doi:10.1006/jmbi.2000.4364 available online at http://www.idealibrary.com on IDEAL® *J. Mol. Biol.* (2001) **305**, 1145–1159

JMB

Contributions of Residue Pairing to β-sheet Formation: Conservation and Covariation of Amino Acid Residue Pairs on Antiparallel β-strands

Yael Mandel-Gutfreund[1], Sydney M. Zaremba[2] and Lydia M. Gregoret[1]*

[1]*Department of Chemistry and Biochemistry*

[2]*Department of Molecular, Cell and Developmental Biology, University of California, Santa Cruz, CA 95064, USA*

**Corresponding author*

In an effort to better understand β-sheet assembly, we have investigated the evolutionary behavior of neighboring residues on adjacent antiparallel β-strands. Residue pairs were classified according to solvent exposure as well as by whether their backbone NH and C═O groups are hydrogen bonded. The conservation and covariation of 19,241 pairs in 219 sequence alignments was analyzed. Buried pairs were found to be the most conserved, while stronger covariation was detected in the solvent-exposed pairs. However, residues on neighboring strands showed a degree of conservation and covariation similar to that of well-separated residues on the same strand, suggesting that evolutionary pressure to maintain complementarity between pairs on neighboring strands is weak. Moreover, in spite of the preference of certain amino acid pairs to occupy neighboring positions on adjacent strands, such favored pairs are neither more strongly mutually conserved nor covary more strongly than pairs of the same type in non-interacting positions. Although the β-sheet pairs did not show outstanding evolutionary coupling, in many protein families significant conservation and covariation patterns were detected for some of the residue pairs. Overall, the weak evolutionary conservation and covariation of the β-sheet pairs indicates that sheet structure is unlikely to be dictated by specific side-chain interactions.

Keywords: β-sheet assembly; residue pairs; conservation; covariation; protein folding

Introduction

Though much progress has been made in the last decade towards understanding the relationship between a protein's sequence and structure, the protein-folding problem remains a biomedically important and captivating puzzle. It is especially intriguing to speculate how β-sheet proteins, having complex topologies and involving numerous contacts between residues distant in sequence, acquire their native structures. Several features of β-sheet protein sequences could be important for efficient folding and stability. One possibility is that hydrophobic collapse is the dominant driving force for folding and stability (Kauzmann, 1959; Dill, 1985), and that the overall hydrophobic/polar pattern of amino acids defines a protein's topology (Eisenberg *et al.*, 1984; Bowie *et al.*, 1990; Kamtekar *et al.*, 1993). A second possibility is that the formation of turns at critical locations is required for antiparallel sheet folding. Supporting this hypothesis are two recent findings by two different laboratories, that residues in the distal loop of SH3 domains are important for nucleating folding (Martinez & Serrano, 1999; Riddle *et al.*, 1999). A third possibility is that recognition between amino acid side-chains on neighboring β-strands guides correct strand register and stabilizes the resulting β-sheet. Optimizing such interactions has led to the successful design of several β-sheet peptides and proteins (Kortemme *et al.*, 1998; de Alba *et al.*, 1999). In reality, different combinations of all three of these features most likely determine β-sheet topology and code for protein stability. Here, we study the relative importance of complementarity between side-chains on neighboring β-strands by examining their evolutionary conservation and covariation in protein families of known structure.

Abbreviations used: HB, hydrogen bonded; NHB, non-hydrogen bonded.

E-mail address of the corresponding author: gregoret@chemistry.ucsc.edu

0022-2836/01/051145–15 $35.00/0

Several statistical studies have analyzed amino acid side-chain pairing preferences in antiparallel β-sheets. In 1980, Lifson and Sander reported that specific amino acid pairs, such as oppositely charged residues, side-chains with hydrogen bonding potential and β-branched residue pairs are statistically favored in antiparallel β-sheets (Lifson & Sander, 1980). Since only 30 representative protein structures were used in that study (due to the small number of solved three-dimensional structures), grouping of amino acids into categories according to their chemical properties was required. More recently, with the tremendous growth of the structural database, it has become possible to derive specific residue pairing propensities for all pairs of amino acids. In two studies, by Wouters and Curmi (1995) and by Hutchinson and co-workers (1998), different pairing preferences were found for distinct types of antiparallel β-sheets pairs: pairs in which the backbone carbonyl and amide groups are hydrogen bonded to one another (HB) and those in which the backbone groups are not hydrogen bonded (NHB) (see Figure 1).

Although both the Wouters & Curmi (1995) and Hutchinson *et al.* (1998) studies demonstrated differences in the pairing preferences between HB and NHB pairs, with the exception of charged and cysteine-cysteine pairs, somewhat different preferences were uncovered in each analysis. The weak correlation between the pairing preferences found in the two studies may suggest that these interactions are not instrumental in determining sheet topology. None-the-less, the derived preferences, particularly those for residue pairs of opposite charge, have been confirmed experimentally to stabilize sheets (Smith & Regan, 1995) and have been useful in designing monomeric β-sheet peptides (Kortemme *et al.*, 1998; de Alba *et al.*, 1999). On the other hand, our own experimental study, in which four different pairs of residues on antiparallel strands of a small protein were randomized and assayed for stability, suggested that overall, pairing preferences are not very strong and are highly modulated by the local environment (Zaremba & Gregoret, 1999).

The increasing number of protein structures from a diverse set of protein families, together with the large numbers of homologous sequences, affords us an opportunity to probe the importance of residue pairing using evolutionary information. In contrast to the previous statistical studies that analyzed the occurrence of amino acid pairs in β-sheets, our goal is to assess the importance of residue pairing in general. We propose that if residue complementarity is important for protein stability or folding, it should be maintained throughout evolution.

It is well established that evolutionary conservation is indicative of the importance of a given residue playing a role in the function or stability of the protein (Dickerson, 1971; Lesk & Chothia, 1980a; Mirny & Shakhnovich, 1999; Hamill *et al.*, 2000a). Many studies have examined whether compensatory changes between residues are indicative of their spatial proximity (e.g. Altschuh *et al.*, 1988; Chelvanayagam *et al.*, 1997). Although mostly weak compensatory covariation signals were found, in some cases it has been shown that correlated mutations can be used to identify residues that contact one another in the tertiary structure of the protein family under investigation (Göbel *et al.*, 1994; Shindyalov *et al.*, 1994; Thomas *et al.*, 1996; Pazos *et al.*, 1997). Furthermore, residue-residue contacts have been used to improve secondary and tertiary structure predictions (Benner & Gerloff, 1991; Gerloff *et al.*, 1997; Olmea & Valencia, 1997; Ortiz *et al.*, 1999). For example, by identifying two compensatory changes in consecutive β-strands in the protein kinase family, Benner & Gerloff (1991) correctly predicted the sheet structure of the small lobe of the cyclic adenosine monophosphate-dependent protein kinase.

Analyses of compensatory base-pair changes and base-pair conservation are commonly used for predicting RNA secondary structures (Eddy & Durbin, 1994; Juan & Wilson, 1999). While amino acid pair interactions are much less specific than RNA base-pairing, an analogy may be made between adjacent β-strands and an RNA double helix. Even though the protein alphabet is much more complex than the nucleic acid alphabet, if residue complementarity is important for β-sheet stability or folding, it may be possible to detect the coordinated evolution of residues on neighboring strands using methods similar to those used for RNA secondary structure prediction.

To study the contribution of residue pairing to β-sheet formation we consider both the conservation and the covariation of the pairs in families of protein sequences aligned to a representative sequence with a known three-dimensional structure. While conservation and covariation result from similar evolutionary constraints, these two terms are, in fact, quite different mathematically. For example, when two residue positions in a multiple sequence alignment are completely conserved, no inference can be made about whether or not they covary. Considering both conservation and covariation therefore should give us a more comprehensive view of the importance of residue pairs to sheet formation.

In addition to studying the concerted evolution of residue pairs on neighboring strands in general, we also compare the behavior of specific types of pairs. One major conclusion reached by Wouters & Curmi (1995) and by Hutchinson *et al.* (1998) is that the HB and NHB positions in antiparallel β-sheets have different amino acid pairing propensities. It was suggested that the pairing preferences result from the different main-chain geometries at the HB and NHB sites that allow, in each case, for only certain amino acid pairs to exhibit intimate side-chain contacts while maintaining their favorable side-chain rotamer conformations (Hutchinson *et al.*, 1998). This is specifically true in the NHB

positions, where the distance between the α-carbon atoms tends to be shorter, forcing side-chains to pack more intimately. Intriguingly, in our limited experimental study of residue pairing, we found that the solvent-exposed HB site was more tolerant of a variety of substitutions, while the NHB site was more restrictive (Zaremba & Gregoret, 1999). Therefore, if the packing interactions are more ideal between NHB pairs, these positions could be more tightly coupled evolutionarily than HB pairs. Here, we investigated this secondary hypothesis based on a large data set of representative sequence alignments.

Results and Discussion

Despite the growing number of solved protein structures, there are still insufficient data to track the evolution of residues that are involved in interstrand β-sheet pairing from structural data alone. It is possible, however, to deduce evolutionary information from sequences of closely related proteins for which the structure of at least one representative protein is known. We obtained sequence alignments for 219 representative antiparallel β-sheet-containing domains of known three-dimensional structure (Table 1). In the alignments, we included only closely homologous proteins with, at minimum, 35 % sequence identity in order to be reasonably confident that the sequence alignments were structurally accurate. To reduce bias from redundant sequences, all sequences with more than 90 % identity were removed from the alignments. Since our analysis covered a diverse set of alignments, each including a different number of sequences varying in sequence identity, we standardized the alignments by weighting the frequencies of each position (column) using the Henikoff algorithm (Henikoff & Henikoff, 1994).

We considered two types of pairs on adjacent antiparallel strands (Figure 1): hydrogen bonded (HB) pairs and non-hydrogen bonded (NHB) pairs. Figure 2 outlines the major steps of the analysis exemplified for the SH3 domain family. From the three-dimensional structure of the representative domain, we extracted all antiparallel β-sheet residue pairs and classified them by their hydrogen bonding pattern. The residues involved in the interstrand pairings were then ascribed to the corresponding positions in the alignment, with the top sequence in the alignment being that of the experimentally determined, representative structure. The conservation and covariation were then calculated for each pair in the alignment. Table 2 summarizes the number of pairs analyzed in each group. Since residues in multistrand sheets can belong to two different groups simultaneously, we also analyzed the conservation and covariation in a subgroup of pairs from β-ribbons (two stranded sheets) in which residues belong to only one pair type (HB or NHB).

In a previous statistical analysis of residue pairing preferences in antiparallel β-sheets, Hutchinson *et al.* (1998) compared the interstrand HB and NHB pairs to intrastrand pairs that are displaced by two residues along the same strand ($i,i+2$). In their study, Hutchinson *et al.* (1998) chose the $i,i+2$ pairs as a control because, although the distance between them is too great to allow for intimate contact, they are expected to have the same β-sheet

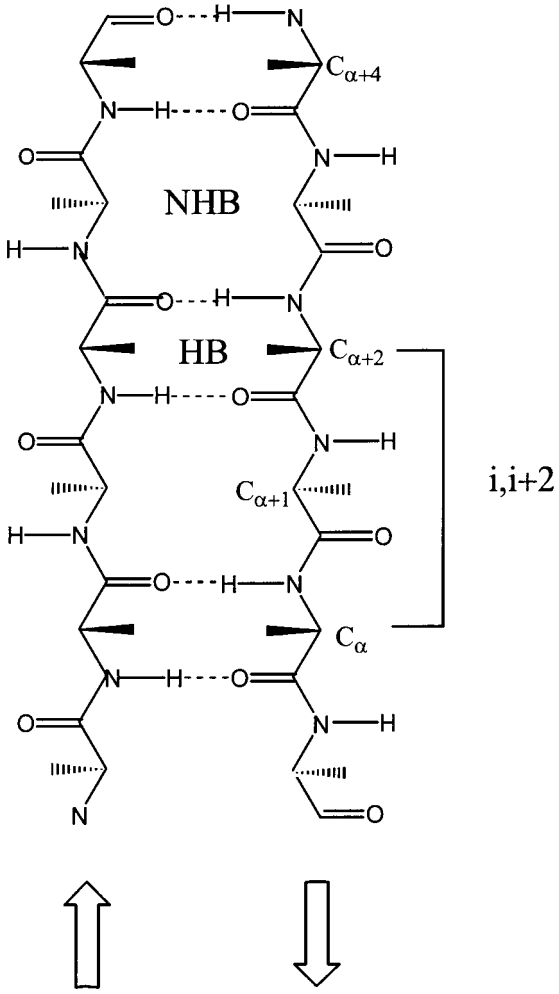

Figure 1. An antiparallel β-sheet with two strands. NHB, a non-hydrogen bonded interstrand pair; HB: a hydrogen-bonded interstrand pair; $i,i+2$, an intrastrand pair with a residue separation of two.

Table 1. Statistics on family alignments used in this study

Number of alignments	219
Minimum number of sequences in alignment	15
Maximum number of sequences in alignment	264
Median sequences in alignment	28
Minimum length of sequences in alignment	25
Maximum length of sequences in alignment	294
Median sequence length	123
Number of unique superfamilies	114
Number of unique folds	96

Figure 2. A representation of the procedure used to extract the β-sheet interstrand pairs from the sequence alignment. The HB (light gray) and NHB (dark gray) β-sheet pairs are illustrated on the three-dimensional model of the representative protein (1sem). The corresponding pairs in the family sequence alignment around the representative protein are shaded in light and dark gray, respectively. The top sequence refers to the representative protein whose structure is known.

propensity as the interstrand pairs and are likely to be in a similar environment on the same face of the sheet. We similarly examined the $i,i+2$ intrastrand pairs as well as pairs displaced by one, three and four residues as controls.

The β-sheet residues analyzed in this study were assigned to different environments according to their location within the protein, i.e. buried in the core or on the surface. A residue's burial is conventionally determined by calculating its normalized solvent-accessible surface area (Lee & Richards, 1971). We used 25% exposure as the threshold value for considering a residue exposed. Most of residue pairs analyzed in our study fell into the buried category (Table 2). Although the interstrand pairs are usually expected to be in a similar environment, ''mixed'' pairs of buried and exposed positions were frequently observed (Table 2). These pairs were usually on the protein surface with one of the residues buried by a neighboring position. We considered these mixed pairs separately.

Table 2. Frequencies of pair types analyzed

	Buried	Exposed	Mixed
HB	1005 (191)	134 (56)	421 (106)
NHB	898 (160)	217 (97)	491 (115)
$i,i+1$	2925	621	1946
$i,i+2$	2584	684	1036
$i,i+3$	1811	363	1277
$i,i+4$	1437	387	1004

Buried pairs have $\leqslant 25\,\%$ surface accessibility; exposed pairs have $>25\,\%$ surface accessibility. Mixed pairs involve one buried and one exposed residue. The number of residues in the subset of pairs originating from β-ribbon only are given in parentheses. HB, hydrogen bonded interstrand β-sheet pairs. NHB, non-hydrogen bonded interstrand β-sheet pairs. $i,i+N$, intrastrand pairs displaced by N residues.

Conservation

The most readily apparent evolutionary information provided by sequence alignments of homologous proteins is the conservation of specific residues. In this study, we used two different measures for computing conservation. First, we computed the conservation of the positions as an average of the conservation of the individual positions. However, since this measure does not take into account any coupling between the residues, we developed a second measure based on the chemical features of the pair as a whole (including combined volume, overall charge and polarity).

In studying the role of evolutionarily conserved residues in β-sheet formation, it is important to distinguish amino acid conservation that is due to structural constraints from conservation of functionally important regions (Mirny & Shakhnovich, 1999). To ensure that our results are not biased by extremely high conservation at the functionally important positions, we eliminated from the alignments positions for which extensive functional studies have been performed (e.g. SH3 domain family, ribosomal proteins) and recalculated the conservation values. Although, as expected, the most conserved residues in the sequence alignments were related to the functionally important groups, no significant differences were detected in our overall results after exclusion of the functionally important residues (not shown).

Conservation of interstrand pairs

We first examined whether interstrand β-sheet pairs on neighboring strands within each environment show stronger conservation values than the intrastrand pairs on the same strand. Figure 3 shows histograms of the average conservation values for the HB (black), NHB (gray), and $i,i+2$ (hatched) residue pairs in the three different environments. As can be seen by the shift of the histograms, the buried positions (Figure 3(a)) are consistently more conserved than the exposed positions (Figure 3(c)) with a higher (though not significantly so) mean value of 0.81 ± 0.15 (calculated for all three groups HB, NHB and $i,i+2$) than for the exposed pairs (0.67 ± 0.16). Mixed pairs (Figure 3(b)) have intermediate values (0.73 ± 0.16). The same dependence of the environment (e.g. buried or exposed) on sequence conservation was found regardless of the type of the residue pairs analyzed. As can be seen by comparing the distribution of the conservation values in the three histograms in Figure 3, the intrastrand ($i,i+2$) pairs (hatched) have a pattern of conservation similar to that of the interstrand (HB, black and NHB, gray) pairs, again with buried positions

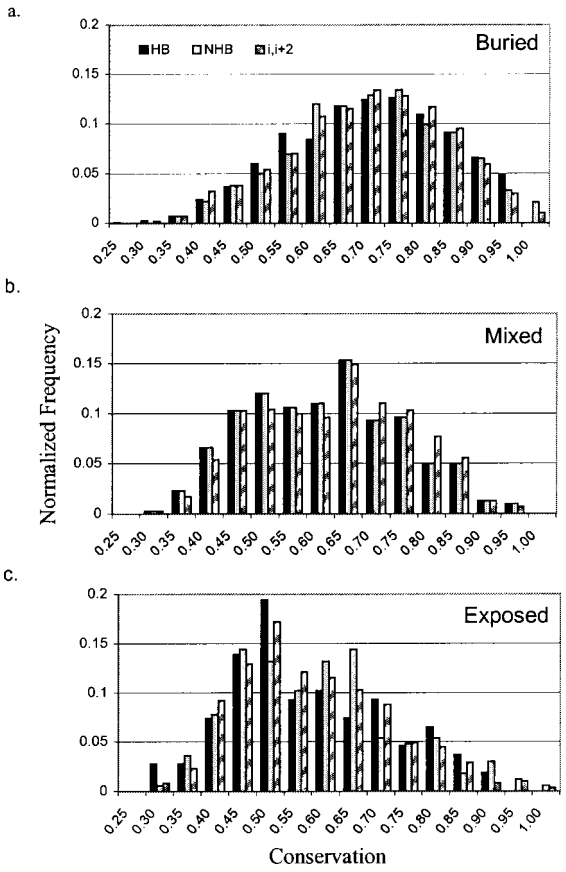

Figure 3. Relative frequencies of average conservation for three types of pairs, HB (black), NHB (gray) and $i,i+2$ (hatched), in three different environments; (a) buried, (b) mixed and (c) exposed. Conservation values were obtained using the McLachlan (1971) amino-acid similarity values. Values higher than 1 represent conservation of rare amino acids (e.g. Trp).

being more strongly conserved than the exposed positions. There was no detectable difference in conservation of the HB *versus* the NHB pairs in our data.

To ensure that our results are not influenced by the fact that residues on the interior strands of a sheet are simultaneously involved in an HB interaction to a residue on one strand and in an NHB interaction with a residue on another strand, we studied the conservation of pairs from two-stranded β-ribbons separately. While there was some difference in the overall distribution of conservation values, as shown in Figure 4, we did not find any significant difference in the conservation of the HB and NHB pairs within β-ribbons. The overall shift toward lower conservation values for the β-ribbon pairs in Figure 4(b) is due to a significantly smaller fraction of buried pairs in the ribbons as compared with sheets overall (18 % *versus* 60 %). When the β-ribbon pairs are broken down into the three different solvent-exposure classes (not shown), their conservation patterns are very similar to the results for all HB and NHB shown in Figure 3.

The conservation results presented in Figures 3 and 4(a) and (b) were obtained using the similarity substitution matrix reported by McLachlan (1971). Very similar results were observed when calculating average conservation based on sequence variation entropy (not shown) (Shenkin *et al.*, 1991). These results suggest that there is no evolutionary force keeping the residues of the β-sheet pairs more conserved than any other pairs of β-strand residues that are not involved in a strand-strand interaction. Moreover, the weak conservation of interstrand pairs does not appear to be a consequence of averaging effects from the β-sheet pairs being involved in more than one pairing interaction (i.e. interacting with a strand on either side): the conservation analysis of HB and NHB pairs from two stranded β-ribbons mirrors the results for HB and NHB pairs overall (Figure 4(a) and (b)).

In a recent analysis, Nagarajaram *et al.* (1999) showed a correlation between the sum of the volume of the two residues in interstrand pairs and the inter-axial distance between the two strands. It is possible that alteration in the distances between β-strands is the mechanism by which the protein compensates for changes in the amino acid properties during evolution, while maintaining the packing requirements of HB and NHB pairs. This could explain the relative weak conservation we observed for both HB and NHB pairs. A similar compensation mechanism was suggested earlier (Lesk & Chothia, 1980a) for residues involved in helix-helix packing.

Pair property conservation

It is possible that we did not observe stronger conservation for residue pairs on adjacent strands because conservation was measured by averaging the conservation of the two individual positions of the pair. Such a calculation does not take into account the possibility that conservation is maintained for the pair as a whole and not for each individual residue separately. We chose to analyze the conservation of residue pairs by measuring their similarity based on the change in the physical-chemical parameters of the two residues con-

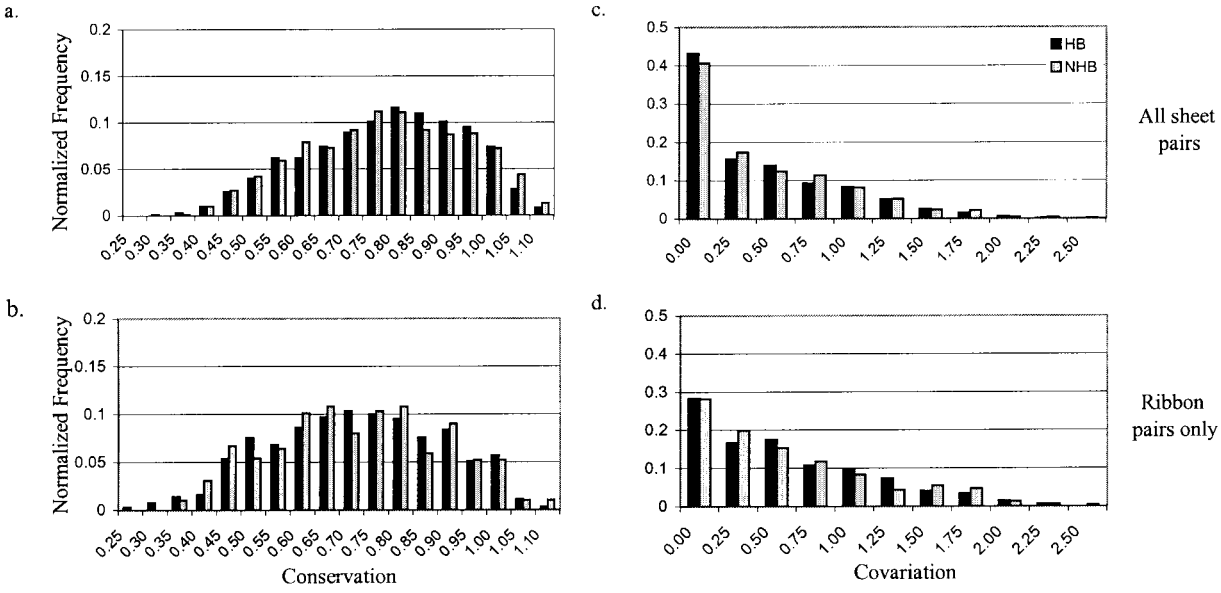

Figure 4. Distribution of the relative frequency of average conservation values and covariation for hydrogen-bonded (black) and non-hydrogen bonded (gray) all β-sheet pairs. (a) Average conservation of pairs in the complete data set (b) Average conservation of pairs in β-ribbons. (c) Covariation (mutual information method) of pairs in the complete data set. (d) Covariation in β-ribbons.

sidered together. Amino acid-amino acid similarity values extracted from substitution matrices have been shown to be correlated, to some extent, with the similarity in their physical-chemical properties (Koshi & Goldstein, 1997). To evaluate the similarity between amino acids pairs, we assigned three properties to each residue pair under investigation in the alignment: average volume (Richards, 1974), overall hydrophobicity (Fauchere & Pliska, 1983), and average charge. We refer to this measure as ''pair property conservation.'' The distance between the properties of all possible residue pairs was calculated. For example the amino acid pair R-E was represented by the vector (0.79, 0.935, 0) describing the average normalized value of *R* (0.91, 1.00, 1.0) and *E* (0.67, 0.87, −1.0). A second amino acid pair, A-V, was represented by the vector (0.53, 0.46, 0.0). If, at a given pair of positions in a sequence alignment, the pair R-E was found in one sequence and the pair, A-V found in another, the similarity between the two pairs was calculated as the distance between the vectors of pairs R-E and A-V (0.55). It is important to note that this measure of conservation is position-independent. In other words, a mutation of pair A-V to pair V-A would have a vector length of zero.

By deriving the distances for all amino acid-amino acid pairs, we computed the average value for all pair substitutions observed for any given pair of positions in the alignment. Using this method, we found a significant difference between the average vector lengths of the buried interstrand pairs (0.12 ± 0.11) compared to that of the exposed pairs (0.30 ± 0.14). Mixed pairs showed intermediate values (0.26 ± 0.14). (Shorter distances indicate higher pair conservation.) As in the average conservation analysis described earlier, we could not detect any overall difference between the distinct types of pairs analyzed (e.g. HB, NHB, and $i,i+2$; not shown). Very similar results were obtained for the small subset of HB and NHB pairs from β-ribbons. However, using the pair property conservation measure, the strong conservation of buried pairs of all types was more apparent.

Conservation of favorable pairs

The results presented thus far indicate that residue pairs on adjacent β-strands (HB and HNB) are no better conserved than residues on the same strand. It is possible that while most residue pairs have a background level of conservation, certain critical pairs are conserved more strongly. If there are only a few such pairs in each alignment, they may not stand out in a broad statistical analysis.

The pairs that could be critical may be those that form particularly good interactions with one another. We thus investigated whether those pairs that were found more frequently than expected in interstrand interactions (Wouters & Curmi, 1995; Hutchinson *et al.*, 1998) show stronger than average pair conservation. We calculated the pair property conservation of all statistically favorable amino acid-amino acid pairs that were in either HB or NHB positions in the template structure using the propensities reported by Hutchinson *et al.* (1998). Table 3 summarizes the mean pair property conservation calculated of each type of pair. As a control, we calculated the mean pair property conservation of these pairs in $i,i+4$ positions, since residues separated by four residues along a β-strand cannot possibly interact.

As shown in Table 3, the most conserved of the favorable pairs were those that involved hydrophobic and aromatic residues in the representative structure (e.g. F-L, F-Y) while positions that were occupied by polar or charged pairs in the representative structure had much higher average pair distances (e.g. K-Q, K-E). These results are in agreement with the general concept that the core of the protein, where the hydrophobic and aromatic residues typically reside, is under strong evolutionary pressure to remain non-polar, while the surface of the protein is allowed much more variability (Lesk & Chothia, 1980b). The same level of conservation was observed, however, for the interacting (HB and NHB) as well as the non-interacting $(i,i+4)$ pairs. Furthermore, the pairs that

Table 3. Pair conservation of statistically significant residue pairs (Hutchinson *et al.*, 1998) in interstrand positions and in intrastrand positions $i,i+4$

Preferred HB pairs		
	HB	$i,i+4$
HH	0.05 (2)	0.41 (3)
KQ	0.42 (5)	0.42 (7)
KS	0.24 (10)	0.33 (13)
KR	0.31 (7)	0.21 (10)
RS	0.26 (14)	0.26 (11)
FL	0.07 (20)	0.12 (21)
FY	0.10 (9)	0.10 (10)
IY	0.24 (14)	0.16 (19)
VY	0.11 (21)	0.16 (23)
VW	0.18 (6)	0.14 (9)
GW	0.06 (3)	0.12 (2)
Preferred NHB pairs		
	NHB	$i,i+4$
TT	0.26 (17)	0.21 (12)
ST	0.23 (16)	0.26 (19)
DH	0.32 (5)	0.26 (4)
DK	0.30 (7)	0.37 (8)
DR	0.26 (7)	0.27 (4)
KE	0.32 (15)	0.31 (12)
KR	0.27 (11)	0.21 (10)
ET	0.28 (13)	0.34 (19)
KN	0.31 (3)	0.37 (5)
KS	0.33 (9)	0.33 (13)
KY	0.27 (11)	0.25 (11)
TR	0.18 (7)	0.28 (11)
FA	0.11 (12)	0.17 (9)
FL	0.10 (14)	0.12 (21)
CC	0.03 (14)	none
FP	0.15 (4)	0.11 (7)
PY	0.19 (9)	0.11 (7)

HB, hydrogen-bonded; NHB, non-hydrogen bonded; i, $i+4$, intrastrand pairs displaced by four residues. Numbers in parentheses designate the observed number of pairs in the representative sequences. Values close to 0 indicate strong conservation.

were found to be significantly favored in the statistical analysis (Hutchinson *et al.*, 1998) were not any more highly conserved than all other types of pairs. Among the pairs that showed statistically significant conservation in our study (p-value <0.05), only 12% of the HB and 13% of the NHB pairs were associated with the favorable pairs in the statistical studies. Thus it appears that even pairs that make ideal interactions are not any more strongly conserved than non-interacting pairs found in similar environments.

While we did not find the favorable pairs to be more strongly conserved, the number of pairs of each type found in the representative sequences did agree, in many cases, with the statistical analyses with respect to preferences of particular pairs for either the HB or NHB positions (Wouters & Curmi, 1995; Hutchinson *et al.*, 1998). The highest correlation was found for cysteine pairs (C-C), which showed a very strong preference to be in the NHB position in the statistical studies (Wouters & Curmi, 1995; Gunasekaran *et al.*, 1997; Hutchinson *et al.*, 1998), and are also found almost exclusively in NHB interstrand positions in our study (14 out of 15 pairs). As indicated in Table 3, these cysteine pairs were almost completely conserved in our alignments (pair property conservation distance = 0.03). Another example is the polar pair Thr-Thr (T-T) that was found preferably in NHB interstrand pairs in the statistical studies as well as in our study (17/1606 in NHB and 4/1560 in HB). However, the pair Ser-Thr (S-T), which is statistically favored in NHB interstrand positions, did not show any preference in our study. These polar pairs (T-T and S-T) were not found to be more strongly conserved than the other types of polar pairs in our analysis. Interestingly, seven out of the 16 S-T pairs in the NHB position did show statistically significant covariation. In another example, the pair Phe-Leu (F-L), which is favored in both HB and NHB positions according to both statistical studies, was observed frequently in our study (pair conservation distances of 0.07 and 0.10 for HB and NHB pairs, respectively). Although the pair conservation of the F-L pair was slightly weaker in the $i,i+4$ positions (0.12), it was still much stronger than in the other types of pairs.

Covariation

In addition to studying the conservation of the neighboring residues we studied the covariation of the pairs. Covariation is identified by correlated changes in two positions in an alignment. We took two different approaches to calculate covariation. The first approach, a method introduced by Göbel *et al.* (1994), represents the amino acid substitutions at an individual position in the alignment by a distance matrix. For each specific pair examined, a standard correlation coefficient value is calculated to determine to what extent the changes observed in the two positions are correlated with each other. In the second approach, covariation is calculated based on the mutual information between any two positions in the alignment, reflecting the dependence between the amino acid identities in the relevant positions (Clarke, 1995).

To ensure that our results are not dependent upon a particular method, we first assessed qualitatively how the two methods relate to each other, comparing the results obtained for several individual alignments. Overall, the two different methods yielded coherent results. Figure 5 illustrates covariation values calculated for β-sheet pairs in two representative proteins belonging to the trefoil beta fold. The structure shown in the two top panels is human fibroblast growth factor (2fgf) and the structure in the bottom panels is human interleukin 1β (1i1b). The interstrand β pairs are colored according to the significance of covariation, ranging from cyan (not significant, p-value <0.1) to red (highly significant, p-value <0.001). (The significance of the covariation was evaluated by calculating a Z-score, which is the number of standard deviation units by which a pair's covariation score deviates for the mean.) In Figure 5(a) and (c), Z-scores were calculated on the basis of the Göbel method (Göbel *et al.*, 1994) and in Figure 5(b) and (d), scores were calculated by the mutual information method (Clarke, 1995). In both proteins, the same pairs are identified as the most strongly covarying by both methods. However, there are differences between the methods in the statistical significance of the most highly covarying pairs. The Z-scores calculated by the Göbel method are less significant for fibroblast growth factor (Figure 5(a) *versus* (b)). For interleukin 1β (Figure 5(c) and (d)), the Z-scores calculated by the Göbel method are more significant and the two methods show very similar results overall. The differences in the significance of pairs observed in the fibroblast growth factor example are most likely a consequence of the different sensitivity of the two methods to the properties of the alignment, such as overall sequence variability, length of alignment, etc. Similar differences between the two methods, in terms of the significance with which they ranked the covariation of particular pairs, were observed in other protein families. These results convinced us to continue using both methods in our analysis.

Covariation in interstrand pairs

Using the Göbel and mutual information methods in tandem, we examined whether interstrand β-sheet pairs covary more strongly than intrastrand pairs. Figure 4(a) and (d) summarize the results obtained by the mutual information method. As shown, similar covariation values for HB and NHB pairs were obtained when analyzing all sheet pairs (Figure 4(c)) or β-ribbons pairs separately (Figure 4(d)). As in Figure 4(b), here again we can see an overall shift of the distribution of the β-ribbons pairs, this time towards higher covariation values, probably reflecting the overall lower

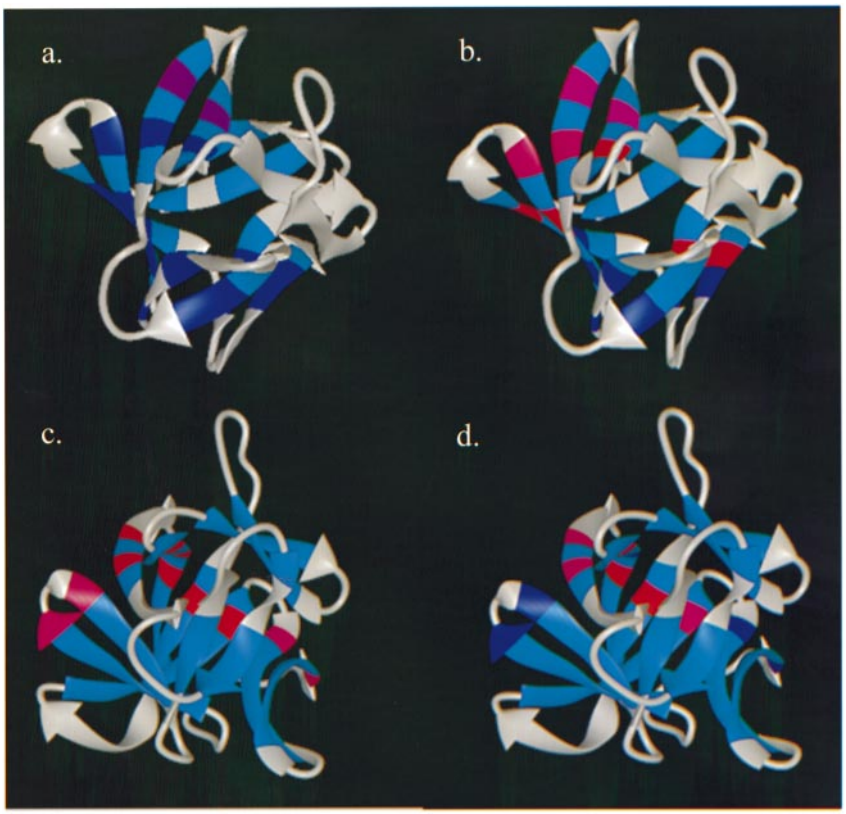

Figure 5. (a) and (b) Ribbon diagrams of human fibroblast growth factor (PDB ID 2fgf) colored by covariation *Z*-scores values from cyan (not significant, $p > 0.1$) to red (highly significant, $p < 0.001$), and calculated based on (a) the Göbel method and (b) the mutual information method. (c) and (d) Ribbon diagrams human recombinant interleukin-1 β (PDB ID 1ilb) colored as above. Results from the Göbel method are shown in (c) and results from the mutual information method are shown in (d).

fraction of buried pairs in the β-ribbon subset (Figure 4(d)). As in the conservation analysis, we calculated the covariation of the different types of pairs according to pair type and environment (buried, mixed, exposed). Figure 6 shows the distribution of the covariation results obtained by the Göbel and the mutual information methods. As observed in the analysis of conservation, the distribution of the relative frequencies (calculated by both methods) is very similar in each of the different types of pairs (HB, NHB, and $i,i+2$), indicating that there is little difference in the covariation of the different pair types. When comparing the covariation of the pairs from different environments, the exposed pairs (Figure 6(c) and (f)) tend to have stronger covariation signals than the buried pairs (Figure 6(a) and (d)) or mixed pairs (Figure 6(b) and (e)). Overall, these covariation results are consistent with our previous conservation results, which showed that all types of buried pairs are more conserved than exposed pairs and thus are expected to covary less because of low information content. These results also agree with an earlier study by Benner *et al.* (1994), which showed that hydrophilic variation tends to be related to surface residues while hydrophobic conservation and variation is characteristic of the protein core. Overall, the conservation and covariation results agree with the large body of data showing that core positions are much less tolerant to mutations than positions on the protein surface.

Since protein cores are highly conserved, it is possible that we did not observe covariation among buried pair positions because there is not enough sequence variability in some of the alignments in our data set to measure covariation. To investigate this possibility, we computed the mean expected covariation values for buried and exposed pairs in the different sequence alignment as a function of the number of sequences in each alignment and the average sequence identity of each alignment. The results are shown in Figure 7. Overall, buried pairs have significantly lower expected covariation values than the exposed pairs. (The mixed pairs, involving one exposed and one buried residue, (not shown) have intermediate values.) However, we could not detect any dependence between the number of sequences in the alignments and the covariation of the buried or exposed pairs. A very weak, though not statistically significant, dependence was observed between the covariation values of the exposed pairs and the average sequence identity of the alignment ($R^2 = 0.174$). No correlation between family sequence identity and expected covariation was observed for the buried pairs.

Given the expected, environment-specific covariation values for all pairs in an alignment, we can evaluate the statistical significance of the covariation observed in the sets of specific buried, exposed, and mixed sheet pairs. For each pair in our data, we calculate a *Z*-score, subtracting the expected covariation (given the environment) from

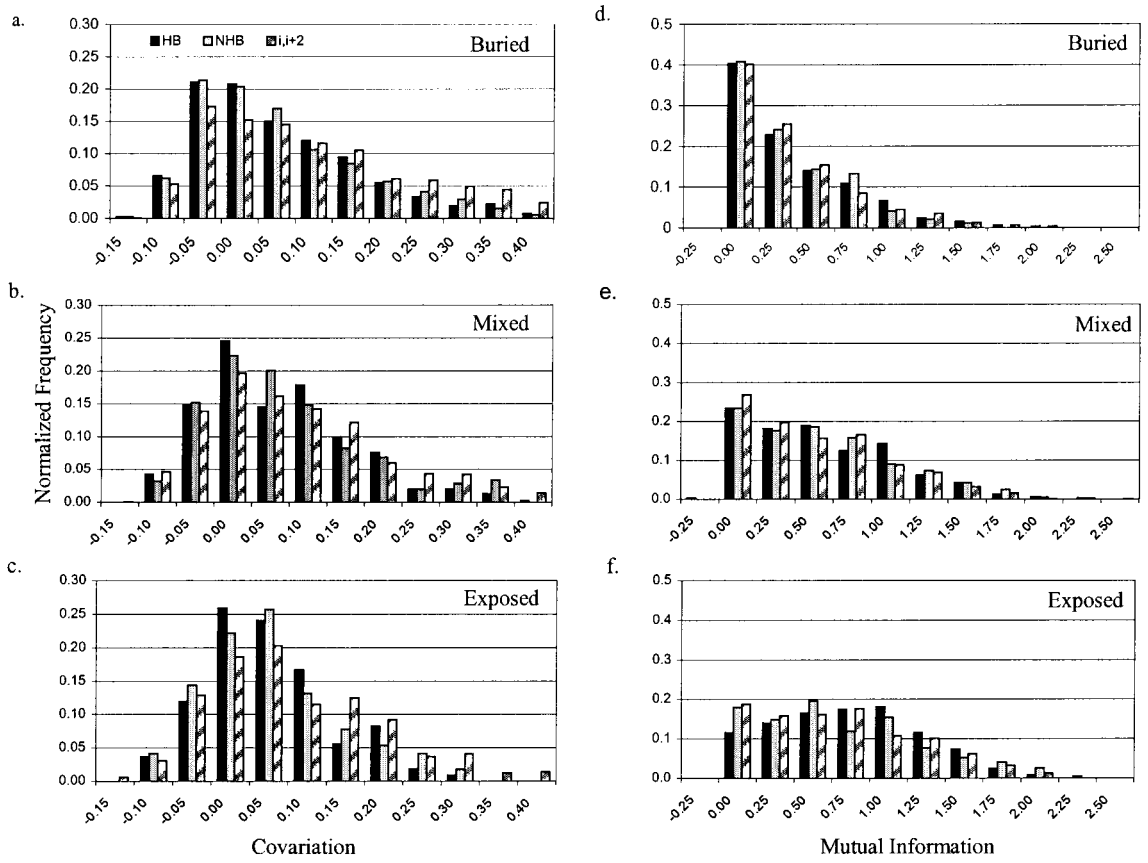

Figure 6. Relative frequencies of covariation values calculated based on the Göbel method (left; (a)-(c)) (values range from −1 to 1) and the mutual information method (right; (d)-(f)) (maximum value = 4.2) for three types of pairs, HB (black), NHB (gray) and $i,i+2$ (hatched), in three different environments, (a) and (d) buried, (b) and (e) mixed and (c) and (f) exposed.

the actual covariation calculated for the pair and dividing by the standard deviation. By using compatible expected values calculated from pairs in the same environment as the pair under investigation, we reduce the effect of the sequence background on the significance of the covariation results. This environment-dependent analysis of interstrand pairs detects 8 % of the all-buried interstrand pairs and 20 % of exposed interstrand pairs as being statistically significant ($Z > 1.65$, p-value <0.05). Thus, even though the conservation of buried pairs is high, it is possible to detect significant covariation among these pairs. However, we found equal numbers of statistically significant covarying pairs among the interstrand (HB and NHB) buried and exposed pairs as among the intrastrand pairs in the same environments. These results are consistent with all other results described here, confirming that β-sheet pairs do not show outstanding evolutionary dependence.

Evolutionary coupling within protein families

While, in general, residue pairs on neighboring strands do not appear to be evolutionarily linked, it is possible that there are a few critical pairs in key locations that play an important role in determining the sheet structure of a given protein family. If only a few such pairs are required in every sequence, their presence may not be detected in a broad statistical analysis.

In 115 of 219 of the protein family alignments, we were able to detect one or more interstrand sheet pairs that showed statistically significant conservation and/or covariation. Highly significant covariation or conservation behavior of some residue pairs could be an indication of their structural importance. In the last few years, a number of studies that have sought to identify folding nuclei in different protein families have been reported (Kim *et al.*, 2000; McCallister *et al.*, 2000; and others reviewed by Plaxco *et al.*, 2000). We have investigated in detail the covariation pattern of the β-sheet pairs in two protein families, the src SH3 domain and fibronectin type III domain of tenascin, for which a large set of experimentally determined φ-values, as well as a large enough number of closely related non-redundant sequences were available (Riddle *et al.*, 1999; Hamill *et al.*, 2000b). In the src SH3 domain we found only two sheet pairs (one HB and one NHB) that showed significant covariation (both, interest-

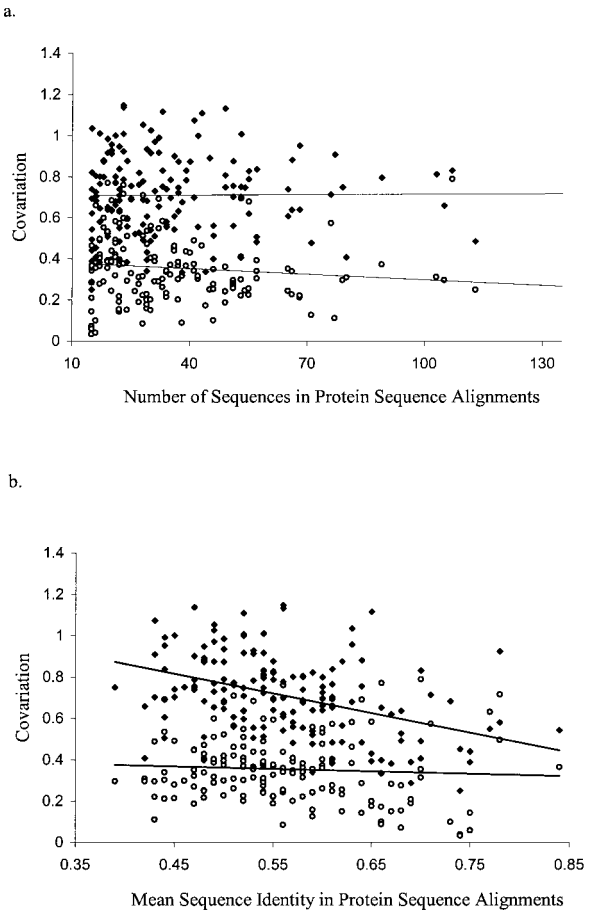

Figure 7. Average covariation (mutual information) values for exposed (◆) and buried (○) pairs in 157 sequence alignments as a dependence of (a) the number of sequences in alignment and (b) the mean pairwise sequence identity of alignment. Linear interpolation lines are plotted for buried and exposed pairs separately. The coefficient of determination (R^2) for the lines in (a) are 0.0000 for the exposed pairs and 0.0152 for the buried pairs. In (b), $R^2 = 0.1742$ for exposed pairs and 0.0040 for buried pairs.

ingly, are buried pairs). These were located in the distal β-hairpin (L44-T53) and in the sheet pair adjacent to the diverging turn (T7-L31). Interestingly, in the experimental study (Riddle *et al.*, 1999) these two regions (distal β-hairpin and the diverging turn) were found to be the most ordered structural elements in the transition state ensemble of the domain and were predicted to be involved in the nucleation of folding. In the TNfn3, the third fibronectin type III domain of human tenascin, we found five statistically significant covarying pairs. Unfortunately, none of the significant pairs that were found in our analysis had been mutated in the experimental study (Hamill *et al.*, 2000b), and no comparison can be made. However, none of the residues that were found to nucleate the folding of TNfn3 was found by us to covary. Our analysis so far could not detect significant correlation between covariation signals and residues that are involved in folding nucleation. Since the available experimental data are very limited, we cannot rule out the possibility that such correlation exists.

Conclusions

In an effort to better understand the driving forces behind β-sheet formation, we studied the conservation and covariation patterns of a large set of antiparallel β-sheet pairs. During the course of evolution, selective pressure on amino acid sequences is expected for maintaining protein function and stability as well as perhaps some degree of folding efficiency. Thus, if residue pairs on neighboring strands of the β-sheet are important for dictating or for maintaining the strand arrangement, we would expect to find evidence for such evolutionary pressure in sequence alignments of protein families. Our results did not reveal any substantial difference in the evolutionary conservation between the interstrand sheet pairs and non-interacting residues on the same β-strand. As expected, in both types of pairs we found a correlation between the evolutionary conservation of the pairs and the environment in which the pair residues, with buried pairs being much more conserved than exposed pairs. Similar results were obtained regardless of whether the average conservation of the two residues involved in the pair was calculated or if the conservation of the overall chemical properties of the pair was measured. We could not detect significant covariation patterns in the interstrand pairs, suggesting that compensatory mutations are no more common in interstrand pairs than in neighboring residues overall.

Despite having different residue pair preferences (Wouters & Curmi, 1995; Hutchinson *et al.*, 1998), the HB pairs and the NHB pairs showed the same degree of conservation and covariation in a large data set of sequence alignments (Figures 3, 4 and 6). To see whether correlated behavior may be strongest for the subset of residue pairs that form statistically favorable interactions, we compared the conservation of favorable residue pairs to the same amino acid pairs in non-interacting, $i,i+4$ positions. We found no evidence for the preferential conservation of favored pairs (Table 3). These results are consistent with the weak correlation of the pairing preferences between the earlier statistical studies. Since different pairing preferences between the HB and the NHB pairs are expected due to the different side-chain packing of these positions, our results imply that the packing geometries are maintained by other compensating mechanisms. It is possible that slight changes in the interstrand distances can accommodate sequence changes (Nagarajaram *et al.*, 1999) similar to the small helix displacements observed in globin protein family (Lesk & Chothia, 1980a).

Mirny & Shakhnovich, (1999) have argued that conservation of residues across non-homologous protein families that share a similar fold can be used to predict the locations of folding nuclei. Recent experimental studies, however, have found little correlation between a residue's evolutionary conservation and its participation in the folding pathway (Plaxco *et al.*, 2000; Hamill *et al.*, 2000b). This debate prompted us to investigate residue pair conservation and covariation within specific families for which there are experimental data on folding nucleation, since it is possible that correlated compensatory mutations would not be detected when analyzing the conservation of individual residues. While we did find some evidence of significant residue pair covariation in one family alignment, the available experimental data are still insufficient to rigorously examine the correlation between these significant pairs and the experimentally determined location of folding nucleation sites.

Overall, our results suggest that, unlike in RNA secondary structure, the specific identity of the neighboring residues on adjacent β-strands plays a minor role in dictating β-sheet formation. Protein structure prediction methods that rely on identifying patterns of concerted evolution are therefore unlikely to improve the prediction of β-sheet topology. However, since we did note a correspondence between the covariation of residue pairs and their involvement in folding nucleation in one of the two protein families we were able to examine, a few key pairs could guide the folding trajectory of β-sheet proteins. It is possible that specific residue-residue interactions are more critical in cases where protein-protein interactions are mediated by β-strand pairing.

Materials and Methods

Data sets

Representative antiparallel β-sheet domains were extracted from the SCOP database (Hubbard *et al.*, 1997) release 1.48. The web tool ASTRAL (Brenner *et al.*, 2000) was used to filter domains with no more than 40 % sequence identity. For each domain in the representative list, a subset of HSSP alignments around the representative sequence (Sander & Schneider, 1993), including only close homologues with >35 % identity, was extracted. The HSSP alignments were further refined by removing all sequences with >90 % identity. Only alignments with >15 sequences were included in our study (Table 1). Within any given alignment, positions with >20 % gaps were not analyzed.

Classification of interstrand pairs

The DSSP algorithm (Kabsch & Sander, 1983) was applied to all representative domains in our database. Residues were identified as belonging to β-sheets and classified according to whether they are hydrogen bonded (HB) or non-hydrogen bonded (NHB). Overall, 1560 HB and 1606 NHB pairs were analyzed, including a subset of 353 HB and 372 NHB pairs that were extracted from β-ribbons (two-stranded sheets) only. Intrastrand pairs displaced by one, two, three and four residues were also analyzed for comparison.

To characterize the environment of the pairs, the solvent exposure of the each residue in the representative structure was defined. Solvent exposure was calculated using the program ACCESS (http://www.cmpharm.ucsf.edu/ ~srp/), based on the Richmond & Richards (1978) algorithm, for calculating solvent-accessibility (*SA*). The *SA* values of each residue were normalized by the solvent exposure value of the specific amino acids (X) in Ala-X-Ala extended triplets (*SA* values for normalization were calculated using the same method). Residues with normalized $SA > 25($ were considered buried. The 25 % cutoff was chosen empirically.

Calculation of covariation

Two approaches were applied to measure the tendency of a pair of positions in a multiple sequence alignment to mutate in concert:

Approach 1

Covariation was calculated as described (Göbel *et al.*, 1994). Briefly, given an alignment with n sequences ($n > 15$), over a length of m residues, for each position in an alignment (from 1 to m), we created a mutation matrix describing all possible $n \times (n-1)$ amino acid-amino acid substitutions at that position (e.g. VL, VF, LF). The amino acid-amino acid substitutions of the mutation matrix were further translated to amino acid similarity scores, which were taken from the McLachlan (1971) 20×20 similarity matrix. This last step generates a distance matrix of substitution scores for each position in the alignment (see Göbel *et al.*, 1994). To evaluate the covariation for any given pair of positions i and j in the alignment, we calculated the correlation coefficient statistic (r_{ij}) between the values generated from matrix i and those generated from matrix j:

$$r_{ij} = \frac{1}{n^2} \sum \frac{W_{kl}(S_{ikl} - \langle S_i \rangle)(S_{jkl} - \langle S_j \rangle)}{\sigma_i \sigma_j} \quad (1)$$

The indices k and l represent the sequence number running from 1 to n. S_{ikl} is the element in the distance matrix of position i representing the similarity between the amino acids in sequence k and l in position i. Accordingly, S_{jkl} is the corresponding element in the distance matrix of position j. $\langle S_i \rangle$ and $\langle S_j \rangle$ are the mean values for the distance matrix i and j, respectively. And σ_i ,σ_j are the standard deviation of S_{ikl} and S_{jkl}. W_{kl} is the weighting factor that describes the overall similarity between sequence k and l, and is presented in order to down-weight information coming from similar pairs of protein. The final r_{ij} is a measure of the correlated mutation behavior between sequence position i and j.

Approach 2

The degree to which amino acid at position i and j covary was calculated based on their mutual information. This term was originally imported from information theory (Cover & Thomas, 1991) and has been applied in the past for calculating covariation in sequence alignments (Korber *et al.*, 1993; Clarke, 1995). The definition for covariation (mutual information) in our study was taken from Clarke (1995) and was calcu-

lated for all pairs in the alignments:

$$\sum_{a_i a_j} (Pa_i a_j)^2 \log_2 (Pa_i a_j / Pa_i Pa_j) \quad (2)$$

Each pair of residues, a_i and a_j, represent the 20 different amino acids. P_{ai} is the probability (observed frequency) of an amino acid type a_i in one position and $Pa_i a_j$ is the probability of the pair of amino acids a_i and a_j in the corresponding positions. $Pa_i a_j$ is therefore the observed frequency of the residue pairs in a given position, while the product $Pa_i Pa_j$ is the expected frequency of this type of pair if they were independent of one another.

As in the previous method, we used a weighting method to reduce bias due to overrepresentation of similar sequences. The overall sequence weights for each pair of sequences in the multiple sequence alignment (calculated as in approach 1) were used to calculate a weighted residue frequency for each position in the alignment (Henikoff & Henikoff, 1994). The weighted frequencies were applied in equation (2) instead of the original probabilities.

Calculation of conservation

Residue conservation

The distance matrices for individual residues in the alignments (previously created for calculating covariation) were used to sum all the amino acid-amino acid substitutions at a given position, resulting in a value of variability for that position. For calculating conservation, the elements of the distance matrices were normalized and the mean value was subtracted from 1.0.

Pair property conservation

Three properties were considered for calculation pair property conservation; volume, hydrophobicity and charge. Normalized van der Waals volumes for the amino acids were taken from Richards (1974). The partition coefficients between water and octanol of the *N*-acetyl amide groups (Fauchere & Pliska, 1983), normalized to range between 0 and 1, were used as the measure of hydrophobicity. Charges were assigned for R and K (+1), E and D (−1) and H (+0.5). The rest of the amino acids charges were set to 0. Based on these properties, all 210 possible amino acid combinations were characterized by a three-dimensional vector, each dimension of the vector representing a property; volume, hydrophobicity and charge. For each pair, the property values were defined as the average values of the two amino acids involved.

The distance between pairs of residues, by means of their chemical properties, was evaluated by calculating the Euclidean distance between the vectors (each pair represented by one vector). These substitution distances were averaged to calculate the mean distance for each pair of positions in our sequence alignments.

Statistical analysis

To evaluate the statistical significance of the covariation of the β-sheet pairs, we compared the covariation calculated for each specific pair to the average covariation of all possible pairs extracted from the same alignment. From each alignment, we first extracted all possible residue pairs. The pairs were then classified according to structural environment (buried, exposed or mixed) and the mean covariation and standard deviation for all pairs in each environment were computed. A standard *Z*-score (known also as the normal deviate) was calculated for each pair, comparing the covariation observed to the expected value given the environment of the pair. The statistical significance of the *Z*-score was determining by the proportion of the distribution (p) that lies beyond the given value.

Acknowledgments

We thank Melissa Cline for mathematical advice, helpful discussions and comments on the manuscript. We thank Alan Davidson for helpful comments on the manuscript. This work was supported by NIH grant GM52885 and by a California Division-American Cancer Society Fellowship.

References

Altschuh, D., Vernet, T., Berti, P., Moras, D. & Nagai, K. (1988). Coordinated amino acid changes in homologous protein families. *Protein Eng.* **2**, 193-199.

Benner, S. A. & Gerloff, D. (1991). Patterns of divergence in homologous proteins as indicators of secondary and tertiary structure: a prediction of the structure of the catalytic domain of protein kinases. *Advan. Enzyme. Regul.* **31**, 121-181.

Benner, S. A., Badcoe, I., Cohen, M. A. & Gerloff, D. L. (1994). *Bona-fide* prediction of aspects of protein conformation - assigning interior and surface residues from patterns of variation and conservation in homologous protein sequences. *J. Mol. Biol.* **235**, 926-958.

Bowie, J. U., Reidhaar-Olson, J. F., Lim, W. A. & Sauer, R. T. (1990). Deciphering the message in protein sequences - tolerance to amino acid substitutions. *Science,* **247**, 1306-1310.

Brenner, S. E., Koehl, P. & Levitt, R. (2000). The ASTRAL compendium for protein structure and sequence analysis. *Nucl. Acids Res.* **28**, 254-256.

Chelvanayagam, G., Eggenschwiler, A., Knecht, L., Gonnet, G. H. & Benner, S. A. (1997). An analysis of simultaneous variation in protein structures. *Protein Eng.* **10**, 307-316.

Clarke, N. D. (1995). Covariation of residues in the homeodomain sequence family. *Protein Sci.* **4**, 2269-2278.

Cover, T. M. & Thomas, J. A. (1991). *Elements of Information Theory* (Schilling, D. L., ed.), John Wiley and Sons, Inc., New York.

de Alba, E., Rico, M. & Jiménez, M. A. (1999). The turn sequence directs beta-strand alignment in designed beta-hairpins. *Protein Sci.* **8**, 2234-2244.

Dickerson, R. E. (1971). The structures of cytochrome *c* and the rates of molecular evolution. *J. Mol. Evol.* **1**, 26-45.

Dill, K. A. (1985). Theory for the folding and stability of globular proteins. *Biochemistry,* **24**, 1501-1509.

Eddy, S. R. & Durbin, R. (1994). RNA sequence analysis using covariance models. *Nucl. Acid. Res.* **22**, 2079-2088.

Eisenberg, D., Schwarz, E., Komaromy, M. & Wall, R. (1984). Analysis of membrane and surface protein

sequences with the hydrophobic moment plot. *J. Mol. Biol.* **179**, 125-142.
Fauchere, J. & Pliska, V. (1983). Hydrophobic parameters of amino acid-side-chains from the partitioning of n-acetyl-amino-acid amides. *Eur. J. Med. Chem.* **18**, 369-375.
Gerloff, D. L., Cohen, F. E., Korostensky, C., Turcotte, M., Gonnet, G. H. & Benner, S. A. (1997). A predicted consensus structure for the N-terminal fragment of the heat shock protein HSP90 family. *Proteins: Struct. Funct. Genet.* **27**, 450-458.
Göbel, U., Sander, C., Schneider, R. & Valencia, A. (1994). Correlated mutations and residue contacts in proteins. *Proteins,* **18**, 309-317.
Gunasekaran, K., Ramakrishnan, C. & Balaram, P. (1997). Beta-hairpins in proteins revisited: lessons for de novo design. *Protein Eng.* **10**, 1131-1141.
Hamill, S. J., Cota, E., Chothia, C. & Clarke, J. (2000a). Conservation of folding and stability within a protein family: the tyrosine corner as an evolutionary cul-de-sac. *J. Mol. Biol.* **295**, 641-649.
Hamill, S. J., Steward, A. & Clarke, J. (2000b). The folding of an immunoglobulin-like Greek key protein is defined by a common-core nucleus and regions constrained by topology. *J. Mol. Biol.* **297**, 165-178.
Henikoff, S. & Henikoff, J. G. (1994). Position-based sequence weights. *J. Mol. Biol.* **243**, 574-578.
Hubbard, T. J. P., Murzin, A. G., Brenner, S. E. & Chothia, C. (1997). SCOP: a structural classification of proteins database. *Nucl. Acids Res.* **25**, 236-239.
Hutchinson, E. G., Sessions, R. B., Thornton, J. M. & Woolfson, D. N. (1998). Determinants of strand register in antiparallel beta-sheets of proteins. *Protein Sci.* **7**, 2287-2300.
Juan, V. & Wilson, C. (1999). RNA secondary structure prediction based on free energy and phylogenetic analysis. *J. Mol. Biol.* **289**, 935-947.
Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers,* **22**, 2577-2637.
Kamtekar, S., Schiffer, J. M., Xiong, H., Babik, J. M. & Hecht, M. H. (1993). Protein design by binary patterning of polar and nonpolar amino acids. *Science,* **262**, 1680-1685.
Kauzmann, W. (1959). Some factors in the interpretations of protein denaturation. *Advan. Protein Chem.* **14**, 1-63.
Kim, D. E., Fisher, C. & Baker, D. (2000). A breakdown of symmetry in the folding transition state of protein L. *J. Mol. Biol.* **298**, 971-984.
Korber, B. T., Farber, R. M., Wolpert, D. H. & Lapedes, A. S. (1993). Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. *Proc. Natl Acad. Sci. USA,* **90**, 7176-7180.
Kortemme, T., Ramírez-Alvarado, M. & Serrano, L. (1998). Design of a 20-amino acid, three-stranded beta-sheet protein. *Science,* **281**, 253-256.
Koshi, J. M. & Goldstein, R. A. (1997). Mutation matrices and physical-chemical properties: correlations and implications. *Proteins,* **27**, 336-344.
Lee, B. & Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **55**, 379-400.
Lesk, A. M. & Chothia, C. (1980a). How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *J. Mol. Biol.* **136**, 225-270.
Lesk, A. M. & Chothia, C. (1980b). Solvent accessibility, protein surfaces, and protein folding. *Biophys. J.* **32**, 35-47.
Lifson, S. & Sander, C. (1980). Specific recognition in the tertiary structure of beta-sheets of proteins. *J. Mol. Biol.* **139**, 627-639.
Martinez, J. C. & Serrano, L. (1999). The folding transition state between SH3 domains is conformationally restricted and evolutionarily conserved. *Nature Struct. Biol.* **6**, 1010-1016.
McCallister, E. L., Alm, E. & Baker, D. (2000). Critical role of beta-hairpin formation in protein G folding. *Nature Struct. Biol.* **7**, 669-673.
McLachlan, A. D. (1971). Tests for comparing related amino-acid sequences. Cytochrome *c* and cytochrome *c* 551. *J. Mol. Biol.* **61**, 409-424.
Mirny, L. A. & Shakhnovich, E. I. (1999). Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. *J. Mol. Biol.* **291**, 177-196.
Nagarajaram, H. A., Reddy, B. V. & Blundell, T. L. (1999). Analysis and prediction of inter-strand packing distances between beta-sheets of globular proteins. *Protein Eng.* **12**, 1055-1062.
Olmea, O. & Valencia, A. (1997). Improving contact predictions by the combination of correlated mutations and other sources of sequence information. *Fold. Des.* **2**, S25-S32.
Ortiz, A. R., Kolinski, A., Rotkiewicz, P., Ilkowski, B. & Skolnick, J. (1999). Ab initio folding of proteins using restraints derived from evolutionary information. *Proteins,* **Suppl 3**, 177-185.
Pazos, F., Helmer-Citterich, M., Ausiello, G. & Valencia, A. (1997). Correlated mutations contain information about protein-protein interaction. *J. Mol. Biol.* **271**, 511-523.
Plaxco, K. W., Larson, S., Runczinski, I., Riddle, D. S., Thayer, E. C., Buchwitz, B., Davidson, A. R. & Baker, D. (2000). Evolutionary conservation in protein folding kinetics. *J. Mol. Biol.* **298**, 303-312.
Richards, F. M. (1974). The interpretation of protein structures: total volume, group volume distributions and packing density. *J. Mol. Biol.* **82**, 1-14.
Richmond, T. J. & Richards, F. M. (1978). Packing of alpha-helices: geometrical constraints and contact areas. *J. Mol. Biol.* **119**, 537-555.
Riddle, D. S., Grantcharova, V. P., Santiago, J. V., Alm, E., Ruczinski, I. & Baker, D. (1999). Experiment and theory highlight role of native state topology in SH3 folding. *Nature Struct. Biol.* **6**, 1016-1024.
Sander, C. & Schneider, R. (1993). The HSSP data base of protein structure-sequence alignments. *Nucl. Acids Res.* **21**, 3105-3109.
Shenkin, P. S., Erman, B. & Mastrandrea, L. D. (1991). Information-theoretical entropy as a measure of sequence variability. *Proteins,* **11**, 297-313.
Shindyalov, I. N., Kolchanov, N. A. & Sander, C. (1994). Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Eng.* **7**, 349-358.
Smith, C. K. & Regan, L. (1995). Guidelines for protein design: the energetics of beta sheet side-chain interactions. *Science,* **270**, 980-982.
Thomas, D. J., Casari, G. & Sander, C. (1996). The prediction of protein contacts from multiple sequence alignments. *Protein Eng.* **9**, 941-948.
West, M. W. & Hecht, M. H. (1995). Binary patterning of polar and non polar amino acids in sequences

and structures of native proteins. *Protein Sci.* **4**, 2032-2039.

Wouters, M. A. & Curmi, P. M. (1995). An analysis of side-chain interactions and pair correlations within antiparallel beta-sheets: the differences between backbone hydrogen-bonded and non-hydrogen-bonded residue pairs. *Proteins: Struct. Funct. Genet.* **22**, 119-131.

Zaremba, S. M. & Gregoret, L. M. (1999). Context-dependence of amino acid residue pairing in antiparallel beta-sheets. *J. Mol. Biol.* **291**, 463-479.

Edited by J. Thornton

(Received 19 July 2000; received in revised form 22 November 2000; accepted 23 November 2000)