalanine and other helix-forming peptides excised from proteins and demonstrate that the contribution of desolvated backbone H-bonds per se might be significant enough to warrant the survival of helical structure. These results stressing the importance of desolvation as a stabilizing factor for H-bonds are in keeping with our own findings (cf. Figs. 2E and 3A,B) and do not contradict but rather complement our results.

Summarizing, we may state that hydrophobic clustering and secondary structure formation are necessarily coupled processes in protein folding. This is so because hydrophobic collapse cannot be dissociated from backbone desolvation, which in turn, has a thermodynamic cost that is dramatically compensated upon backbone H-bonding. Furthermore, the desolvation of hydrogen bonds is statistically a very thorough one, implying that such bonds are not adventitious but crucial contributors to the over-all stability of the protein and a guidance to the folding process.

Our results on the actual cost of burying the backbone, reliant as they are on energetic or thermodynamic parameters, might be subject to revision. On the other hand, the building constraints arising from the coupling of hydrophobic clustering to hydrogen-bond formation (Fig. 3) stem from mere geometric analysis, i.e. counting hydrophobes in desolvation domains, and as such is expected to remain impervious to revision.

The analysis put forth in this work is essentially kinetic: It would not be surprising to learn that fully folded stable proteins which bury large surface areas also have more desolvated backbone H-bonds than those which bury smaller areas. The striking feature is that the area buried is commensurate with the total number of backbone hydrogen bonds (Fig. 3) and that the building constraint that defines the commensurability holds all along the folding pathway. This implies that the protein builds its structure in the same way all along starting from the unfolded ensemble.

We may finally ask, what do the few insufficiently desolvated hydrogen bonds in the native structure signal? Recent work reveals that such bonds are determinants of binding sites [24], and thus become properly desolvated upon protein complexation.

Acknowledgements: A.F. thanks Prof. Y. Duan for his authorization to use the Duan-Kollman trajectory for the purpose of study. Enlightening discussions with Profs. Robert Huber, R. Stephen Berry, Tobin, R. Sosnick and Stuart A. Rice are gratefully acknowledged.

References

[1] Fiebig, K.M. and Dill, K.A. (1993) J. Chem. Phys. 98, 3475–3487.
[2] Baldwin, R.L. (2002) Science 295, 1657–1658.
[3] Kataoka, M., Nishii, I., Fujisawa, T., Ueki, T., Tokunaga, F. and Goto, Y. (1995) J. Mol. Biol. 249, 215–228.
[4] Ghosh, A., Elber, R. and Scheraga, H.A. (2002) Proc. Natl. Acad. Sci. USA 99, 10394–10398.
[5] Fernández, A., Sosnick, T.R. and Colubri, A. (2002) J. Mol. Biol. 321, 659–675.
[6] Krantz, B.A., Srivastava, A.K., Nauli, S., Baker, D., Sauer, R.T. and Sosnick, T.R. (2002) Nat. Struct. Biol. 9, 458–463.
[7] Roseman, M.A. (1988) J. Mol. Biol. 201, 621–623.
[8] Duan, Y. and Kollman, P.A. (1998) Science 282, 740–744.
[9] Avbelj, F. and Baldwin, R.L. (2002) Proc. Natl. Acad. Sci. USA 99, 1309–1313.
[10] Dill, K.A., Fiebig, K.M. and Chan, H.S. (1993) Proc. Natl. Acad. Sci. USA 90, 1942–1946.
[11] Akiyama, S., Takahashi, S., Kimura, T., Ishimori, K., Morishima, I., Nishikawa, Y. and Fujisawa, T. (2002) Proc. Natl. Acad. Sci. USA 99, 1329–1334.
[12] Yang, A.S. and Honig, B. (1995) J. Mol. Biol. 252, 351–365.
[13] Makhatadze, G.I. and Privalov, P.L. (1995) Adv. Protein Chem. 47, 307–425.
[14] Fraczkiewicz, R. and Braun, W. (1998) J. Comp. Chem. 19, 319–333.
[15] Ooi, T. (1994) Adv. Biophys. 30, 105–154.
[16] Fernández, A. (2002) Phys. Lett. A 299, 217–220.
[17] Ooi, T., Oobatake, M., Nemethy, G. and Scheraga, H.A. (1987) Proc. Natl. Acad. Sci. USA 84, 3086–3090.
[18] Fernández, A. (2001) J. Chem. Phys. 114, 2489–2502.
[19] Shi, Z., Krantz, B.A., Kallenbach, N. and Sosnick, T.R. (2002) Biochemistry 41, 2120–2129.
[20] Pace, C.N. (2001) Biochemistry 40, 310–313.
[21] Pace, C.N., Shirley, B.A., McNutt, M. and Gajiwala, K. (1996) FASEB J. 10, 75–83.
[22] Ben-Tal, N., Sitkoff, D., Topol, I.A., Yang, A.-S., Burt, S.E. and Honig, B. (1997) J. Phys. Chem. B 101, 450–457.
[23] Avbelj, F., Luo, P. and Baldwin, R.L. (2000) Proc. Natl. Acad. Sci. USA 97, 10786–10791.
[24] Fernández, A. and Scheraga, H.A. (2003) Proc. Natl. Acad. Sci. USA 100, 113–118.