GXXXG and AXXXA: Common α-Helical Interaction Motifs in Proteins, Particularly in Extremophiles†

Gary Kleiger, Robert Grothe, Parag Mallick, and David Eisenberg*

Howard Hughes Medical Institute, UCLA-DOE Laboratory of Structural Biology and Molecular Medicine, Molecular Biology Institute, University of California at Los Angeles, P.O. Box 951570, Los Angeles, California 90095-1570

Received January 24, 2002; Revised Manuscript Received March 20, 2002

ABSTRACT: The GXXXG motif is a frequently occurring sequence of residues that is known to favor helix−helix interactions in membrane proteins. Here we show that the GXXXG motif is also prevalent in soluble proteins whose structures have been determined. Some 152 proteins from a nonredundant PDB set contain at least one α-helix with the GXXXG motif, 41 ± 9% more than expected if glycine residues were uniformly distributed in those α-helices. More than 50% of the GXXXG-containing α-helices participate in helix−helix interactions. In fact, 26 of those helix−helix interactions are structurally similar to the helix−helix interaction of the glycophorin A dimer, where two transmembrane helices associate to form a dimer stabilized by the GXXXG motif. As for the glycophorin A structure, we find backbone-to-backbone atomic contacts of the Cα−H···O type in each of these 26 helix−helix interactions that display the stereochemical hallmarks of hydrogen bond formation. These glycophorin A-like helix−helix interactions are enriched in the general set of helix−helix interactions containing the GXXXG motif, suggesting that the inferred Cα−H···O hydrogen bonds stabilize the helix−helix interactions. In addition to the GXXXG motif, some 808 proteins from the nonredundant PDB set contain at least one α-helix with the AXXXA motif (30 ± 3% greater than expected). Both the GXXXG and AXXXA motifs occur frequently in predicted α-helices from 24 fully sequenced genomes. Occurrence of the AXXXA motif is enhanced to a greater extent in thermophiles than in mesophiles, suggesting that helical interaction based on the AXXXA motif may be a common mechanism of thermostability in protein structures. We conclude that the GXXXG sequence motif stabilizes helix−helix interactions in proteins, and that the AXXXA sequence motif also stabilizes the folded state of proteins.

Common amino acid sequence motifs in proteins offer windows into protein function (*1*). One well-known example is that of Cys_2His_2 zinc fingers, one of the most common eukaryotic DNA-binding motifs; these are detected by the consensus sequence (F/Y)-X-C-X_{2-5}-C-X_3-(F/Y)-X_5-φ-X_2-H-X_{3-5}-H, where X represents any amino acid and φ is a hydrophobic one (*2*). In the three-dimensional structure of this motif, the two cysteine and two histidine residues coordinate a zinc ion, stabilizing the structure for binding.

The amino acid sequence motif GXXXG stabilizes helix−helix interactions in membrane proteins (*3*). In fact, GXXXG is the most over-represented sequence motif (considering all possible motifs of the form ZX^nZ, where Z is any amino acid) in a database of transmembrane helices, occurring nearly 32% above expectation (*4*). GXXXG is also the dominant motif found in an in vivo transmembrane oligomerization selection system (*5*).

The three-dimensional structure of the transmembrane domain of glycophorin A provides insight into how the glycyl residues of the GXXXG motif stabilize helix−helix interactions (*6*). Here two single-transmembrane helices containing the GXXXG motif associate to form a symmetric, right-handed homodimer. All four glycines (two from each helix) can be found at the helix−helix interface. The four-residue separation of glycyl residues in the GXXXG motif serves to align them on one face of the helix, providing a flat platform for the two glycyl residues of the other helix, and permits the backbones to bind closely. In this close configuration, the two glycyl Cα atoms are able to donate their hydrogen atoms to form hydrogen bonds with carbonyl oxygen atoms on the adjacent helix, stabilizing the glycophorin A dimer (*7*). The occurrence of four glycyl residues at the helix−helix interface also minimizes the loss of side chain entropy upon binding, stabilizing the bound state of the dimer (*8*).

Recently, we discovered a helix−helix interaction at the intermolecular interface of the E1α and E1β subunits of pyruvate dehydrogenase which is stabilized by two glycyl residues from the E1β subunit that form a GXXXG motif (*9*). Both glycines are conserved in all known E1β protein sequences. The two glycyl residues stabilize the intermolecular interface in a manner analogous to that of glycophorin A. While both helices of the helix−helix interaction of glycophorin A contain the GXXXG motif, only the α-helix from the E1β subunit contains the GXXXG motif in the pyruvate dehydrogenase example, demonstrating that only one helix of the helix−helix interaction needs to contain the GXXXG motif to promote association in a glycophorin

† The work of G.K. was funded in part by U.S. Public Health Service Training Grant GM07185. This work was supported by the National Institutes of Health and BER of the Department of Energy.

* To whom correspondence should be addressed. Phone: (310) 825-3754. Fax: (310) 206-3914. E-mail: david@mbi.ucla.edu.

10.1021/bi0200763 CCC: $22.00
Published on Web 04/20/2002

A-like manner. The E1β example is the first helix−helix interaction stabilized by the GXXXG motif in a soluble protein and prompted us to ask how many known protein structures contain an α-helix with the GXXXG motif and whether the GXXXG motif functions in stability. Here we document that 152 proteins in the nonredundant Protein Data Bank (1611 proteins total) contain at least one helix with the GXXXG motif, and suggest that the function of both the GXXXG motif and the related AXXXA motif is to stabilize both intermolecular interactions between protein subunits and intramolecular interactions within the same protein.

MATERIALS AND METHODS

Databases. To account for the over-representation of certain protein families in the PDB,[1] we used a nonredundant PDB set (www.fccc.edu/research/-labs/dunbrack/culledpdb.html) (*10*). In this set, only X-ray crystal structures of >2.5 Å resolution are included. In addition, no two sequences in the nonredundant set are greater than 30% identical to each other. The culled PDB (nonredundant PDB) set contains 1731 chains corresponding to 1611 proteins.

Counting the Occurrences of Amino Acid Sequence Motifs. We counted the occurrences of amino acid sequence motifs in α-helices. Starting with the atomic coordinates for each chain in our nonredundant PDB set, we assigned secondary structure to each residue using the program DSSP (*11*). Secondary structure elements, defined as a string of at least five consecutive residues with identical secondary structure, were extracted. Secondary structure elements longer than 25 residues were discarded because it was computationally intractable to permute such long sequences. Duplicate copies of sequences were discarded. The final number of α-helices from the nonredundant PDB is 12 802, and the final number of β-strands is 10 249. Occurrences of all 20 symmetrical amino acid sequence motifs were then calculated. Note that two motifs may overlap, yet both are counted. For example, in the following sequence, GXGXGXG, where X represents any amino acid other than glycine, the GXXXG motif occurs twice.

Calculating the Expected Value and Variance for Each Amino Acid Sequence Motif. The method used to calculate the expected value and variance for each amino acid sequence motif has been described elsewhere (*4*). Briefly, the expectation and variance for any symmetrical amino acid sequence motif (e.g., GXXXG and AXXXA) were calculated for each α-helix in our database. For a helix of known length and composition, we used bit strings to represent all possible sequences (e.g., 1's for glycines, 0's for non-glycines). If it is assumed that residues are uniformly distributed in the helix, the probability for the occurrence of a given number of motifs is computed by counting the occurrences of the motif in each enumerated sequence. The expected number of motifs is then given by

$$\langle X \rangle = \sum_x xp(x) \tag{1}$$

where X is a random variable corresponding to the number of occurrences of the amino acid sequence motif in a given helix and $p(x)$ is the probability of observing x occurrences. Similarly, the variance is given by

$$\mathrm{Var}(x) = \langle X^2 \rangle - \langle X \rangle^2 \tag{2}$$

The expectation and variance of each α-helix were then summed over all helices to calculate the total expectation and variance for a particular amino acid sequence motif for the entire database. Note that this method is applicable to motifs of any given spacing, n.

We calculated two quantities which compare the number of observed and expected motifs: the enrichment of observed over expected occurrences and the Z score. The enrichment is

$$\frac{X_{\mathrm{obs}} - \langle X \rangle}{\langle X \rangle} \tag{3}$$

where X_{obs} is the observed occurrences of the motif. The standard deviation of the enrichment is

$$\frac{1}{\sqrt{\langle X \rangle}} \tag{4}$$

When the expected number of motifs is small, the enrichment may exhibit large fluctuations. For a small sample, it may be the case that the observed enrichment is merely a statistical fluctuation. The Z score assesses how likely this is. The Z score is given by

$$Z = \frac{X_{\mathrm{obs}} - \langle X \rangle}{\sigma} \tag{5}$$

where σ is the standard deviation.

Structural Comparison of Helix−Helix Interactions, Identifying Atomic Contacts of the Cα−H···O Type within Those Interactions, and Calculating the Interhelical Axial Distances between Interacting Helices. The program ALIGN, version 2 (*12*), was used to superimpose the structure of one helix−helix interaction onto the other. Two helices were considered to be interacting if six or more residues were in contact. Using the definition of Chothia et al. (*13*), a contact is defined as two atoms within 0.6 Å of the sum of their van der Waals radii. Helix−helix interactions containing either the GXXXG motif or the AXXXA motif were identified using custom PERL software. Histograms were generated using custom PERL software and plotted using the program XMGR. The standard deviation of the frequency for any bin i in the histogram, σ_{f_i}, is given by

$$\sigma_{f_i} = \sqrt{\frac{f_i^{\mathrm{obs}}}{N}} \tag{6}$$

where f_i^{obs} is the observed frequency for the ith bin and N is the total number of observations.

The coordinates for hydrogen atoms were calculated using the CCP4 program HGEN, which uses standard geometry and a bond length of 1.0 Å. The structures of helix−helix interactions containing the GXXXG motif were analyzed and searched for atomic contacts of the Cα−H···O type using the program O. Interhelical axial distances and packing angles were calculated using the CCP4 program HELIXANG.

[1] Abbreviations: rms, root-mean-square; PDB, Protein Data Bank.

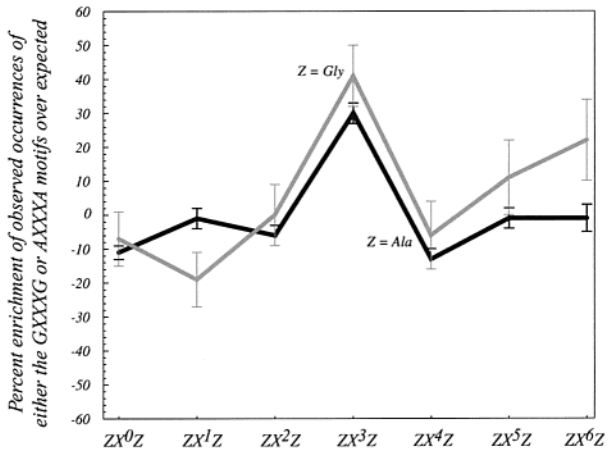

FIGURE 1: Percent enrichment of observed over expected occurrences of two symmetrical amino acid sequence motifs in α-helices of the form ZX^nZ, where Z is either glycine (gray line) or alanine (black line). Notice the peaks at $n = 3$, which aligns both residues of the motif on the same side of the helix. This spacing of the residues creates a flat interaction surface and is consistent with the observation that the GXXXG and AXXXA sequences function as helical interaction motifs.

Calculating the Enrichment of Observed over Expected Occurrences for both the GXXXG and AXXXA Motifs for All ORFs in a Genome. The putative open reading frames from 24 fully sequenced genomes were downloaded from the entrez genome search and retrieval system of the National Center for Biotechnology Information (NCBI, www.ncbi.nlm.nih.gov). Secondary structure was predicted for each open reading frame using the program PHD (*14*). Secondary structure elements were extracted in a manner identical to that with the nonredundant PDB set. The program SEG was used to find and remove low-complexity amino acid repeats (*15*). *Z* scores and the enrichment of observed occurrences over expected were calculated in a manner identical to that with the PDB set.

RESULTS

We counted 172 occurrences of the GXXXG motif in all α-helices from the nonredundant PDB, and discarded 11 of these as being in low-complexity repeats. Some 152 proteins from the nonredundant PDB contain at least one helix with the GXXXG motif, or nearly one in 10 proteins.

To estimate the significance of the number of occurrences of GXXXG, we also calculated how many occurrences of the GXXXG motif we expect at random. Specifically, we ask the following question. If glycyl residues were distributed uniformly in α-helices of the observed lengths, what fraction of these helices would contain the GXXXG motif? These values can be calculated using the TMSTAT method, in which all possible permutations of glycyl residues in each helix are enumerated and searched for the GXXXG motif (*4*).

Occurrences of the GXXXG motif in α-helices from the nonredundant PDB are enriched, occurring 41 ± 9% more often than expected if glycyl residues were uniformly distributed in α-helices (Figure 1 and Table 1). One would expect only 114 occurrences rather than 161 of the GXXXG motif in the 12 802 α-helices from the nonredundant PDB, given the length and number of glycines in each helix.

Table 1: Statistics for 14 Symmetrical Amino Acid Sequence Motifs, Showing That GXXXG and AXXXA Are the Only Motifs from This Group that Are Significantly Enriched[a]

amino acid sequence motif	no. of observed occurrences	no. of expected occurrences	enrichment (%)	σ	*Z* score
GG	149	161	−7 ± 8	11.2	−1.0
GXG	117	145	−19 ± 8	10.8	−2.6
GXXG	130	130	0 ± 9	10.3	0
GXXXG	161	114	**41 ± 9**	9.8	**4.8**
GXXXXG	93	99	−6 ± 10	9.2	−0.6
GXXXXXG	94	84	11 ± 11	8.5	1.1
GXXXXXXG	88	72	22 ± 12	7.9	2.0
AA	1512	1700	−11 ± 2	33.4	−5.6
AXA	1506	1526	−1 ± 3	32.4	−0.6
AXXA	1271	1352	−6 ± 3	30.9	−2.6
AXXXA	1533	1179	**30 ± 3**	29.3	**12.1**
AXXXXA	876	1005	−13 ± 3	27.4	−4.7
AXXXXXA	835	847	−1 ± 3	25.4	−0.5
AXXXXXXA	700	705	−1 ± 4	23.3	−0.2

[a] The number of observed occurrences for each motif in 12 802 α-helices from the nonredundant PDB is given. The expected number of occurrences is calculated assuming uniformly distributed residues in the same α-helices. The percent enrichment is the number of observed occurrences for a given motif minus the number of expected occurrences, divided by the number of expected occurrences. The standard deviation and *Z* score are also given (see Materials and Methods). Notice the unusual values for the enrichment and *Z* score for GXXXG and AXXXA (in bold).

While the 161 occurrences of the GXXXG motif in α-helices are significant, no other spacing for symmetrical sequence motifs containing glycyl residues shows similar enrichment. For example, the GXXG motif occurs 130 times in the 12 802 α-helices from the nonredundant PDB, which is equal to the expected number of occurrences if glycyl residues were uniformly distributed. The observed and expected occurrences for symmetrical sequence motifs containing glycyl residues, corresponding to a range of spacings from zero residues between glycyl residues (GX^0G) to six residues (GX^6G), are given in Table 1. Therefore, the GXXXG motif in α-helices is the only symmetrical amino acid sequence motif with glycyl residues that shows significant enrichment over expected occurrences.

The observed occurrences of the AXXXA motif in α-helices from the nonredundant PDB are also enriched over expectation (Figure 1 and Table 1). The AXXXA motif occurs 1533 times in helices from the nonredundant PDB. Some 808 proteins from the nonredundant PDB contain at least one helix with the AXXXA motif, corresponding to more than half of the proteins in the nonredundant PDB. However, only 1179 occurrences of the AXXXA motif are expected given the length and composition of each helix in the database, corresponding to an enrichment of 30 ± 3%. Similar to the GXXXG motif, the three-residue spacing between alanine residues in the AXXXA motif is the only spacing for symmetrical sequence motifs containing alanine residues with significant enrichment of observed over expected occurrences (Table 1).

There are other symmetrical amino acid sequence motifs that show enrichment of observed occurrences on the same side of amphipathic helices (e.g., LXXL and LXXXL), but the spacing dependence for these other motifs is less stringent than for the GX^nG and AX^nA motifs. This suggests that the roles of the GXXXG and AXXXA motifs are not simply to

Table 2: A List of the 26 Proteins from the Nonredundant PDB That Contain both a Helix with the GXXXG Motif and a Helix−Helix Interaction Similar to the Helix−Helix Interaction of Glycophorin A[a]

PDB entry	protein	rmsd from GpA	packing angle (deg)	axial distance (Å)
1afo	glycophorin A	−	−33	7.0
1hjr	resolvase	1.1	−53	6.9
1luc	luciferase	0.4	−49	6.3
1qs0	oxidoreductase	1.0	−58	6.7
1f8y	transferase	0.8	−35	6.3
1b0p	oxidoreductase	0.8	−45	6.8
1b24	endonuclease	0.6	−31	6.5
1dfa	hydrolase PI-SCEI	0.8	−32	6.6
1dq3	hydrolase PI-PFUI	0.7	−36	6.5
1eex	diol dehydratase	1.1	−36	6.8
1gak	cell adhesion sp18	1.2	−49	7.0
1zpd	pyrrole decarboxylase	0.7	−48	6.3
1qmg	oxidoreductase	1.2	−40	5.9
1d0v	transferase	1.2	−51	6.9
1d3v	arginase	1.1	−35	6.8
1evy	glycerol-1,3-phosphate dehydrogenase	0.9	−42	5.7
1fx8	glycerol facilitator	0.8	−45	6.8
1e39	flavocytochrome	0.8	−32	7.0
1qla	fumarate reductase	1.2	−42	6.7
1dbt	orotidine decarboxylase	0.8	−34	6.5
1e3a	penicillin amidase	0.7	−40	6.1
1vns	haloperoxidase	1.2	−36	6.4
8abp	L-arabinose binding protein	1.0	−39	6.2
1gca	galactose binding protein	1.2	−38	6.4
2dri	D-ribose binding protein	1.3	−37	6.3
1bfd	benzoylformate decarboxylase	0.6	−45	6.4
1dbq	DNA binding protein	1.2	−37	5.7

[a] The backbone atoms from each helix−helix interaction were superimposed onto the backbone atoms of the GpA dimer; the rms distance for aligned main chain atoms (rmsd from GpA) is given. The interhelical packing angle and axial distance are given for each helix−helix interaction.

provide helical amphipathicity. To gain insight into the function of the GXXXG and AXXXA motifs, we analyzed the structures of those proteins containing the GXXXG or AXXXA motifs from the nonredundant PDB.

Structural Analysis of Proteins Containing α-Helices with the GXXXG Motif: Glycophorin A-like Helix−Helix Interactions. Some 97 of the 172 α-helices containing the GXXXG motif form at least one helix−helix interaction in their respective proteins (56%). Of these helix−helix interactions, 26 are similar in structure to the helix−helix interaction of glycophorin A in that they share the following structural features. (1) Each of the 26 helix−helix interactions structurally superimposes with the helix−helix interaction of glycophorin A such that the aligned main chain atoms deviate by no greater than 1.2 Å (rms deviation). This means that the helix−helix crossing is of the right-handed type (*7*). (2) At least one of the potential backbone-to-backbone Cα−H···O type hydrogen bonds observed in the glycophorin A dimer is also observed in the helix−helix interaction. It is notable that 23 of these 26 examples of helix−helix interactions contain only one GXXXG helix, consistent with our finding for E1α and E1*β* that one helix containing the GXXXG motif is sufficient to promote proximity of the helices and a Cα−H···O contact (*9*).

As an example of a glycophorin-like helix−helix interaction, consider resolvase (1hjr), which forms a homodimer in which two symmetry-related helices containing the GXXXG motif interact across the intermolecular interface (Figure 2a) in a right-handed fashion. In fact, the helix−helix interactions of resolvase and the glycophorin A dimer are nearly identical (Figure 2b). The structures superimpose with an rms deviation of 1.1 Å for 124 main chain atoms. In addition, all four glycyl residues forming the GXXXG motif in the glycophorin A dimer are structurally equivalent to the glycyl residues forming the GXXXG motif in resolvase. Finally, in the helix−helix interaction of resolvase, three backbone-to-backbone atomic contacts of the Cα−H···O type have geometries consistent with hydrogen bond formation (Figure 2b). Together, all of the 26 helix−helix interactions that contain the GXXXG motif and are structurally similar to the glycophorin A dimer contain 102 backbone-to-backbone contacts of the Cα−H···O type that have geometries consistent with interhelical hydrogen bond formation (see the Supporting Information).

The 26 helix−helix interactions that are structurally similar to glycophorin A are unusual in that the interhelical axial distance in each helix−helix interaction is short (Figure 3b and Table 2). We therefore ask if the distribution of interhelical distances for helix−helix interactions containing the GXXXG motif is distinct from the general distribution of interhelical distances. This probability can be measured using the χ^2 test (*16*). The distributions of interhelical axial distances for both helix−helix interactions containing the GXXXG motif (black bars) and interactions not containing the motif (white bars) are shown in Figure 3a. Comparing these two distributions gives a χ^2 value of 32.1 (13 degrees of freedom), corresponding to a probability of <0.01 that both sets of interactions came from the same population.

In addition to being distinct from the general set of helix−helix interactions, the distribution of interactions containing the GXXXG motif is enriched with short interhelical distances over helix−helix interactions not containing the GXXXG motif. For example, the frequency of helix−helix interactions containing the GXXXG motif with interhelical distances in the range of 6.5−7.0 Å is 13.5%; the frequency is only 4.9% for helix−helix interactions not containing the GXXXG motif in the same range (Z score = 8.6).

Interhelical Axial Distances for Helix−Helix Interactions Containing the AXXXA Motif. Unlike the distribution of the GXXXG motif, the distribution of interhelical distances for helix−helix interactions containing the AXXXA motif is not distinct from the general distribution of interhelical distances. The histograms of interhelical distances for both helix−helix interactions containing the AXXXA motif (black bars) and not containing the AXXXA motif (white bars) are shown in Figure 3c. Comparing these two distributions gives a χ^2 value of 21 (16 degrees of freedom), corresponding to a probability of <0.2 that both sets of interactions came from the same population. In addition, only a modest enrichment of helix−helix interactions with interhelical distances in the range of 7.5−8.0 Å is observed for interactions containing the AXXXA motif (Z score = 4.5).

Calculating the Enrichment of Observed Occurrences of the GXXXG and AXXXA Motifs over Expected Occurrences for Fully Sequenced Genomes. Observed occurrences for both the GXXXG and AXXXA motifs are enriched over expected occurrences from the predicted α-helices of 24 fully sequenced genomes (Figure 4). For example, all proteins from the thermophilic organism *Thermatoga maritima* are pre-

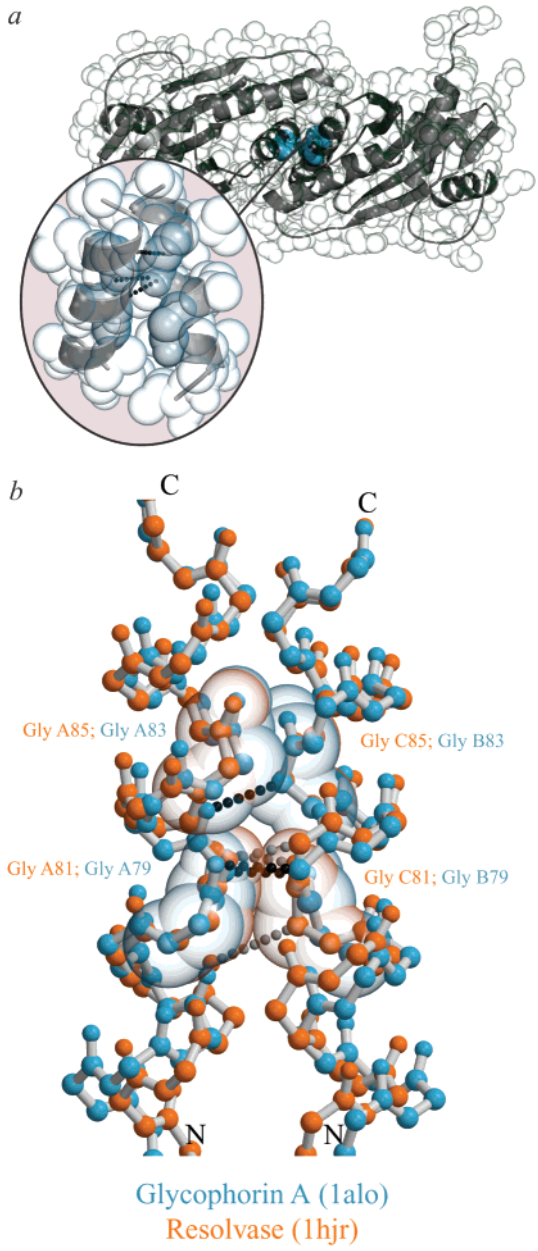

FIGURE 2: (a) Space-filling model of the homodimeric Holliday-junction protein resolvase (*25*) (1hjr). A helix with the GXXXG motif interacts with its symmetry mate at the intermolecular interface. The glycyl residues of the GXXXG motif are blue. Closer inspection of the helix−helix interaction shows that the glycyl residues of the GXXXG motif on one helix interact against the glycycl residues of the GXXXG motif on the adjacent helix. The interaction of these residues allows close association of the two helices and promotes backbone-to-backbone contacts between Cα−H···O atoms that have geometries consistent with hydrogen bond formation. Potential hydrogen bonds are shown as dotted lines. (b) Structural superposition of the helix−helix interaction of resolvase (orange) and the helix−helix interaction of the glycophorin A dimer (blue). The four glycyl residues of the GXXXG motifs are shown as space-filled spheres (blue for glycophorin A and orange for resolvase). Notice that the glycyl residues of the GXXXG motif of glycophorin A and resolvase are structurally equivalent. Black dotted lines highlight atomic contacts in the glycophorin A dimer with appropriate geometries for backbone-to-backbone hydrogen bonds of the Cα−H···O type. Gray dotted lines denote similar contacts in the resolvase helix−helix interaction. The helix−helix interactions are nearly identical in structure, demonstrating that the GXXXG motif can stabilize helix−helix interactions in both membrane proteins and soluble proteins.

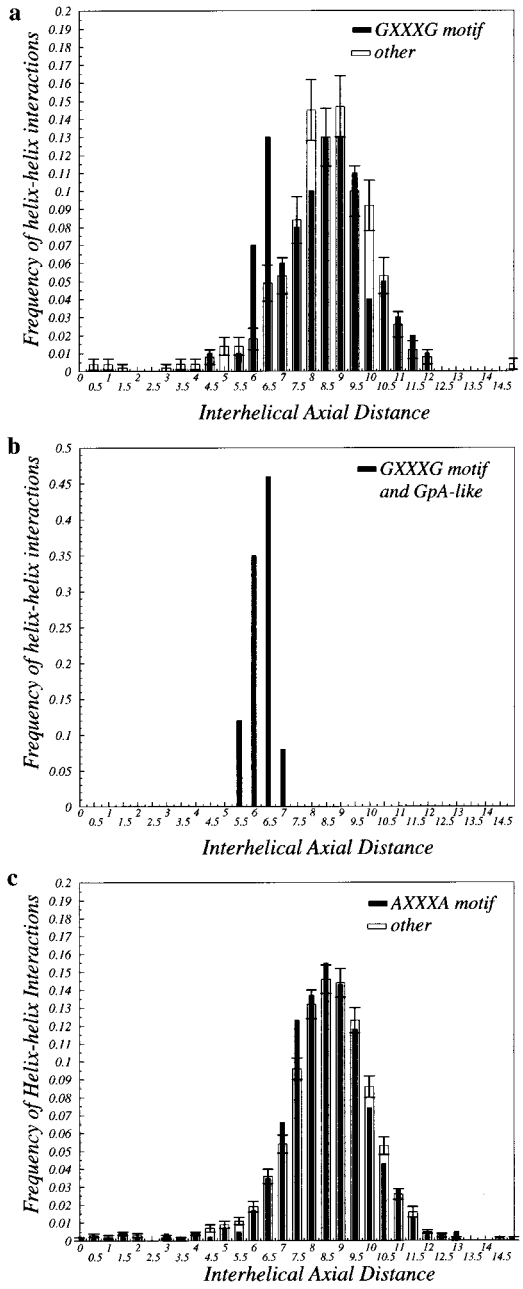

FIGURE 3: Histogram of interhelical axial distances. (a) Some 171 helix−helix interactions from the nonredundant PDB, where at least one helix in the interaction contains the GXXXG motif (black bars, mean interhelical axial distance of 8.8 ± 3 Å). The histogram of interhelical axial distances for helix−helix interactions that do not contain the GXXXG motif (extracted from the same proteins) is also shown (white bars, mean interhelical axial distance of 9.0 ± 3.2 Å). Notice that the helix−helix interactions containing the GXXXG motif are enriched with interactions of interhelical axial distances between 6.0 and 7.0 Å as compared to helix−helix interactions that do not contain the GXXXG motif. (b) A subset of the 171 helix−helix interactions containing the GXXXG motif, consisting of 26 helix−helix interactions, where at least one helix in the interaction contains the GXXXG motif, and the helix−helix interaction is structurally similar to the glycophorin A dimer. Notice that these helix−helix interactions have a narrow range of short interhelical axial distances ranging from 6.0 to 7.0 Å. (c) Similar to panel a, except the histogram is composed of 1106 helix−helix interactions containing the AXXXA motif (black bars, mean interhelical axial distance of 8.8 ± 2 Å). Unlike the helix−helix interactions containing the GXXXG motif, helix−helix interactions containing the AXXXA motif appear to have interhelical distances consistent with helix−helix interactions not containing the AXXXA motif (white bars, mean = 8.9 ± 2.5 Å).

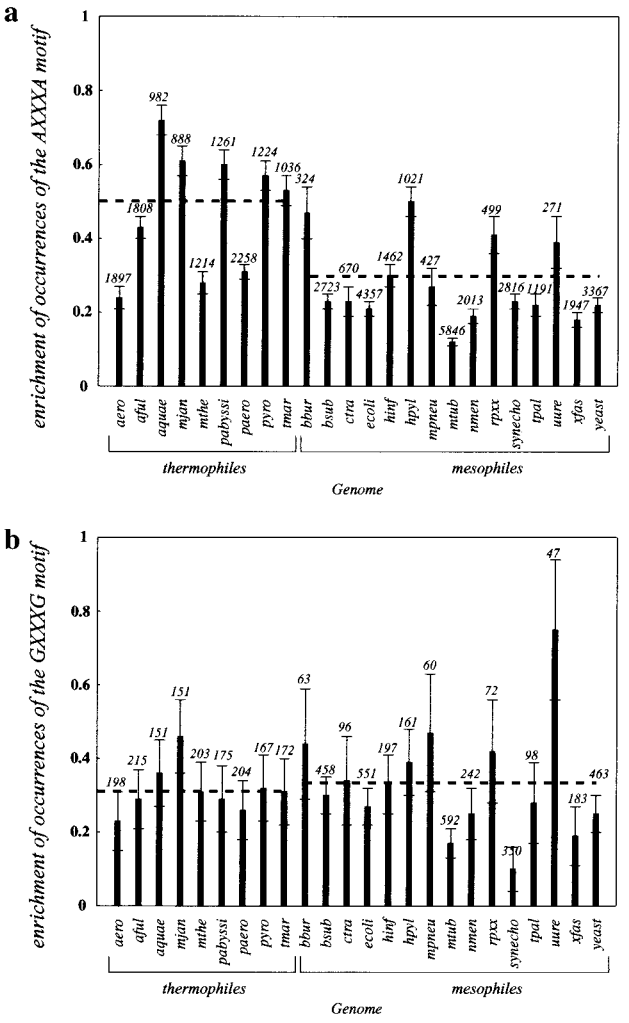

FIGURE 4: Enrichment of observed occurrences of the (a) AXXXA and (b) GXXXG amino acid sequence motifs over expected occurrences in α-helices from fully sequenced genomes. The genomes have been classified into two categories: thermophiles and mesophiles. The dashed lines represent the average enrichment of observed occurrences of the motif over expected for either the thermophilic group or the mesophilic one. Notice that the average enrichment for the AXXXA motif is greater for thermophiles than for mesophiles, suggesting that the AXXXA motif may function in thermophiles to increase the thermostability of proteins: aero, *A. pernix*; aful, *Archaeoglobus fulgidus*; aquae, *Aq. aeolicus*; mjan, *Methanococcus jannaschii*; mthe, *M. thermoautotrophicum*; pabyssi, *Pyrococcus abyssi*; paero, *P. aerophilum*; pyro, *Pyrococcus horikoshii*; tmar, *T. maritima*; bbur, *Borrelia burgdorferi*; bsub, *Bacillus subtilis*; ctra, *Chlamydia trachomatix*; ecoli, *Escherichia coli*; hinf, *Haemophilus influenzae*; hpyl, *Helicobacter pylori*; mpneu, *Mycoplasma pneumoniae*; mtub, *Mycobacterium tuberculosis*; nmen, *Neisseria meningitidis*; rpxx, *Richettsia prowazekii*; synecho, *Synechocystis* sp.; tpal, *Treponema pallidum*; uure, *Ureaplasma urealyticum*; xfas, *Xyella fastidiosa*; and yeast, *Saccharomyces cerevisiae*.

dicted to contain 14 702 α-helices (between 5 and 25 residues in length). The GXXXG motif is observed 172 times in the α-helices. One would expect only 132 occurrences of the GXXXG motif given the length and composition of the helices, corresponding to an enrichment of 30 ± 9% for observed occurrences over expected.

As for the GXXXG motif, occurrences of the AXXXA motif in α-helices from *T. maritima* are also enriched over expected occurrences. The AXXXA motif is observed 1036 times in the 14 702 predicted α-helices. One would expect only 679 occurrences of the AXXXA motif, corresponding to an enrichment of 53 ± 4%. Results for all 24 sequenced genomes are shown in Figure 4.

Thermophilic organisms tend to have greater enrichment of observed occurrences of the AXXXA motif over expected occurrences than in mesophiles. The average enrichment of observed occurrences of the AXXXA motif over expected is 0.48 for the 9 thermophiles, whereas the average enrichment is 0.28 for the 15 mesophiles. Unlike the case for the AXXXA motif, there appears to be no correlation between enrichment of the GXXXG motif and the thermophilicity of the organism.

DISCUSSION

The GXXXG Motif and the Cα−H···O Hydrogen Bond Stabilize Helix−Helix Interactions in Protein Structures. The GXXXG motif has been well characterized as a helix−helix stabilizing motif in membrane proteins. Here we show that the GXXXG motif stabilizes helix−helix interactions in soluble proteins as well. In fact, we show that 26 helix−helix interactions are similar to the helix−helix interaction of glycophorin A. This is significant because glycophorin A-like helix−helix interactions display backbone-to-backbone contacts of the Cα−H···O type that have the geometric hallmarks of hydrogen bond formation, suggesting that Cα−H···O hydrogen bonds stabilize helix−helix interactions in soluble proteins as well as membrane proteins.

The existence of C−H hydrogen bond donor groups in biological molecular structures has received an increasing amount of attention (*17*). Recently, the energy of the Cα−H···O type hydrogen bond was estimated to be 2.5−3.0 kcal/mol in vacuo, or approximately one-half the energy of a N−H···O hydrogen bond (*18*). Therefore, several coordinated Cα−H···O hydrogen bonds could contribute significantly to protein stability. Senes et al. (*7*) compiled a list of 145 potential hydrogen bonds of the Cα−H···O type from a structural database of transmembrane helix−helix interactions. Notably, the GXXXG motif was found at several of the helical interfaces. The occurrence of Cα−H···O hydrogen bonds has been reported between β-strands in soluble protein structures, although few Cα−H···O hydrogen bonds were found between α-helices in those proteins (*19*). To our knowledge, this is the first report to enumerate extensive examples from the PDB of potential Cα−H···O hydrogen bonds between α-helices.

Are the Cα−H···O contacts observed in the 26 helix−helix interactions from the PDB contributing to the stability of the helix−helix interaction, or are the contacts incidental to the close interaction of the helices promoted by the GXXXG motif? One clue to this question is provided by the histogram of interhelical axial distances for helix−helix interactions containing the GXXXG motif (Figure 3a). A statistically significant enrichment of helix−helix interactions occurs in the range of interhelical distances of 6.0−7.0 Å. The 26 glycophorin A-like helix−helix interactions have interhelical distances in the same range (Figure 3b). The enrichment of this class of helix−helix interactions implies some biological significance, and we believe that the function of this interaction is to stabilize helix−helix interactions with important biological roles.

An example of an important biological role involving helix−helix interactions is the stabilization of protein−protein interfaces. The GXXXG motif in the E1β subunit of pyruvate dehydrogenase stabilizes a helix−helix interaction between the E1β and E1α subunits and contributes to the overall stability of the protein−protein interaction (*9*). The interaction of these two subunits is imperative for the proper function of the dehydrogenase. Together, 5 of the 26 helix−helix interactions help stabilize protein−protein interactions. Other potentially important roles involving helix−helix interactions include stabilizing the active site of an enzyme or stabilizing a folding intermediate along the folding pathway for a protein.

The AXXXA Motif May Increase the Thermostability of Proteins in Thermophiles. While the occurrence of the AXXXA motif is enhanced in all of the 24 fully sequenced genomes in our study, occurrences from thermophiles appear to be enhanced to a greater extent than those from mesophiles. In fact, *Aquifex aeolicus*, one of the most thermophilic bacteria known (*20*), has the largest enrichment of observed occurrences of the AXXXA motif over expected occurrences (72%) than any of the 24 genomes.

Why are observations of the AXXXA motif and not the GXXXG motif enriched in thermophiles over mesophiles? Although it is dependent upon the position on the α-helix, alanine can stabilize the folded state of a helix-containing protein as much as 2.0 kcal/mol more than glycine (*21*, *22*). Therefore, incorporation of the GXXXG motif in an α-helix of a protein might be energetically unfavorable to the folded state, but locally would stabilize the helix−helix interaction. To make a protein thermostable, it is necessary to optimize the energy of the folded state, and therefore, the GXXXG motif would not be as suitable.

How then does the AXXXA motif provide stability to proteins? It is unlikely that Cα−H···O hydrogen bonds are involved to the same extent as with GXXXG, mainly because of the increased interhelical distance imposed on the helix−helix interaction by alanine at the helix−helix interface as compared to glycine. In addition, only minor enrichment of short interhelical distances was observed for helix−helix interactions containing the AXXXA motif that would be suggestive of backbone-to-backbone hydrogen bonds. However, the AXXXA motif, like the GXXXG motif, provides a complementary surface for interaction. Second, it permits the two helices to come into close contact with each other, facilitating contact between the other interfacial residues. Finally, there is no loss of side chain entropy for alanine residues upon helix−helix association, whereas residues with many conformations would lose entropy upon association.

Why do not all of the thermophiles show greater enhancement for occurrences of the AXXXA motif than the mesophiles? Three of the thermophiles (*Aeropyrun pernix*, *Methanobacterium thermoautotrophicum*, *and Pyrobaculum aerophilum*) show percent increases similar to those of mesophiles; these organisms are evolutionarily more similar to each other and form a branch on the 16S RNA-based tree of life (*23*). Therefore, it is plausible that the mechanism of thermostability that utilizes the AXXXA motif was evolved after a common ancestor to these three organism diverged from other thermophiles. There are other mechanisms for thermostability, such as disulfide bonds, which these organisms can use instead. In addition, *Heliobacter pylori* is the only mesophile with a percent increase similar to those of thermophiles. In fact, *H. pylori* is an extremophile itself and lives at low pH (*24*), suggesting the possibility that the AXXXA motif may also be useful for enhancing the stability of proteins in this environment.

ACKNOWLEDGMENT

We thank Dr. Tom Graeber, Dr. Michael Sawaya, and Jennifer Padilla for helpful discussion and Dr. Lukasz Salwinski for assistance in computer programming.

SUPPORTING INFORMATION AVAILABLE

A table documenting the 102 potential Cα−H···O hydrogen bonds in the 26 helix−helix interactions similar in structure to the helix−helix interaction of glycophorin A, including statistics providing distances and angles for each hydrogen bond. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES

1. Hofmann, K., Bucher, P., Falquet, L., and Bairoch, A. (1999) *Nucleic Acids Res. 27*, 215−219.
2. Wolfe, S. A., Nekludova, L., and Pabo, C. O. (2000) *Annu. Rev. Biophys. Biomol. Struct. 29*, 183−212.
3. Russ, W. P., and Engelman, D. M. (2000) *J. Mol. Biol. 296*, 911−919.
4. Senes, A., Gerstein, M., and Engelman, D. M. (2000) *J. Mol. Biol. 296*, 921−936.
5. Russ, W. P., and Engelman, D. M. (1999) *Proc. Natl. Acad. Sci. U.S.A. 96*, 863−868.
6. MacKenzie, K. R., Prestegard, J. H., and Engelman, D. M. (1997) *Science 276*, 131−133.
7. Senes, A., Ubarretxena-Belandia, I., and Engelman, D. M. (2001) *Proc. Natl. Acad. Sci. U.S.A. 98*, 9056−9061.
8. MacKenzie, K. R., and Engelman, D. M. (1998) *Proc. Natl. Acad. Sci. U.S.A. 95*, 3583−3590.
9. Kleiger, G., Perry, J., and Eisenberg, D. (2001) *Biochemistry 40*, 14484−14492.
10. Hobohm, U., Scharf, M., and Schneider, R. (1993) *Protein Sci. 1*, 409−417.
11. Kabsch, W., and Sander, C. (1983) *Biopolymers 22*, 2577−2637.
12. Cohen, G. H., Padlan, E. A., and Davies, D. R. (1986) *J. Mol. Biol. 190*, 593−604.
13. Chothia, C., Levitt, M., and Richardson, D. (1981) *J. Mol. Biol. 25*, 215−250.
14. Rost, B. (1996) *Methods Enzymol. 266*, 525−539.
15. Devereux, J., Haeberli, P., and Smithies, O. (1984) *Nucleic Acids Res. 12*, 387−395.
16. Cramér, H. (1946) *Mathematical methods of statistics*, Princeton University Press, Princeton, NJ.
17. Weiss, M. S., Brandl, M., Suhnel, J., Pal, D., and Hilgenfeld, R. (2001) *Trends Biochem. Sci. 26*, 521−523.
18. Scheiner, S., Kar, T., and Gu, Y. (2001) *J. Biol. Chem. 276*, 9832−9837.
19. Derewenda, Z. S., Lee, L., and Derewenda, U. (1995) *J. Mol. Biol. 252*, 248−262.
20. Deckert, G., Warren, P. V., Gaasterland, T., Young, W. G., Lenox, A. L., Graham, D. E., Overbeek, R., Snead, M. A., Keller, M., Aujay, M., Huber, R., Feldman, R. A., Short, J. M., Olsen, G. J., and Swanson, R. V. (1998) *Nature 392*, 353−358.
21. Serrano, L., Neira, J. L., Sancho, J., and Fersht, A. R. (1992) *Nature 356*, 453−455.
22. Chakrabartty, A., Schellman, J. A., and Baldwin, R. L. (1991) *Nature 351*, 586−588.

23. Woese, C. R. (2000) *Proc. Natl. Acad. Sci. U.S.A. 97*, 8392−8396.
24. Tomb, J. F., White, O., Kerlavage, A. R., Clayton, R. A., Sutton, G. G., Fleischmann, R. D., Ketchum, K. A., Klenk, H. P., Gill, S., Dougherty, B. A., Nelson, K., Quackenbush, J., Zhou, L., Kirkness, E. F., Peterson, S., Loftus, B., Richardson, D., Dodson, R., Khalak, H. G., Glodek, A., McKenney, K., Fitzegerald, L. M., Lee, N., Adams, M. D., Venter, J. C., et al. (1997) *Nature 388*, 539−547.
25. Ariyoshi, M., Vassylyev, D. G., Iwasaki, H., Nakamura, H., Shinagawa, H., and Morikawa, K. (1994) *Cell 78*, 1063−1072.

BI0200763