Amino Acid Residues at Protein−Protein Interfaces: Why Is Propensity so Different from Relative Abundance?

Ariel Fernández,*,[†] L. Ridgway Scott,[‡] and Harold A. Scheraga[§]

Institute for Biophysical Dynamics, The University of Chicago, Chicago, Illinois 60637, Department of Computer Science, Department of Mathematics, The University of Chicago, Chicago, Illinois 60637, and Baker Laboratory of Chemistry and Chemical Biology, Cornell University, Ithaca, New York 14853-1301

Received: March 28, 2003; In Final Form: June 23, 2003

It is known that nonpolar residues on the surface of proteins are likely to be at interfaces in protein−protein association. However, as one normalizes the relative abundance at the interface to the overall abundance, one finds that the residues with the highest *propensity* to be at interfaces are as follows (in decreasing order): Asn, Thr, Gly, Ser, Asp, Ala, and Cys. None of these residues is distinctly nonpolar. We show here that these residues also have the highest propensity to be engaged in hydrated backbone hydrogen bonds of the monomeric structure, acting as either proton donors or acceptors. Such preformed hydrogen bonds, in turn, are known to be determinants of protein−protein association as they are stabilized by water removal, upon formation of a complex. A linear correlation is reported between the two independently determined propensities. In addition, the seven residues with the highest propensity for being engaged in hydrated hydrogen bonds all have at most one torsional degree of freedom in their side chain. Thus, upon protein association, the thermodynamically advantageous intermolecular dehydration of preformed hydrogen bonds operates synergistically with a minimization of the entropic loss resulting from the conformational hindrance of the side chains and of contact with nearby nonpolar side chains.

Introduction

The site specificity of protein−protein interactions is poorly understood despite its paramount biological importance.[1−7] It is known that nonpolar residues such as Leu and Val are more abundant at protein−protein interfaces, but it is also true that such residues have a relatively high frequency of occurrence in proteins as a whole. In accordance with the prevalence of nonpolar patches at interfaces, the removal of water surrounding nonpolar residues on the protein surface, leading to hydrophobic interactions, has been taken to be a driving force for protein−protein association.[2,5,8] However, a closer examination of interfaces reveals that nonpolar residues often appear to be mismatched across the interface with polar residues from the binding partner.[9]

In addition, if one computes the ratio of the abundance of residues at the interfaces to the overall abundance of the residues, a different picture emerges: the residues that have the highest tendency to be located at the interface are typically polar. Thus, we are faced with a question: How could the removal of surrounding water upon protein−protein association be favorable for such residues?

Rather than assessing only abundance, our analysis is also aimed at obtaining the *propensity*, f, of each amino acid residue to be at a binding site. The value of f for each amino acid residue is defined as the quotient of the interface abundance over the overall abundance. Abundances and propensities are normalized and reported as percentages. To obtain f, a sample of 12 665 nonredundant PDB entries containing protein−protein complexes was interrogated, and the residues at the interface were recorded, while the overall amino acid composition of the present protein universe was obtained independently, from the Swiss-Prot database.[10] The examination of the PDB was selective in the sense that the protein sequences were obtained from the OWL composite database.[11] These database emphasizes nonredundancy (operationally defined here as less than 66% homology), an important factor in assessing the frequency of a structural pattern. Structural redundancies, on the other hand, were avoided by intersecting our original database with that containing only representative proteins used for protein structure alignment by incremental combinatorial extension of the optimal path.[12]

We find that Leu is prevalent at interfaces to a considerable extent simply because it happens to be the overall most abundant amino acid residue. On the other hand, the residue with the highest f-propensity to be at a protein−protein interface is Asn. The appropriate statistics are shown in Figure 1. The propensity is normalized and reported as a percentage, and the data are reported as differences, $f - \langle f \rangle$, from the mean $\langle f \rangle = 5$ taken over all amino acid residues. The dispersion in f-propensities is $\sigma = 0.58$.

Figure 1 reveals that the residues with a marked propensity ($[f - \langle f \rangle] > \sigma/2$) to be at the interface are as follows (in decreasing order): Asn, Thr, Gly, Ser, Asp, Ala, and Cys. None of these residues can be classified as nonpolar.[13] In fact, the only nonpolar residue that has an appreciable propensity to be at an interface is Val (Figure 1), precisely the one with the lowest number of side-chain rotameric states (Met, on the other hand, has a marginal propensity: $[f - \langle f \rangle] < \sigma/2$). Thus, the loss of rotational freedom for Val upon binding entails the lowest loss in conformational entropy.

* Telephone: 773 702 4908. Fax: 773 702 8487. E-mail: ariel@uchicago.edu.

[†] Institute for Biophysical Dynamics, The University of Chicago.

[‡] Department of Computer Science, Department of Mathematics, The University of Chicago.

[§] Baker Laboratory of Chemistry and Chemical Biology, Cornell University.

10.1021/jp0348124 CCC: $25.00 © 2003 American Chemical Society
Published on Web 08/06/2003

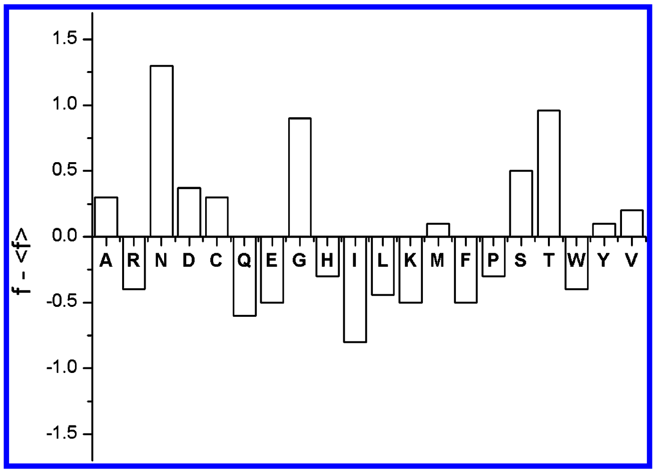

Figure 1. Propensity to be part of protein−protein interfaces for each amino acid type ($\langle f\rangle = 5$). The propensity f, given as a percentage, is computed as the quotient of the abundance of each amino acid residue at protein−protein interfaces divided by its overall abundance.

Figure 1 also shows that, because overexposed nonpolar residues on the protein surface occur infrequently relative to their overall abundance, they do not actually reflect an inherent tendency to be part of binding regions.

We are thus poised to answer the following questions: "How can we explain the trend revealed by Figure 1?" Furthermore, "how do we justify the propensity of polar residues to be at protein−protein interfaces, where their hydration is hindered?"

It was recently reported[7] and experimentally confirmed[14] that the extent of intramolecular dehydration of preformed backbone hydrogen bonds of monomeric molecules is an important feature in determining protein−ligand and, in particular, protein−protein interfaces.[7] The possibility of intermolecular dehydration of side-chain hydrogen bonds upon binding probably also contributes to some degree to define association sites, but the assessment of their dehydration is not as straightforward as that for backbone hydrogen bonds (see Methods). Hydrogen bonds which are insufficiently wrapped by surrounding nonpolar groups can be strengthened (stabilized) upon binding by the removal of surrounding water. The strengthening (stabilization) is a consequence of (a) the decrease in the dielectric constant (leading to a decrease in charge screening), (b) the reduction of hydration of polar groups[7,15−17] upon water removal, and (c) the hydrophobic interactions among the nonpolar groups clustered to protect the hydrogen bond from water attack.

In other words, packing defects in the monomeric structure in the form of underwrapped hydrogen bonds are "adhesive". This effective attraction exerted by the insufficiently dehydrated hydrogen bond on an external nonpolar group from the binding partner yields a stabilization of the former which effectively overcomes the work required to remove residual water dehydrating the hydrogen-bonded polar groups.[7,14]

Methods

To explain the relatively high propensity of certain polar residues to be at binding interfaces, we need to (a) determine the thermodynamic factor that compensates for the loss in hydration of the polar residues upon protein association; and (b) assess the interplay between this factor and the loss in side-chain conformational entropy which results upon binding.

The loss of entropy when the rotational motion of a side chain is restricted upon formation of a hydrogen bond is about 5 eu per *covalent bond*.[8,18] When a nonpolar side chain is involved in a hydrophobic interaction with the nonpolar portion of a side chain of a residue capable of forming a hydrogen bond (see Figure 1 of ref 7), the entropy loss is about 0 to 1 eu per *side chain*.[8,15] Thus, the entropy loss is smaller for shorter side chains.

The protection of hydrogen bonds requires that nonpolar groups (CH_n, $n = 1, 2, 3$) be clustered in the vicinity of intramolecular hydrogen bonds, "wrapping" them as needed for proper exclusion of surrounding water and restricting side-chain rotational freedom.[7,8,15] While the extent of dehydration of backbone hydrogen bonds can be determined analytically (see below), the dehydration of side-chain hydrogen bonds has to be assessed on a case-by-case basis. Here, we focus exclusively on backbone hydrogen bonds.[7,17]

Different ways of determining the extent of dehydration are possible.[7,17] To interrogate vast amounts of structural data, as done in this paper, we are forced to adopt a fast efficient way of assessing the extent to which water is excluded from the microenvironment surrounding the hydrogen bond. Thus, we define a "dehydration domain" associated with the hydrogen bond and count the number of nonpolar groups contained in it. A much more complicated but somewhat more precise description of the microenvironment requires that we compute the sensitivity of the hydrogen bond electrostatics to water removal, a way of assessing the extent of solvent exposure:[19] Approximately 4% of all the hydrogen bonds on the protein surface are *sufficiently* but very *unevenly* wrapped by hydrophobic groups. Thus, the criterion adopted in this paper would count them as well dehydrated, at odds with the computation given in ref 19, which would spot them as being sensitive to water removal. This fact introduces a small source of discrepancy since only about 9% of such bonds lie at known protein−ligand interfaces, thus amounting to 0.36% of the total number of surface hydrogen bonds interrogated.

A dehydration domain for a backbone hydrogen bond may be defined as the union of two intersecting spheres, with a diameter of 6.4 Å, centered at the α-carbons of the residues paired by the hydrogen bond. The number, ρ, of side-chain nonpolar groups within such domains averaged over all the backbone hydrogen bonds is maximal ($\rho = 20.8$) in the least interactive proteins such as myoglobin, and minimal in highly interactive ones, such as lysozyme ($\rho = 11.7$), with the most deficient wrappers of their hydrogen bonds being certain toxins and the prion proteins in their cellular conformation.[7,14] Intramolecularly underwrapped hydrogen bonds (UWHBs) are typically defined by the inequality $\rho < 11$ and become significantly stabilized by removal of water. Thus, they typically signal binding sites in soluble proteins.[7] The statistics of hydrogen-bond dehydration vary, of course, with the choice of dehydration radius; however, the results are robust within the range 6.0−7.5 Å in the sense that the same UWHBs are invariably singled out.

Results

In Figure 2, we report the statistics on the propensity, f_d, of different residues to form underwrapped backbone hydrogen bonds. This propensity, attributed to each amino acid residue, represents the percentage of UWHBs from all PDB monomers which involve the particular amino acid residue as a proton donor or acceptor divided by the overall abundance of that amino acid residue. To construct Figure 2, all nonredundant PDB entries representing monomeric proteins were interrogated, as well as those containing complexes. In the latter case, only the intramolecular wrapping or dehydration of backbone hydrogen bonds in the monomers was computed (see Methods).

The residues which are most likely engaged in an UWHB ($[f_d - \langle f_d\rangle] > \sigma/2$, with $\sigma = 0.61$, $\langle f_d\rangle = 5$) are (in decreasing

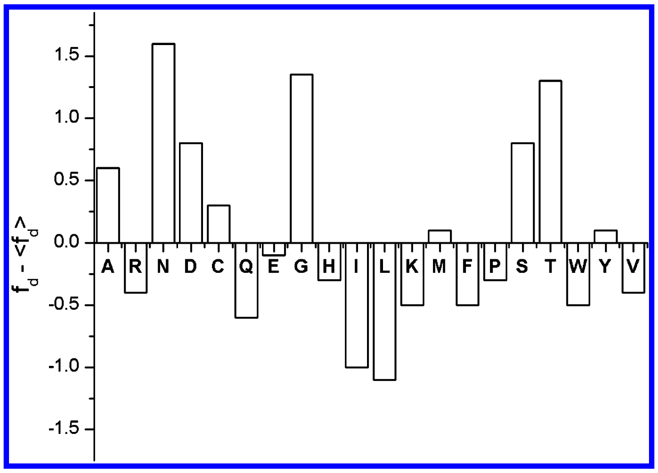

Figure 2. Propensity to be involved in an underwrapped hydrogen bond as proton donor or acceptor for each amino acid type ($\langle f_d \rangle = 5$). The propensity, f_d, given as a percentage, is computed as the quotient of the frequency with which the particular amino acid residue is involved in UWHBs divided by its overall abundance.

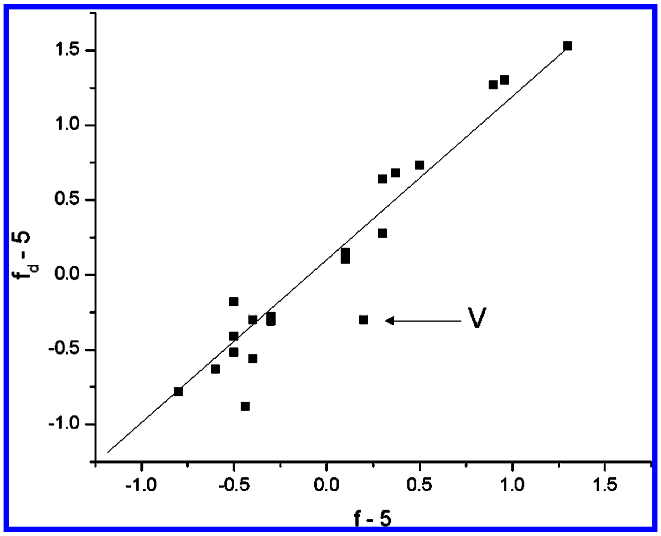

Figure 3. Propensity to be engaged in an UWHB, f_d, plotted against the propensity to be at a protein−protein interface, f. The largest distance from the best-fit line is encountered by Val, which also shows opposing propensities, as revealed by comparing Figures 1 and 2.

order): Asn, Gly, Thr, Ser, Asp, Ala and Cys, that is, *the same amino acid residues with the highest propensity to be at a protein−protein interface*, with one transposition (Gly ↔ Thr) in the order.

Strikingly, a clear-cut linear correlation exists between f_d, the propensity for wrapping, and f, the propensity for being part of a binding interface (Figure 3). The single most conspicuous outlier is Val (cf. Figures 1−3): its hydrophobicity indicates that it is a good wrapper of its backbone hydrogen bonds, while its side-chain rigidity makes it prone to be at protein−protein interfaces since it can minimize the entropic cost of conformational hindrance upon binding.

To rationalize the results shown in Figure 2, we note that the residues with a minimal distance of their polar groups from the backbone (Asn, Asp, Thr, Ser, and Cys), with the entropy loss upon hydrogen bonding indicated above, and with minimal contact of their nonpolar positions with nearby nonpolar side chains, are likely to be engaged in UWHBs, not only because they themselves have fewer nonpolar groups but also because the relative proximity of their polar groups to a backbone hydrogen bond prevents further clustering of nonpolar groups around the bond.

This hindrance of hydrophobic clustering around the backbone hydrogen bond is extreme for Asn and Asp because each has two polar groups closest to the backbone. However, Asp often carries a net charge and is thus comparatively less prone to be buried at interfaces, as revealed by contrasting Figures 1 and 2 (it is most likely to be buried only if it can form a stable salt bridge upon protein−protein association). It is observed that while Asp has a high propensity to be engaged in a UWHB, it has a relatively lower propensity to be at an interface**.** On the other hand, Ala and Gly are themselves underwrappers, and thus, backbone hydrogen bonds in which they are involved require a cooperative organization of the chain for proper wrapping with substantial nonlocal contributions to hydrogen-bond desolvation.

There is a synergistic effect which reinforces the propensity of the seven most common amino acids residues involved in UWHBs to be part of protein−protein interfaces: these residues have side chains with at most one torsional degree of freedom. This implies that the entropic cost associated with their conformational hindrance upon binding[8,15,18] is minimized. Thus, the intermolecular desolvation of the UWHBs concurrent with protein association takes place at a minimal entropic expense.

Conclusions

The exclusion of water from exposed nonpolar residues has been assumed to be a major factor driving protein−protein association. However, polar−nonpolar mismatches are frequent at interfaces.[9] Furthermore, while exposed nonpolar patches are relatively abundant at interfaces, the actual propensities to be at interfaces, determined by normalizing to the over-all abundances, indicate a different trend: certain *polar* residues such as Asn and Thr actually have the highest propensity to be at interfaces. This trend is reported in this work and poses another problem: what factor may be acting in protein−protein association that compensates for the hydration hindrance of polar residues such as Asn or Thr which takes place as the monomers bind to each other and remove water from the interface?

Here, we showed that the residues with the highest propensity to be at the interface are exactly those with the highest propensity to be engaged in preformed backbone hydrogen bonds which are insufficiently dehydrated or underwrapped within the monomeric structure. Such bonds become strengthened and stabilized as intermolecular dehydration takes place upon protein−protein association.[7,14] This is the thermodynamic factor that compensates for the hydration hindrance of the polar residues which takes place upon protein−protein association, but it is necessary that the polar residues be engaged in a preformed underwrapped hydrogen bond.[7,14]

Once this factor is properly incorporated, the discrepancy between abundance and propensity to be at the binding interface may be accounted for: the residues with the highest propensity to be at protein−protein interfaces are not the most abundant at the interface; instead, they are the ones which have the highest propensity to be engaged in preformed hydrogen bonds which are underwrapped or underdesolvated intramolecularly.

Acknowledgment. This work was supported by the National Science Foundation, Grant MCB00-03722.

References and Notes

(1) Jones, S.; Thornton, J. M. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 13−20.

(2) Jones, S.; Thornton, J. M. *J. Mol. Biol.* **1997**, *272*, 133−143.

(3) DeLano, W. L.; Ultsch, M. H.; de Vos, A. M.; Wells, J. A. *Science* **2000**, *287*, 1279−1283.

(4) Hu, Z.; Ma, B.; Wolfson, H.; Nussinov, R. *Proteins: Struct. Funct. Gen.* **2000**, *39*, 331−342.

(5) Glaser, F.; Steinberg, D. M.; Vakser, I. A.; Ben-Tal, N. *Proteins: Struct. Funct. Gen.* **2001**, *43*, 89−102.

(6) Smith, G. R.; Sternberg, M. J. *Curr. Opin. Struct. Biol.* **2002**, *12*, 28−35.

(7) Fernández, A.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 113−118.

(8) Némethy, G.; Scheraga, H. A. *J. Phys. Chem.* **1962**, *66*, 1773−1789; Erratum *J. Phys. Chem.* **1963**, *67*, 2888.

(9) Sondermann, P.; Huber, R.; Oosthuizen, V.; Jacob, U. *Nature* **2000**, *406*, 267−273.

(10) Boeckmann, B.; Bairoch, A.; Apweiler, R.; Blatter, M.-C.; Estreicher, A.; Gasteiger, E.; Martin, M. J.; Michoud, K.; O'Donovan, C.; Phan, I.; Pilbout, S.; Schneider, M. *Nucleic Acids Res.* **2003**, *31*, 365−370.

(11) Bleasby, A. J.; Akrigg, D.; Attwood, T. K. *Nucleic Acids Res.* **1994**, *22*, 3574−3577.

(12) Shindyalov, I. N.; Bourne, P. E. *Protein Eng.* **1998**, *11*, 739−747.

(13) Bahar, I.; Jernigan, R. L. *J. Mol. Biol.* **1997**, *266*, 195−214.

(14) Fernández, A.; Berry, R. S. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 2391−2396.

(15) Némethy, G.; Steinberg, I. Z.; Scheraga, H. A. *Biopolymers* **1963**, *1*, 43−69.

(16) Makhatadze, G. I.; Privalov, P. L. *Adv. Protein Chem.* **1995**, *47*, 307−425.

(17) Fernández, A.; Sosnick, T. R.; Colubri *J. Mol. Biol.* **2002**, *321*, 659−675.

(18) Laskowski, M., Jr.; Scheraga, H. A. *J. Am. Chem. Soc.* **1954**, *76*, 6305−6319.

(19) Fernández, A.; Scott, L. R. *Biophys. J.* **2003**,, in press.