The Dependence of Amino Acid Pair Correlations on Structural Environment

Adrian P. Cootes,[1] **Paul M.G. Curmi,**[2] **Ross Cunningham,**[3] **Christine Donnelly,**[3] **and Andrew E. Torda**[1]*
[1]*Research School of Chemistry, The Australian National University, Canberra, Australia*
[2]*Initiative in Biomolecular Structure, School of Physics, The University of New South Wales, Sydney, Australia*
[3]*Statistical Consulting Unit, The Australian National University, Canberra, Australia*

ABSTRACT **A statistical analysis was performed to determine to what extent an amino acid determines the identity of its neighbors and to what extent this is determined by the structural environment. Log-linear analysis was used to discriminate chance occurrence from statistically meaningful correlations. The classification of structures was arbitrary, but was also tested for significance. A list of statistically significant interaction types was selected and then ranked according to apparent importance for applications such as protein design. This showed that, in general, nonlocal, through-space interactions were more important than those between residues near in the protein sequence. The highest ranked nonlocal interactions involved residues in β-sheet structures. Of the local interactions, those between residues *i* and *i* + *2* were the most important in both α-helices and β-strands. Some surprisingly strong correlations were discovered within β-sheets between residues and sites sequentially near to their bridging partners. The results have a clear bearing on protein engineering studies, but also have implications for the construction of knowledge-based force fields. Proteins 32:175–189, 1998.**

Key words: pairwise statistics; secondary structure; nonlocal interactions

INTRODUCTION

It is still not yet known how a protein sequence determines its own fold. Consequently, attempts to design sequences to fold to a specified structure have only shown promise recently.[1] Central to solving these problems is the need to determine which intramolecular interactions make the largest contributions to the specificity of a sequence for its native conformation. Theoretical analyses of lattice models[2–5] and experiments[6–9] have suggested that nonlocal interactions generally make a greater contribution to a sequence's structural specificity than do local interactions, although there is some evidence to the contrary,[10] at least when considering events at the protein surface.[11,12] However, a more detailed determination of those interactions of greatest significance to real proteins is necessary if the important problems of fold recognition and sequence design are to be solved.

The aim of this paper is to rank the types of amino acid pairwise interaction in order of importance via a statistical analysis of the protein structure database. Viewed anthropomorphically, it can be asked to what extent does an amino acid determine its neighbor and to what extent does the resulting pair determine its environment class. This approach is physically naive, but it should avoid many assumptions about what are the most important contributions to protein composition.

Statistical analysis of proteins has a long history. Structures have been studied extensively for residue propensities in various physical environments[13–15] and for significant factors in amino acid substitutions in structural homologs.[16–18] However, there has been relatively little statistical analysis for significant factors in pairwise interactions.[19–22,40]

We applied log-linear analysis[23–26] of pairwise amino acid statistics to determine both their significance and relative dependence on structural environment. This is a more general approach than is used for knowledge-based force fields and does not rely on statistical mechanics for its derivation.[41–42] Log-linear analysis determines whether variables are dependent on each other by constructing a model that assumes independence of those variables and assesses the fit of that model to the data. The discrepancy of the model from the data is quantified by a measure, termed the "mean deviance," which is calculated from a χ^2 distributed, log-likelihood ratio statistic.

The validity of a classification of pairwise amino acid interactions with respect to a series of structural variables, such as secondary structure, can be tested using log-linear analysis and quantified by a mean deviance statistic ($\overline{D_{123}}$). This should yield a final classification of apparently independent interaction classes.

*Correspondence to: Andrew Torda, Research School of Chemistry, The Australian National University, Canberra ACT 0200, Australia. E-mail: Andrew.Torda@anu.edu.au

Received 7 November 1997; Accepted 16 March 1998

Each interaction class can be tested for dependence of the residues on their partner residue using a less general mean deviance statistic ($\overline{D_{12}}$). If there is a sufficiently high probability that the pair of residues for that interaction class are dependent, or associated, then that class is significant. Those interaction classes can be assessed for the degree of association between amino acids, or the relative ability of one residue to determine the identity of its pair residue, using an association measure ($\overline{d_{12}}$). Simple hypotheses can also be tested on those remaining classes, such as whether or not the association between the pair of amino acids is symmetrical about those amino acids ($\overline{D_{sym}}$).

In our study, an arbitrary decision was made to divide proteins into secondary structures and tabulate data within and between these elements. This classification may be quite arbitrary, but its validity can be tested. For example, this analysis can quantitatively state whether it is valid to collect separate statistics for α-helices and β-sheets. Continuing in this vein, it can be determined whether it is profitable to further divide β-sheet data into parallel and anti-parallel categories. Applying this strategy further, the importance of peptide chain directionality can also be assessed, a feature included in some protein design and fold recognition methods and neglected in others.

MATERIALS AND METHODS

Protein Set

The set of Protein Data Bank[27] structures used for analysis was a subset of the list from Hobohm et al.,[28] August release, 1996. The structures were selected such that there was less than 25% sequence identity between any two proteins. Structures that consisted only of α-carbon coordinates and of those complexed with nucleotides were not selected. This resulted in 494 chains (listed in Appendix A).

Structural Definitions

Counts were collected for each possible pair of amino acid types and then separated into a series of categories. Categories resulting in an average of less than three counts per residue pair were not considered. The first categories were the structural motifs defined by the DSSP program:[29] α-helix, β-sheet, loop, turn, 3–10-helix, and isolated β-bridge. These were further subdivided according to pair separation *n* along the amino acid sequence. Counts for separations *n* of up to five residues were obtained for each secondary structure type. Between β-strands, pairs were counted for each residue and its bridging partner and each residue and its various diagonal partners (Fig. 1). Finally, residue pairs within an arbitrary cutoff spatial separation were counted, as defined below.

α-*Helices*

The interaction classes considered for helices were similar to those studied by Klinger and Brutlag,[21] except that a distinction was made between "edge" and "internal" residues within a helix. The helix edge residues were defined as the first and last four residues (the first and last turn) in each helix. No distinction was made between N- and C-most edges of the helix in this study. The remaining residues were defined to be internal helix residues. Each possible combination of pairs of edge and internal residues were considered.

β-*Strands*

Tallies for residues in β-strands were also separated into edge and internal categories but using a definition appropriate to β-sheets. Edge pairs were defined as those in which both residues had one or less bridge partners, and internal pairs were those in which both residues have two bridge partners. Edge pairs were also partitioned with respect to their strands environment into "parallel," in which the edge strand formed a parallel ladder with the remainder of the sheet, "antiparallel," and "no bridge," where the β-strand was isolated.

Internal pairs were also further divided with respect to strand environment into "both parallel," "both antiparallel," and "mixed," where the pairs were on strands participating in two parallel sheet ladders, two antiparallel ladders, or one of each, respectively.

Division of β-strand pairs into edge and internal categories occurred for sequence separations of up to four. The total number of counts for sequence separations above this were too low to give reliable statistics for each subcategory. Likewise, sequence separations of more than two could have no further subdivision of the edge and internal categories due to low counts.

Loops

Those regions defined by DSSP to have no backbone hydrogen-bonded structure were counted as loops. Individual loop residues were grouped into "low" and "high" backbone curvature categories. High curvature residues were those that were either defined by DSSP to be at the center of a bend or were displaced less than three residues from a bend center. Low curvature residues were all those loop residues that did not meet the above requirement. Each possible combination of pairs of low and high residues were considered.

Cross-strand β-*sheet interactions*

Cross-strand pair counts were collected in a similar fashion to Hubbard,[30] but with greater structural detail. Pairs were collected of the form $i \rightarrow j + n$, where *i* is a β-sheet residue on the N-most strand of the ladder formed by the adjacent strands, *j* is the

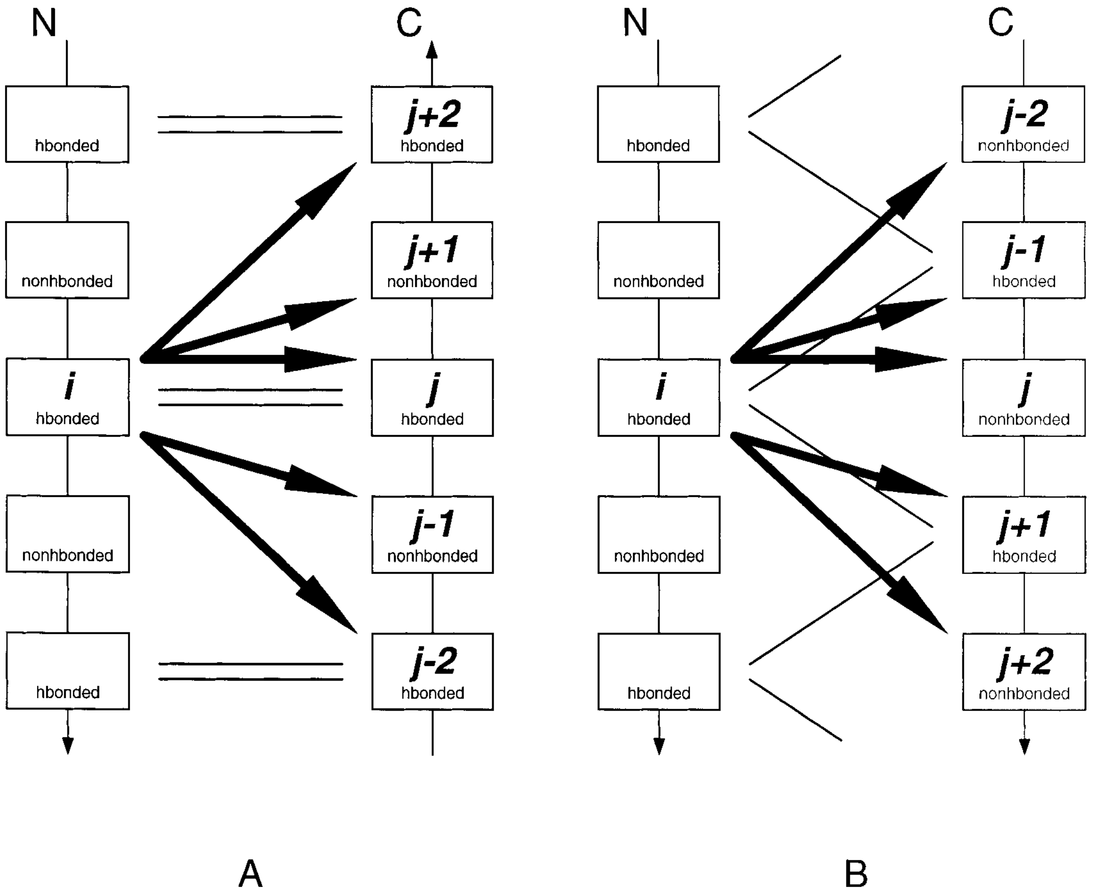

Fig. 1. Representations of cross-strand interactions in β-sheets. The different hydrogen bonding classes are (**A**) antiparallel β-sheet and (**B**) parallel β-sheet. Interaction classes are depicted with respect to a hydrogen bonded residue *i* on the left strand. A similar set of interactions was also considered using a nonhydrogen bonded residue as the reference residue. Each box represents one residue and thick arrows the interactions considered in Results. Thin lines between strands represent hydrogen bonds.

bridge partner of residue *i* and is on the C-most strand, and *n* the number of residues separated along the strand from the bridge partner, neglecting bulges (Fig. 1).

Cross-strand pairs were partitioned into those that participate in parallel or antiparallel ladders. Edge and internal residues were defined as in Wouters and Curmi,[22] with edge residues defined to be those that had only one bridge partner and internal residues were defined to be those with two. "H-bonded" residues were those that participated in hydrogen bonding within the ladder considered. "Non-H-bonded" were those that did not meet this requirement (Fig. 1). Hydrogen bonding between backbone C=O and N-H groups was defined according to the criterion used by the DSSP program.[29] To classify residues into h-bonded or non-h-bonded categories within a β-sheet ladder, both the type of ladder and the DSSP-defined hydrogen bonding of backbone groups had to be considered. For antiparallel ladders, a residue is considered to be h-bonded if both the donor and acceptor backbone groups are hydrogen bonded to that residue's bridge partner; it is non-h-bonded otherwise. For parallel sheets, the residue is considered to be h-bonded if the donor and acceptor backbone groups are hydrogen bonded to residues adjacent to the bridge partner; it is non-h-bonded otherwise.

Bulge residues that are inserted between residues participating in the ladder but that do not participate in the ladder themselves are avoided for cases where $n \neq 0$. This is achieved by searching along the C-most strand participating in that ladder from the bridge partner until the residue with the appropriate hydrogen bonding class (h-bonded or non-h-bonded) expected for that type of ladder and that separation *n* in a "bulgeless" sheet is found.

All possible combinations of edge/internal and h-bonded/non-h-bonded properties for residue pairs were considered for antiparallel ladders. Only combinations of h-bonded/non-h-bonded properties were considered for parallel ladders.

Sidechain contacts

Those amino acid pairs that were not considered in the secondary structure-dependent local interactions (with the exception of β-strand local interaction classes with sequence separation of greater than

TABLE I. Sidechain Radii r^{sc} for Each Amino Acid

Amino acid	r^{sc} (Å)
Arginine	2.63
Tryptophan	2.57
Tyrosine	2.37
Lysine	2.27
Phenylalanine	2.03
Methionine	1.86
Glutamine	1.86
Glutamate	1.84
Histidine	1.83
Isoleucine	1.68
Leucine	1.64
Asparagine	1.59
Aspartate	1.57
Valine	1.37
Threonine	1.33
Proline	1.31
Cysteine	0.91
Serine	0.71
Alanine	0.00
Glycine	0.00

two) or the cross-strand β-sheet interactions, but were spatially local, were considered separately as a sidechain contact interaction class. The ranges of sequence separation within each type of secondary structure element in which pairs are said to be "local" were ±2 residues within β-strands and ±5 residues within all other secondary structures.

In order to calculate sidechain contacts, a simple definition was constructed which could later be applied using only the coordinates of backbone atoms. First, an ideal sidechain centroid position was calculated for each type of amino acid (except for glycine, in which the C^α atom was defined as the "sidechain centroid"). For each occurrence of a residue type within the test set, the geometric mean of the sidechain heavy atoms was calculated using a frame of reference defined by the backbone atoms N, C^α, and C. Next, for each type of amino acid a sidechain radius, r^{sc}, was calculated by surveying each occurrence of the amino acid in the database. The average and standard deviation of the distance from each sidechain atom to the residue's centroid was calculated. The radius was taken as the average + one standard deviation ($r^{SC}_{glycine}$ was defined to be zero) and values are given in Table I.

Given this construction, a nonlocal contact between residues i and j was recognized, as shown in Figure 2, when the distance between the calculated sidechain centroids was less than $r_i^{SC} + r_j^{SC} + 4\text{Å}$. The value of 4Å was an arbitrary cutoff.

Each interaction type was partitioned with respect to the secondary structure element in which each amino acid was found. Each amino acid was classified into α-helix, β-sheet, loop, turn, and "other." Each pair was then classified into α-helix–α-helix, α-helix–β-sheet, and so on. The category "other" is the combination of the remaining secondary structure types not already included. These categories were combined due to their relatively low occurrence in the dataset.

Log-Linear Model for Assessing Significant Structural Variables

This kind of model can be formalized by defining a three-dimensional matrix where two of the variables are amino acid identities and the third is the structural variable of interest. The observed frequency of each combination of these three variables is given by N_{ijk}, where i and j correspond to amino acid types I and J, the identities of the amino acid variables 1 and 2, respectively. k corresponds to the structural class K, the identity of the structural variable 3. The number of amino acid types, N_{aa}, is the number of categories for variables 1 and 2, and N_{str}, the number of structural classes, is the number of classes of variable 3. The expected frequency E_{ijk} of each combination of variables 1, 2, and 3, assuming a model of independence of amino acid interaction from the structural variable, is given by:

$$\ln E_{ijk} = u_0 + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{23(jk)} + u_{13(ik)} \quad (1)$$

where

$$u_0 = \frac{\sum_i \sum_j \sum_k \ln E_{ijk}}{N_{aa} \cdot N_{aa} \cdot N_{str}} \quad (2)$$

representing a "baseline count" and

$$u_{1(i)} = \frac{\sum_j \sum_k \ln E_{ijk}}{N_{aa} \cdot N_{str}} - u_0 \quad (3)$$

$u_{2(j)}$ and $u_{3(k)}$ are similar terms. These account for the relative distributions of the ith category of variable 1, the jth category of variable 2, and the kth category of variable 3, respectively.

$$u_{12(ij)} = \frac{\sum_k \ln E_{ijk}}{N_{aa}} - u_{1(i)} - u_{2(j)} - u_0 \quad (4)$$

and $u_{23(jk)}$ and $u_{13(ik)}$ are similar terms. These account for the dependence of the ith category of variable 1 on the jth category of variable 2, the jth category of variable 2 on the kth category of variable 3, and the ith category of variable 1 on the kth category of variable 3, respectively.

Each of the parameters for this model, and hence the expected frequencies E_{ijk}, were obtained by an iterative reweighted least squares fitting procedure using the Genstat statistical package.[31] The fit of the model was evaluated by measuring the discrepancy of the observed values N_{ijk} from the expected values

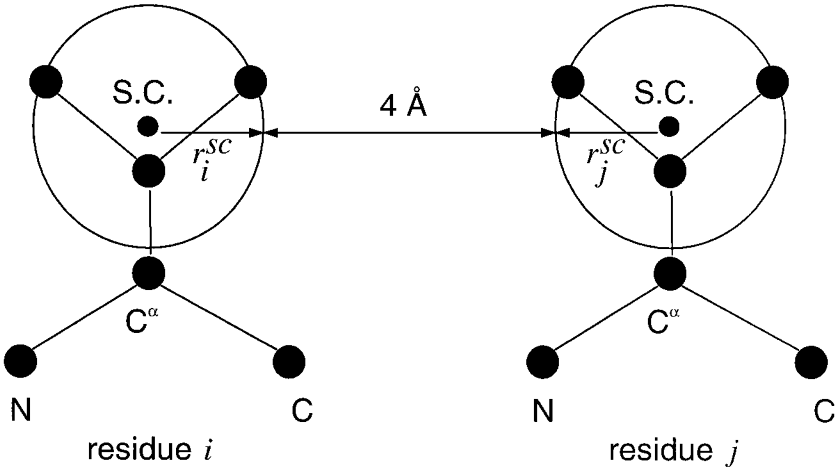

Fig. 2. Definition of nonlocal sidechain contacts. r_i^{SC} and r_j^{SC} are sidechain radii of *i* and *j* given for each residue in Table I. Sidechain centroid position (S.C.) is the average center calculated for each type of residue, as described in Materials and Methods. The 4 Å distance represents the cutoff criterion used for sidechain contacts.

E_{ijk}. The deviance D_{123} is a likelihood ratio statistic, which follows a χ^2 distribution with v_{123} degrees of freedom, where:

$$D_{123} = 2 \sum_i \sum_j \sum_k N_{ijk} \ln\left(\frac{N_{ijk}}{E_{ijk}}\right) \tag{5}$$

and

$$v_{123} = (N_{aa} - 1)(N_{aa} - 1)(N_{str} - 1) \tag{6}$$

The mean deviance, $\overline{D_{123}}$, is given by:

$$\overline{D_{123}} = \frac{D_{123}}{v_{123}} \tag{7}$$

and is used as the test statistic for the fit of the above model.

Interaction Significance

The level of confidence in an interaction class' pairwise residue dependence can be assessed by fitting a model similar to that of the previous section but which contains only two variables (the two amino acids). The modeled expectation frequencies E_{ij} and the mean deviance statistic $\overline{D_{12}}$ are calculated in an analogous fashion to that indicated in Equations 1–7 but without considering degrees of freedom over structural states k and with no model term for the dependence between amino acids u_{12}.

Pairwise Amino Acid Association

To isolate the degree of association between interacting amino acids, the mean deviance $\overline{D_{12}}$ is scaled by the total number of pair counts ($\overline{D_{12}}$ is proportional to the total number of pair counts for the model considered here). This gives an appropriate association measure, the scaled mean deviance $\overline{d_{12}}$:

$$\overline{d_{12}} = \frac{\overline{D_{12}}}{N_{12}} \tag{8}$$

where N_{12} is the total number of pair counts.

Interaction Symmetry

To check the significance of interaction symmetry, D_{sym} is calculated in a similar fashion to $\overline{D_{12}}$ with the exception that the constraint:

$$u_{12(ij)} = u_{12(ji)} \tag{9}$$

for all i and j is added to the model and the degrees of freedom v_{sym} is now given by:

$$\mathrm{M}_{sym} = \frac{(N_{aa} - 1)\,(N_{aa} - 2)}{2} \tag{10}$$

Pair Distances

The average distance between pairs of residues participating in cross-strand interaction classes were calculated and are given in full in Appendix B. The distance x^{ca} for each interaction class was evaluated by averaging the distance between the α-carbons of each pair participating in that interaction class. The distance x^{sc} was found by averaging the distances between each of the residues sidechain "shells" (defined previously) for each pair.

RESULTS

We begin by considering the structural classes (α-helix, β-sheet...) and assessing whether they affect the distribution of amino acids pair types. This leads to a selection of structural classes which have a statistically meaningful effect. Finally, these are ranked according to the degree to which each amino acid appears to determine the type of its partner.

Significant Structural Variables in Interactions

Initially, residue pairs were divided into the broad categories of local and nonlocal depending on sequence separation. The nonlocal pairs were further subdivided into cross-strand interactions where the

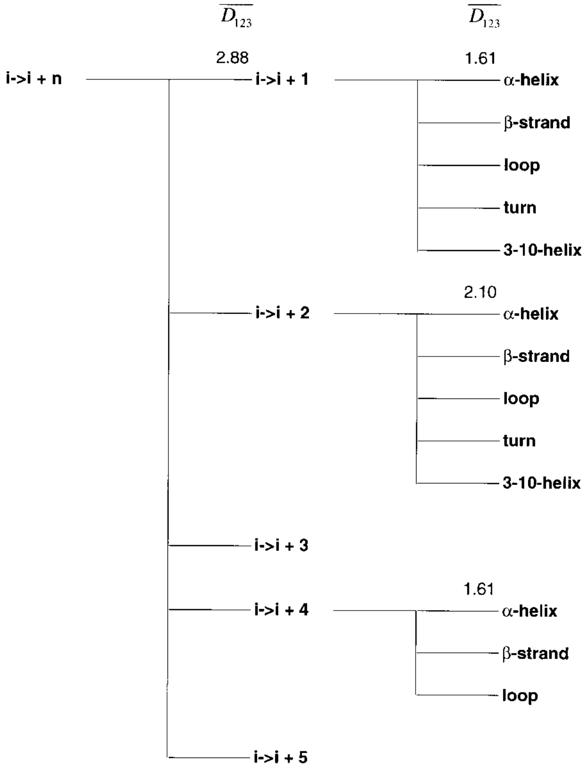

Fig. 3. Significant classifications of local interactions. Each branch shows the classification into structural categories. Only significant classifications are shown. Significance was determined by the mean deviance, $\overline{D_{123}}$ shown above each branch. A cutoff value of 1.5 was used to select for significance.

residues within the pair were located on neighboring strands in a β-sheet, and sidechain contact for cases where the pair of residues did not reside in a single secondary structure element but were proximal in space.

Within these three categories, residue pairs were subdivided into structural classes, as described below. Each structural division was assessed for statistical significance by calculating the appropriate mean deviance $\overline{D_{123}}$. This gives a measure of confidence in the model (that the structure does not affect pairwise correlation). A $\overline{D_{123}} = 1$ corresponds to a 50% probability that the pairwise correlations are independent of structure. A large $\overline{D_{123}}$, usually set at 1.5, indicates a probability of less than 10^{-8} and a distinctly significant structural class. Those classifications with $\overline{D_{123}} > 1.5$ are shown in Figures 3, 4, and 5. All remaining classifications are listed in Appendix C.

Local interactions

Local interactions were defined as those between residues $i \rightarrow i + n$, where n is some small number and both i and $i + n$ were within a single secondary structural element. Local interactions were analyzed for the effect of sequence separation, secondary structure, and subsecondary structures on pairwise amino acid interactions. Definitions of all categories are given in Materials and Methods.

The dependence of local interaction classes on sequence separation, n, was examined for separations of up to five residues. If we take $\overline{D_{123}} > 1.5$ as an indicator of significance, then the $\overline{D_{123}}$ for the dependence of amino acid interactions on sequence separation n was found to be greater than the minimum value required for significance (Fig. 3). Thus, local interactions were found to have significant dependence on sequence separation.

For each sequence separation n, the effect of secondary structure type on residue interactions was examined. The $i \rightarrow i + 1$, $i \rightarrow i + 2$, and $i \rightarrow i + 4$ interaction classes were found to be most dependent on secondary structure (Fig. 3). However, the $i \rightarrow i +$ 3 and i $\rightarrow i + 5$ classes were relatively independent of secondary structure ($\overline{D_{123}} = 1.3$ and $\overline{D_{123}} = 1.2$, respectively).

There was only weak dependence of pairwise interaction on any of the subsecondary structure variables tested for any separation n (all $\overline{D_{123}} < 1.5$, Appendix C).

Cross-strand interactions

Cross-strand interactions were defined as those interactions between pairs of amino acids located on adjacent strands within β-sheets (Fig. 1). Cross-strand interactions were analyzed for the effect of ladder type (parallel or antiparallel), hydrogen bonding, strand positioning in the sheet (whether each strand was edge or internal to the sheet), and sequence separation n from bridge partners on pairwise amino acid interactions (Fig. 1).

As expected, the dependence of these pairwise cross-strand interactions on the size of $|n|$ was found to be highly significant (Fig. 4), but for $|n| = 1$ and $|n| = 2$, the effect of the direction of separation along the strand was found to be insignificant ($\overline{D_{123}} = 1.1$ and $\overline{D_{123}} = 1.3$, respectively).

For bridge partner pairs ($n = 0$), the distinction between antiparallel pairs and parallel pairs is not significant (Fig. 4). However, there is significant variation between the separate hydrogen bonding categories for both the antiparallel (h-bonded–h-bonded and non-h-bonded–non-h-bonded) and parallel (h-bonded–non-h-bonded and non-h-bonded–h-bonded) cases. Antiparallel pairs also exhibited dependence on strand positioning in the sheet (edge–edge, edge–internal, internal–edge, and internal–internal categories) but less so than the dependence on hydrogen bonding.

For pairs separated from the bridge partner by one residue ($|n| = 1$), the division between antiparallel and parallel pairs was found to be insignificant (Fig. 4). However, this masks one useful classification. Of the antiparallel strands, there is significance in separating pairs based on chain direction ($n = 1$ vs.

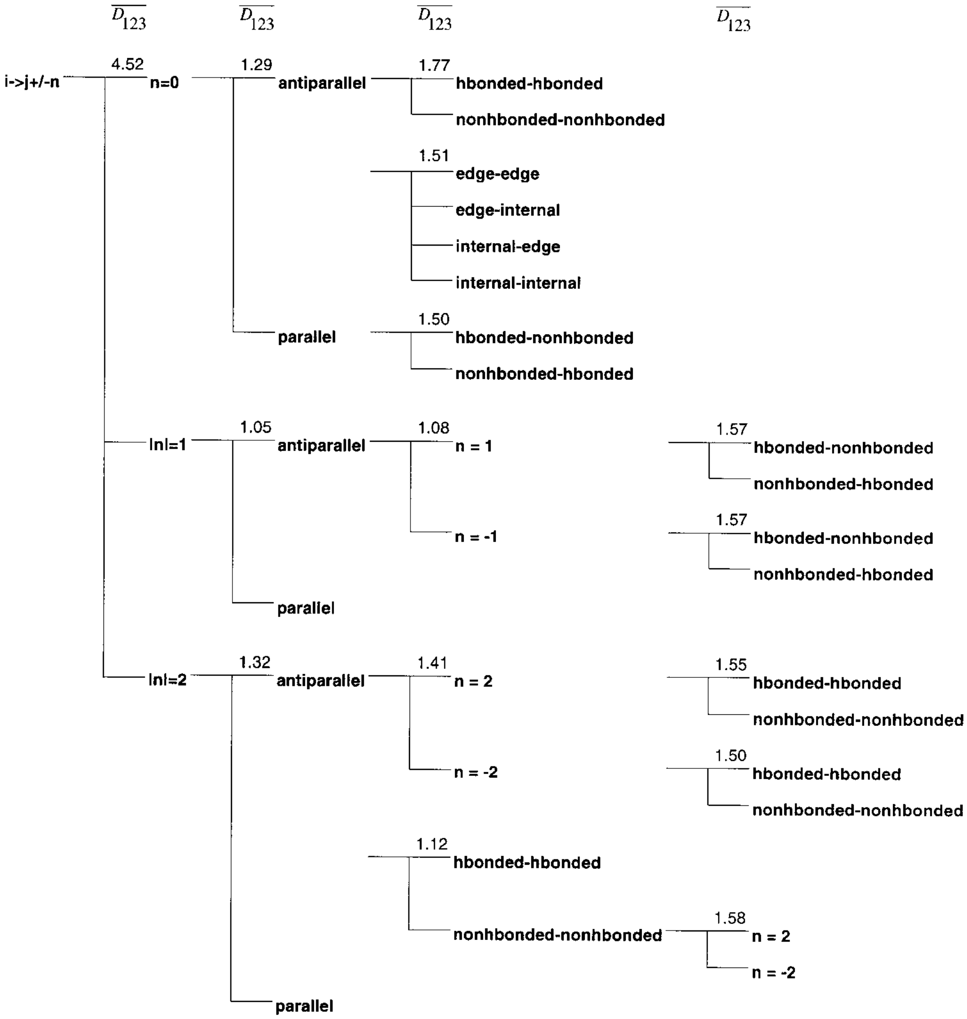

Fig. 4. Partitioning of cross-strand interactions in β-sheets. Interactions are initially divided according to types *i*→ *j* ± *n*, where *j* is the bridging partner of *i* and *n* is the distance along the sequence from *j*. These are subsequently divided into parallel and antiparallel sheet types. These are further subdivided into subclasses which still show statistical significance. Layout and labels as in Figure 3.

$n = -1$) and further by hydrogen bonding classes. This can be explained by considering Figure 1a. The pattern of hydrogen bonds viewed from residue *i* is quite different from that seen from a non-h-bonded position one residue up or down on the strand.

Similarly, one can consider pairs in antiparallel strands with $|n| = 2$. As with the $|n| = 1$ interactions, the division by chain direction and hydrogen bonding class was significant. Furthermore, from Figure 4 it can be seen that data should be collected separately for h-bonded–h-bonded and non-h-bonded–non-h-bonded pairs. The non-h-bonded–non-h-bonded pairs should then be classified according to chain direction ($n = 2$ vs. $n = -2$) .

Sidechain contact interactions

Sidechain contact interactions were those between amino acids not within the secondary structure-dependent local sequence range of each other (see Materials and Methods) or considered a cross-strand interaction but that have spatially proximal sidechains (described in Materials and Methods) (Fig. 2). Sidechain contact interactions were assessed for the effect of secondary structure type on their pairwise residue distributions. Chain direction was not taken into account, so interaction pair distributions of the form X-Y (where both X and Y are one of the secondary structure groups α-helix, β-strand, loop, turn, or other) are simply the transpose of the pair distribution for the interaction class Y-X.

Division of pairwise contact interactions into secondary structure categories, dependent on the secondary structure of just one of the amino acids, was found to be significant (Fig. 5). Further division of each of these categories with respect to the second-

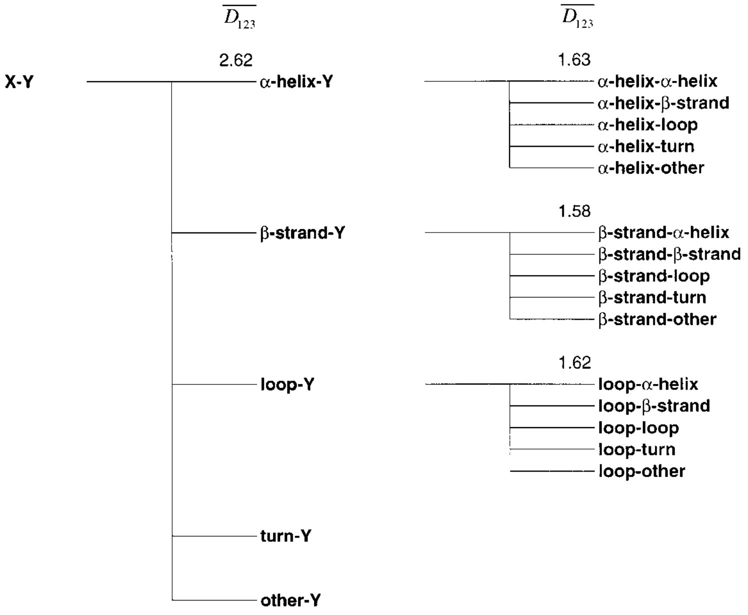

Fig. 5. Significant classifications of the nonlocal sidechain contact interactions. X and Y represent the five secondary structure classes used. Layout and labels as in Figure 3.

ary structure of the other amino acid was significant for the α-helix-Y, β-strand-Y and loop-Y categories (Fig. 5) but not the turn-Y or other-Y categories ($\overline{D_{123}} = 1.3$ and $\overline{D_{123}} = 1.1$, respectively). It could be argued that the pairwise associations are merely reflecting tendencies in distances between secondary structure elements. In fact, the average pairwise distance between sidechain shells (within the 4 Å cutoff) was between 2.4 Å and 2.6 Å for all secondary structure classes, although the distributions may differ. The data suggests that statistics for pairwise interactions should be treated differently for different secondary structure contexts.

Significant Interaction Classes and Ranking Interaction Classes

Those structural variables found to have $\overline{D_{123}}$ of greater than or equal to 1.5 in the previous section can be deemed to constitute a list of pairwise interaction classes, each with distinct pairwise preferences. However, this has not shown that the pairwise distributions within each class are not merely the result of noise. This was tested using the measure $\overline{D_{12}}$. The classes were then ranked by their $\overline{D_{12}}$ (Table II) where a $\overline{D_{12}}$ of 1.5 corresponds to a probability of about 10^{-8}.

This could be seen as a first method for detecting classifications which are statistically reliable and classes in which residues influence their neighbors in a statistical sense.

The classes with the highest significance values were typically those with the highest number of total counts. This is because $\overline{D_{12}}$ reflects both the quantity of data and the strength of amino acid association. To isolate the effect of amino acid association, the classes with $\overline{D_{12}} > 1.5$ were ranked by scaled mean deviance $\overline{d_{12}}$, as defined in Materials and Methods. This means that the candidate classes are statistically significant and ranked by their apparent physical importance judged by the extent to which members of residue pairs influence each other. This ranking is given in Table III.

Nonlocal interaction classes (sidechain contact and cross-strand classes) were generally ranked higher than local interaction classes. Of the nonlocal interaction classes, those between only β-strand residues (cross-strand and β-strand–β-strand sidechain contacts) were ranked higher than those between residues in other secondary structures.

Cross-strand interaction classes were labeled as $i \rightarrow j \pm n$ where j is the bridge partner of residue i and n is the displacement along the sequence from j. The parallel sheet $i \rightarrow j$ bridge partner interaction classes were ranked higher than the antiparallel sheet $i \rightarrow j$ interaction classes. This could be because parallel β-sheets place bridge partner residue sidechains closer than antiparallel sheets, but this is not generally the case (Appendix B).

The $i \rightarrow j$ bridge partner interaction classes were ranked higher than those interaction classes between residues displaced along the strand from the other's bridge partner ($i \rightarrow j + n$, where $n = -2,-1,1,2$), with the exception of the antiparallel sheet $i \rightarrow j - 2$ interaction class where both i and $j - 2$ residues are not hydrogen bonded to residues in that ladder (non-h-bonded–non-h-bonded). The sidechain shells participating in this class are separated on average by 3.0 Å (Appendix B). This distance is comparable with those of the bridge partner interaction classes, which are separated on average by 2.4 Å and 2.8 Å for the non-h-bonded–non-h-bonded and h-bonded–h-bonded cases, respectively. The $i \rightarrow j - 2$ (non-h-bonded–non-h-bonded) intershell distance is also much less than the distance for

TABLE II. Interaction Significance $\overline{D_{12}}$ and Symmetry $\overline{D_{sym}}$

Interaction	$\overline{D_{12}}$[a]	$\overline{D_{sym}}$[b]	Count[c]
Contact (α-helix–loop)	10.18	2.54	44160
Contact (turn-Y)	10.16	1.97	56324
Contact (α-helix–α-helix)	8.74	n/a	28976
Contact (β-strand–loop)	8.71	2.11	40640
Contact (α-helix–β-strand)	7.04	1.47	25100
Local i → i + 2 (α-helix)	6.99	1.27	30096
Contact (other-Y)	6.31	1.18	31938
Contact (β-strand–β-strand)	6.28	n/a	15044
Local i → i + 4 (α-helix)	4.96	1.73	23136
Contact (loop–loop)	4.66	n/a	21268
X-strand i → j* (antiparallel, nonhbonded–nonhbonded)	3.59	1.02	6380
Local i → i + 3	3.22	1.72	42441
Local i → i + 1 (loop)	2.76	1.48	25344
X-strand i → j*-2 (antiparallel, nonhbonded–nonhbonded)	2.68	1.47	3572
X-strand i → j* (antiparallel, hbonded–hbonded)	2.49	1.13	6396
Local i → i + 5	2.42	1.09	26639
Local i → i + 1 (β-strand)	2.41	0.96	15665
Local i → i + 1 (α-helix)	2.37	1.97	33488
X-strand i → j* (antiparallel, internal–internal)	2.09	1.00	3047
X-strand i → j* (antiparallel, internal–edge)	2.06	1.13	3380
X-strand i → j* (antiparallel, edge–internal)	2.04	1.03	3395
Local i → i + 2 (β-strand)	1.96	0.91	8546
X-strand i → j* (antiparallel, edge–edge)	1.94	1.29	2955
X-strand i → j* (parallel, nonhbonded–hbonded)	1.93	1.10	2535
X-strand i → j* (parallel, hbonded–nonhbonded)	1.77	1.21	2437
X-strand i → j*-1 (antiparallel, hbonded–nonhbonded)	1.72	1.08	4802
X-strand i → j* + 1 (antiparallel, hbonded–nonhbonded)	1.66	0.85	5028
X-strand i → j* ± 2 (antiparallel, hbonded–hbonded)	1.57	1.10	7249
Local i → i + 1 (turn)	1.56	1.39	7680
X-strand i → j* ± 2 (parallel)	1.52	1.21	4783
X-strand i → j* + 1 (antiparallel, nonhbonded–hbonded)	1.45	1.21	4799
X-strand i → j* ± 1 (parallel)	1.42	1.04	7296
Local i → i + 1 (3–10-helix)	1.40	1.53	3329
Local i → i + 2 (loop)	1.35	1.17	16704
Local i → i + 2 (turn)	1.35	1.28	1869
Local i → i + 2 (3–10-helix)	1.23	0.98	1923
X-strand i → j* − 1 (antiparallel, nonhbonded–hbonded)	1.22	0.84	5036
Local i → i + 4 (β-strand)	1.16	1.12	2273
Local i → i + 4 (loop)	1.14	1.04	7561
X-strand i → j* + 2 (antiparallel, nonhbonded–nonhbonded)	1.11	1.16	3589

*j = β-bridge partner.

[a] $\overline{D_{12}}$ is interaction significance and [b] $\overline{D_{sym}}$ is interaction symmetry as defined in Materials and Methods.

[c] Count is the number of observations within each interaction type. For local and cross-strand interactions classes where both interaction partners have identical structural properties, except for relative sequence direction, asymmetry reflects the effect of chain direction. For all other cases, asymmetry reflects the partitioning of each interaction partner into the different environments as well as chain direction.

TABLE III. Interactions Ranked by Pairwise Residue Association

Interaction	$\overline{d_{12}}$ $(\times 10^{-4})$[a]
X-strand i → j* (parallel, nonhbonded–hbonded)	7.61
X-strand i → j*-2 (antiparallel, nonhbonded-nonhbonded)	7.50
X-strand i → j* (parallel, hbonded-nonhbonded)	7.26
X-strand i → j* (antiparallel, internal-internal)	6.86
X-strand i → j* (antiparallel, edge-edge)	6.57
X-strand i → j* (antiparallel, internal-edge)	6.09
X-strand i → j* (antiparallel, edge-internal)	6.01
X-strand i → j* (antiparallel, nonhbonded-nonhbonded)	5.63
Contact (β-strand–β-strand)	4.17
X-strand i → j* (antiparallel, hbonded-hbonded)	3.89
X-strand i → j* − 1 (antiparallel, hbonded-nonhbonded)	3.58
X-strand i → j* + 1 (antiparallel, hbonded-nonhbonded)	3.30
X-strand i → j* ± 2 (parallel)	3.18
Contact (α-helix–α-helix)	3.02
Contact (α-helix–β-strand)	2.80
Local i → i + 2 (α-helix)	2.32
Contact (α-helix–loop)	2.31
Local i → i + 2 (β-strand)	2.29
Contact (loop–loop)	2.19
X-strand i → j* ± 2 (antiparallel, hbonded-hbonded)	2.17
Contact (β-strand–loop)	2.14
Local i → i + 4 (α-helix)	2.14
Local i → i + 1 (turn)	2.03
Contact (other-Y)	1.97
Contact (turn-Y)	1.80
Local i → i + 1 (β-strand)	1.54
Local i → i + 1 (loop)	1.09
Local i → i + 5	0.91
Local i → i + 3	0.76
Local i → i + 1 (α-helix)	0.71

*j = β-bridge partner.

[a] $\overline{d_{12}}$ refers to pairwise residue association, as defined in Materials and Methods.

the antiparallel $i \rightarrow j - 2$ (h-bonded–h-bonded) interaction class, whose average intershell distance is 6.9 Å, and the $i \rightarrow j + 2$ (non-h-bonded–non-h-bonded) interaction class, whose average intershell distance is 4.5 Å (Appendix B).

Of the local interaction classes, α-helix and β-strand $i \rightarrow i + 2$ classes were comparable and had the highest residue associations. The α-helix $i \rightarrow i + 2$ interaction class was comparable with the $i \rightarrow i + 4$ α-helix class. Generally, α-helix local interaction classes were ranked lower than the α-helix nonlocal classes. The $i \rightarrow i + 1$ interaction classes within loops, β-strands, and especially turns have higher residue associations than the $i \rightarrow i + 3$ and $i \rightarrow i + 5$ classes.

Interaction Symmetry

Once it has been determined that an interaction class exhibits dependence between amino acids, a simple extension of the model used in the previous section can be fitted to the data to determine whether or not that dependence is symmetric. That is, is the pair IJ just as favorable or as unfavorable as the pair JI for all possible I and J?

Significant interaction classes were tested for their interaction symmetry by evaluating $\overline{D_{sym}}$ for each interaction class (Table II). The mean deviance $\overline{D_{sym}}$ provides an indicator of the interaction class' symmetry, increasing with the asymmetry of the pairwise interactions. A $\overline{D_{sym}}$ of 1.5 corresponds to a probability of about 10^{-5} that the interaction class is symmetric under the assumptions discussed previously. Apart from the nonlocal sidechain contact interaction classes, those classes where the two residue types are structurally symmetric, the asymmetry measures sensitivity of amino acid interactions to chain direction.

Almost all the cross-strand interaction classes were found to be symmetric, although the antiparallel $i \rightarrow j - 2$ (non-h-bonded–non-h-bonded) class may be significantly asymmetric. Local interaction classes, however, varied in symmetry properties. The α-helix $i \rightarrow i + 1$ and $i \rightarrow i + 4$ interaction classes were quite sensitive to chain direction but the $i \rightarrow i + 2$ class was not. The β-sheet $i \rightarrow i + 1$ and $i \rightarrow i + 2$ interaction classes were both found to be reasonably symmetric, as was the loop $i \rightarrow i + 1$. The $i \rightarrow i + 5$ interaction class was symmetric, while the $i \rightarrow i + 3$ class was not.

Nonlocal sidechain contact interaction classes between residues from identical secondary structure groupings are symmetric by construction. Those between dissimilar secondary structure categories are asymmetric except for α-helix–β-strand contacts (which almost meet the criterion for asymmetry) and the other-Y contacts.

DISCUSSION

The structure to which a sequence folds appears to be largely determined by nonlocal interactions. This is a statistical observation and may or may not have any bearing on the order of events in protein folding and possible initial assembly of secondary structure elements. This result is consistent with experiment[6–9] and theoretical studies of lattice models.[2–5] It may not agree with other proposals.[10–12] This has serious implications for protein design. When allocating a residue to a given position in a target structure, it is more important that this residue interacts favorably with other residues that are spatially close but sequentially distant than with sequentially near residues.

Of the nonlocal interaction classes, those involving residues only within β-strands (cross-strand and β-strand–β-strand contacts) were found to be generally more important than those involving residues in other secondary structures. This result is consistent with previous studies[6–8] indicating that the identity of residues found on β-strands has a greater dependence on tertiary context than residues found on other secondary structures. For cross-strand interaction classes, those involving bridge partner residues are the most important. Of these, the parallel interaction classes are more important than antiparallel classes, possibly because parallel sheets tend to be buried within a structure to a greater extent than antiparallel sheets.[32] For example, bridge partners with opposite charge within parallel sheets are more strongly associated than those in antiparallel sheets in order to avoid destabilizing isolated charges in the protein interior. The relative strength of the antiparallel bridge partner (non-h-bonded–non-h-bonded) interaction class compared to the antiparallel bridge partner (h-bonded–h-bonded) class is probably simply due to the geometry of the antiparallel sheet. The non-h-bonded–non-h-bonded pair's average sidechain–sidechain distance is approximately 0.5 Å less than that of the h-bonded–h-bonded pair (Appendix B).

The importance of cross-strand interaction classes generally decreases with the increase in the separation of each residue from the bridge partner of the other (that is, $i \rightarrow j$ interaction classes are more important than $i \rightarrow j \pm 1$, which are more important than $i \rightarrow j \pm 2$). The exception to this rule is the antiparallel $i \rightarrow j - 2$ (non-h-bonded–non-h-bonded) class which is more important than all but one $i \rightarrow j$ cross-strand interaction class. This is surprising for two reasons.

First, a residue in a sheet might be expected to be most influenced by the closest residue in the paired strand (residue j, not $j - 2$). Second, the interaction is distinctly directional. This means that, statistically, the $i \rightarrow j - 2$ interaction class is far more important than the $i \rightarrow j + 2$ class. Although, intuitively, this is not reasonable, it can be explained geometrically. Given the operational definition of sidechain radii (under Materials and Methods), the average sidechain–sidechain distance for those residue pairs in the $i \rightarrow j - 2$ (non-h-bonded–non-h-bonded) interaction class is 3.0 Å. This is close to 2.4 Å and 2.8 Å, the distances found for the $i \rightarrow j$ non-h-bonded–non-h-bonded and h-bonded–h-bonded interaction classes, respectively (Appendix B). This reasoning can also explain the directionality of the interaction. The average sidechain–sidechain distance for those residue pairs in the $i \rightarrow j + 2$ (non-h-bonded–non-h-bonded) interaction class is 4.5 Å. The most significant amino acid pairs for the $i \rightarrow j - 2$ (non-h-bonded–non-h-bonded) interaction class are ion pairs and large aromatic residues based on a Pearson χ^2 statistic (Appendix D). These types of amino acid pairs are also typical of the most significant pairs found in antiparallel $i \rightarrow j$ interaction classes, which have been shown to result from direct physical interaction between sidechains.[22] In more general terms, this result is consistent with a twisting of antiparallel β-sheet and a shearing of adjacent antiparallel β-strands toward their C-termini,[22] bringing residues i and $j - 2$ close together in space, perhaps in order to optimize sidechain packing in the sheet.

Of the local interaction classes considered, the α-helix $i \rightarrow i + 2$ and the β-strand $i \rightarrow i + 2$ interaction classes are the most important. The α-helix $i \rightarrow i + 2$ interaction class was found to be comparable with the α-helix $i \rightarrow i + 4$ interaction class, indicating that the identity of a residue in an α-helix is influenced by the residue on the opposite side of the α-helix to the same extent as the residue adjacent to it. This is a surprising result, as the $i \rightarrow i + 2$ interaction class is thought to simply partition residues of opposite hydropathy on either side of the α-helix, whereas the $i \rightarrow i + 4$ class is thought to favor residue pairs of like hydropathy but also involves direct physical interaction between sidechains.[21] This result indicates that partitioning of residues on opposite sides of an α-helix is stronger than local physical interaction within the α-helix. β-strand local interaction classes, however, behave as might be expected, with $i \rightarrow i + 2$ class (both residues on the same side of the sheet) ranking higher in importance than the $i \rightarrow i + 1$ class (residues on opposite sides of the sheet).

Although these results are objectively correct, they must be interpreted with some caution. On the basis of the statistics, α-helix $i \rightarrow i + 2$ pairs seem to be slightly more important than α-helix $i \rightarrow i + 4$ pairs. Unfortunately, they are not truly independent. Using similar analytical techniques and the same significance threshold described in this work, it was found that the $i \rightarrow i + 2 \rightarrow i + 4$ statistics can be explained by the $i \rightarrow i + 2$ and $i + 2 \rightarrow i + 4$ interactions, without considering the $i \rightarrow i + 4$ interactions (although the counts were slightly below the threshold representation required in this study). This implies that a large part of the $i \rightarrow i + 4$ statistics can be attributed to $i \rightarrow i + 2$ interactions alone. This is just a particularly clear example of an endemic problem. Each apparently independent pairwise association may actually reflect indirect effects of third (and even higher order) neighbors.

For some sequence separations (1, 2, and 4) local interactions were found to be dependent on secondary structure, but interactions with other sequence separations (3 and 5) were not. It may be that the representation of the $i \rightarrow i + 5$ interaction is too low in each secondary structural category to distinguish between them. However, lack of representation should not be the reason for the independence of the $i \rightarrow i + 3$ interaction class from secondary structure. For α-helices, the $i \rightarrow i + 3$ pairwise amino acid distribution should reflect physical interaction between sidechains of similar hydropathy.[21] The β-strand $i \rightarrow i + 3$ pairwise distribution has been shown to reflect the tendency for the partitioning of polar and nonpolar residues onto opposite sides of the β-sheet.[21] Thus, the $i \rightarrow i + 3$ interactions within α-helices, β-strands, and loops might have been expected to be distinguishable to a similar level of significance as interactions considered between amino acids with other sequence separations. However, it is not dependent on secondary structure to the same degree as other local interactions.

From Table III, nonlocal interaction types involving α-helices are significant, but are less important than those involving β-strands. Like β-strand residues, they also exhibit a greater dependence on spatially close, non-local residues than they do on local residues. The α-helix–α-helix, α-helix–β-strand and α-helix–loop sidechain contact interaction classes are all more important, or comparable to, the most important α-helix local interaction class.

Nonlocal sidechain contact interactions are secondary structure-dependent. Thus, the different geometrical arrangements of residues about nonlocal contact residue pairs (which are found in a fixed spatial relationship to low resolution, as described in Methods), corresponding to each type of secondary structure that those residues occur in, appear to influence how that pair interacts. This result suggests that interactions between the pair of residues and the surrounding environment are quite important in determining nonlocal sidechain contact interactions.

The results here covered about 150 structural classes, but this is only a fraction of the possibilities and ignores features such as helix capping.[33,34] Table III, however, shows some of the most important interactions which should be taken into account in protein design and fold recognition. There are also implications for knowledge-based force fields.[35–39] For example, the results show where it is useful to collect data respecting peptide chain direction and where this is not justified. The results also suggest that nonlocal residue pairs interact differently, depending on secondary structure context. This may reflect the fact that the pairwise approach in this work neglects higher order interactions.[40]

Residue pair correlation statistics, including those used in this study, can be obtained at http://www.rsc.anu.edu.au/~cootes/protein_stats.html, as well as Appendices A, B, C, and D.

REFERENCES

1. Dahiyat, B.I., Mayo, S.L. De novo protein design: Fully automated sequence selection. Science 278:82–87, 1997.
2. Thomas, P.D., Dill, K.A. Local and nonlocal interactions in globular proteins and mechanisms of alcohol denaturation. Protein Sci. 2:2050–2065, 1993.
3. Abkevich, V.I., Gutin, A.M., Shakhnovich, E.I. Impact of local and non-local interactions on thermodynamics and kinetics of protein folding. J. Mol. Biol. 252:460–471, 1995.
4. Dill, K.A., Bromberg, S., Yue, K. et al. Principles of protein folding — A perspective from simple exact models. Protein Sci. 4:561–602, 1995.
5. Govindajaran, S., Goldstein, R.A. Optimal local propensities for model proteins. Proteins 22:413–418, 1995.

6. Minor, D.L. Jr., Kim, P.S. Context is a major determinant of β-sheet propensity. Nature 371:264–267, 1994.
7. Smith, C.K., Regan, L. Guidelines for protein design: The energetics of β-sheet sidechain interactions. Science 270: 980–982, 1995.
8. Minor, D.L. Jr., Kim, P.S. Context-dependent secondary structure formation of a designed protein sequence. Nature 380:730–734, 1996.
9. Muñoz, V., Cronet, P., López-Hernandez, E., Serrano, L. Analysis of the effect of local interactions on protein stability. Fold. Des. 1:167–178, 1996.
10. Rooman, M.J., Kocher, J.-P.A., Wodak, S.J. Extracting information on folding from the amino acid sequence: Accurate predictions for protein regions with preferred conformation in the absence of tertiary interactions. Biochemistry 31:10226–10238, 1992.
11. Gilis, D., Rooman, M. Stability changes upon mutation of solvent-accessible residues in proteins evaluated by database-derived potentials. J. Mol. Biol. 257:1112–1126, 1996.
12. Gilis, D., Rooman, M. Predicting protein stability changes upon mutation using database-derived potentials: Solvent accessibility determines the importance of local versus non-local interactions along the sequence. J. Mol. Biol. 272:276–290, 1997.
13. Chou, P.Y., Fasman, G.D. Prediction of protein conformation. Biochemistry 13:222–245, 1974.
14. Gamier, J., Osguthorpe, D.J., Robson, B. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. J. Mol. Biol. 120:97–120, 1978.
15. Muñoz, V., Serrano, L. Intrinsic secondary structure propensities of the amino acids, using statistical (-(matrices: Comparison with experimental scales. Proteins 20:301–311, 1994.
16. Overington, J., Johnson, M.S., Šali, A., Blundell, T.L. Tertiary structural constraints on protein evolutionary diversity: Templates, key residues and structure prediction. Proc. R. Soc. Lond. B 241:132–145, 1990.
17. Overington, J., Donelly, D., Johnson, M.S., Šali, A., Blundell, T.L. Environment-specific amino acid substitution tables: Tertiary templates and prediction of protein folds. Protein Sci. 1:216–226, 1992.
18. Tomii, K., Kanehisa, M. Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. Protein Eng. 9:27–36, 1996.
19. von Heijne, G., Blomberg, C. The β structure: Inter-strand correlations. J. Mol. Biol. 117:821–824, 1977.
20. Lifson, S., Sander, C. Specific recognition in the tertiary structure of β-sheets of proteins. J. Mol. Biol. 139:627–639, 1980.
21. Klinger, T.M., Brutlag, D.L. Discovering structural correlations in alpha-helices. Protein Sci. 3:1847–1857, 1994.
22. Wouters, M.A., Curmi, P.M.G. An analysis of side chain interactions and pair correlations within antiparallel β-sheets: The difference between backbone hydrogen-bonded and non-hydrogen bonded residue pairs. Proteins 22:119–131, 1995.
23. Bishop, Y.M.M., Fienberg, S.E., Holland, P.W. "Discrete Multivariate Analysis." Cambridge, MA: The MIT Press, 1975.
24. Everitt, B.S. "The Analysis of Contingency Tables." London: Chapman and Hall, 1977.
25. Fingleton, B. "Models of Category Counts." Cambridge, UK: Cambridge University Press, 1984.
26. Andersen, E.B. "The Statistical Analysis of Categorical Data." Berlin: Springer-Veriag, 1990.
27. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B. et al. The Protein Data Bank: A computer-based archival file for macromolecular structures. Eur. J. Biochem 80:319–324, 1977.
28. Hobohm, U., Scharf, M., Schneider, R., Sander, C. Selection of representative protein data sets. Protein Sci. 1:409–417, 1992.
29. Kabsch, W., Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637, 1983.
30. Hubbard, T.J.P. Use of β-strand interaction pseudo-potentials in protein structure prediction and modelling. In: "Proceedings of the Biotechnology Computing Track, Protein Structure Prediction Minitrack of the 27th HICSS." Lathrop, R.H., ed. IEEE Computer Society Press. 1994:336–354.
31. Payne, R.W. "Genstat 5 Reference Manual." New York: Oxford University Press, 1989.
32. Richardson, J.S. β-sheet topology and the relatedness of proteins. Nature 268:495–500, 1977.
33. Presta, L.G., Rose, G.D. Helix signals in proteins. Science 240:1632–1641, 1988.
34. Richardson, J.S., Richardson, D.C. Amino acid preferences for specific locations at the ends of alpha helices. Science 240:1648–1652, 1988.
35. Sippl, M.J. Knowledge-based potentials for proteins. Curr. Opin. Struct. Biol. 5:229–235, 1995.
36. Böhm, G. New approaches in molecular structure prediction. Biophys. Chem. 59:1–32, 1996.
37. Jones, D.T., Thornton, J.M. Potential energy functions for threading. Curr. Opin. Struct. Biol. 6:210–216, 1996.
38. Jernigan, R.L., Bahar, I. Structure-derived potentials and protein simulations. Curr. Opin. Struct. Biol. 6:195–209, 1996.
39. Torda, A.E. Perspectives in protein fold recognition. Curr. Opin. Struct. Biol. 7:200–205, 1997.
40. Munson, P.J., Singh, R.K. Statistical significance of hierarchical multi-body potentials based on Delaunay tessellation and their application in sequence-structure alignment. Protein Sci. 6:1467–1481, 1997.
41. Sippl, M.J. Calculations of conformational ensembles from potentials of mean force: An approach to the knowledge-based prediction of local structures in globular proteins. J. Mol. Biol. 213:859–883, 1990.
42. Jones, D.T., Taylor, W.R., Thornton, J.M. A new approach to protein fold recognition. Nature 358:86–89, 1990.

APPENDIX A. The List of Protein Chains From Which Statistics Were Collected

131L_, 153L_, 193L_, 1AAF_, 1AAK_, 1ABRB, 1ADEA, 1ADMA, 1ADT_, 1AEP_, 1AERA, 1AFB1, 1AMG_, 1AMP_, 1AORA, 1AOZA, 1APS_, 1ARB_, 1ARS_, 1ARV_, 1ASH_, 1ATE_, 1ATLA, 1ATPE, 1BAM_, 1BBPA, 1BBT1, 1BBT2, 1BBT3, 1BCFA, 1BCO_, 1BDMB, 1BEC_, 1BGC_, 1BGLA, 1BGW_, 1BIA_, 1BIP_, 1BNCB, 1BNDA, 1BNH_, 1BP2_, 1BPB_, 1BRIC, 1BRLA, 1BTL_, 1BUCA, 1BVP1, 1BW4_, 1BYB_, 1CAUA, 1CAUB, 1CCR_, 1CDOA, 1CEAA, 1CELA, 1CEO_, 1CEWI, 1CFB_, 1CHD_, 1CHKA, 1CHMA, 1CID_, 1CKIA, 1CKSB, 1CLC_, 1CMBA, 1CNSA, 1CPCA, 1CPCB, 1CPT_, 1CRL_, 1CSEI, 1CSH_, 1CSMA, 1CTN_, 1CTT_, 1CUS_, 1CYG_, 1CYU_, 1CYX_, 1DAAA, 1DAR_, 1DDT_, 1DEAA, 1DHR_, 1DHX_, 1DIH_, 1DLC_, 1DLHA, 1DLHB, 1DNPA, 1DOI_, 1DPB_, 1DPGA, 1DPPA, 1DSBA, 1DTR_, 1DTS_, 1DUPA, 1DVRA, 1DYNA, 1DYR_, 1ECA_, 1ECL_, 1ECPA, 1EDE_, 1EFT_, 1EPAB, 1ESC_, 1ESFB, 1ESL_, 1ETC_, 1EXH_, 1FBAA, 1FBR_, 1FC2D, 1FCDA, 1FCDC, 1FIM_, 1FJMA, 1FKJ_, 1FKX_, 1FNC_, 1FNF_, 1FPS_, 1FRPA, 1FRUA, 1FUJA, 1FVL_, 1GADO, 1GARA, 1GCA_, 1GCB_, 1GHR_, 1GKY_, 1GLCG, 1GLN_, 1GMFA, 1GOF_, 1GPB_, 1GPC_, 1GPH1, 1GPMA, 1GPR_, 1GRJ_, 1GSA_, 1GTQA, 1HAN_, 1HAR_, 1HBQ_, 1HC4_, 1HCE_, 1HCNA, 1HCNB, 1HDCA, 1HFH_, 1HGEA, 1HJRA, 1HLB_, 1HMPA, 1HMT_, 1HMY_, 1HNGA, 1HPM_, 1HQAA, 1HRM_, 1HSLA, 1HTMD, 1HTP_, 1HUCB, 1HUW_, 1HVD_, 1HVKA, 1HXN_, 1I1B_, 1IAE_, 1ICEA, 1ICEB, 1IDO_, 1ILK_, 1INP_, 1IRK_, 1IRL_, 1ISCA, 1ITG_, 1JAPA, 1JCV_, 1JELP, 1KIFA, 1KNB_, 1KPBA, 1KPTA, 1LBA_, 1LCPA, 1LCT_, 1LENA, 1LGR_, 1LIS_, 1LKI_, 1LKKA, 1LLO_, 1LPBB, 1LPE_, 1LPT_, 1LTDA, 1LTSA, 1LTSD, 1LXA_, 1LYLA, 1MAL_, 1MAT_, 1MHCA, 1MHLA, 1MHLC, 1MINA, 1MINB, 1MKAA, 1MLA_, 1MLDA, 1MML_, 1MMOB, 1MMOD, 1MMOG, 1MOLA, 1MPP_, 1MRJ_, 1MSAA, 1MSC_, 1MUCA, 1MUP_, 1MUT_, 1MXA_, 1NAL1, 1NAR_, 1NBAA, 1NCFA, 1NCHA, 1NDH_, 1NEF_, 1NFP_, 1NHKL, 1NHP_, 1NIF_, 1NIPA, 1NNC_, 1OACA, 1OMP_, 1ONC_, 1OROB, 1OVAA, 1OXA_, 1OYC_, 1PBA_, 1PBE_, 1PBGA, 1PBN_, 1PBXA, 1PCO_, 1PCRH, 1PDA_, 1PDGA, 1PEA_, 1PFKA, 1PGS_, 1PHG_, 1PHR_, 1PII_, 1PKM_, 1PKP_, 1PLQ_, 1PLS_, 1PMA1, 1PMAA, 1PNE_, 1PNKA, 1PNKB, 1POC_, 1POXA, 1POY1, 1PRCC, 1PRCM, 1PRR_, 1PRTA, 1PRTB, 1PRTD, 1PRTF, 1PSDA, 1PTD_, 1PTVA, 1PTX_, 1PUT_, 1PVC1, 1PVC2, 1PVDA, 1PXTB, 1PYAB, 1PYP_, 1QORA, 1QPG_, 1QUK_, 1RCB_, 1RCF_, 1RCI_, 1RCPA, 1REC_, 1REGX, 1RFBA, 1RIBA, 1RPA_, 1RRGA, 1RSY_, 1RTG_, 1RTP1, 1SACA, 1SAT_, 1SBP_, 1SCHA, 1SCUA, 1SCUB, 1SESA, 1SLTB, 1SLUA, 1SMD_, 1SMNA, 1SNC_, 1SRA_, 1SRIB, 1STD_, 1SVA1, 1SVB_, 1SVR_, 1TAG_, 1TAHA, 1TAM_, 1TCA_, 1TFS_, 1THTA, 1THV_, 1THX_, 1TIV_, 1TLK_, 1TML_, 1TNRA, 1TPG_, 1TPLA, 1TRKA, 1TRY_, 1TSP_, 1TSSA, 1TTBA, 1TYS_, 1IUBSA, 1UBSB, 1UKZ_, 1UMUA, 1VCAA, 1VCC_, 1VHH_, 1VHRA, 1VID_, 1VIL_, 1VMOA, 1VSD_, 1VSGA, 1WAS_, 1WBC_, 1WDCC, 1WHTA, 1WHTB, 1XAA_, 1XNB_, 1XYZA, 1YPTB, 1YUA_, 256BA, 2ABD_, 2ABK_,

APPENDIX A. (Continued)

2ACG_, 2ACQ_, 2ALP_, 2AYH_, 2AZAA, 2BBKH, 2BGU_, 2BLTA, 2BRD_, 2BTFA, 2CAS_, 2CBA_, 2CCYA, 2CDV_, 2CHSA, 2CPL_, 2CTC_, 2CTX_, 2CWGA, 2CYP_, 2DKB_, 2DLDA, 2DLN_, 2DRI_, 2EBN_, 2END_, 2ER7E, 2FAL_, 2FD2_, 2GDM_, 2GSQ_, 2GSTA, 2HBG_, 2HFT_, 2HHMA, 2HMZA, 2HPDA, 2HTS_, 2KAUB, 2KAUC, 2LIV_, 2MADL, 2MEV1, 2MNR_, 2MTAC, 2NACA, 2OLBA, 2OMF_, 2ORA_, 2PCDM, 2PGD_, 2PHY_, 2PIA_, 2PII_, 2PLEA, 2POLA, 2POR_, 2PRD_, 2PRK_, 2PSPA, 2REB_, 2RN2_, 2RSLB, 2SAS_, 2SCPA, 2SIL_, 2SNV_, 2STV_, 2TCT_, 2TGI_, 2TMDA, 3AAHA, 3BCL_, 3CD4_, 3CHY_, 3CLA_, 3COX_, 3DFR_, 3GLY_, 3GRS_, 3HHRC, 3PGA1, 3PGM_, 3PMGA, 3PTE_, 3SDHA, 3SICI, 3TGL_, 4ENL_, 4FGF_, 4FXN_, 4GCR_, 4MT2_, 4RHV3, 4SBVA, 4TS1A, 4XIAA, 5P21_, 5RUBA, 5RXN_, 5TIMA, 6FABL, 6TAA_, 7PCY_, 7RSA_, 8ABP_, 8ACN_, 8ATCA, 8ATCB, 8CATA, 8FABB, 8RUCI, 8TLNE, 9LDTA, 9PAP_, 9RNT_.

APPENDIX B. Average Pair Distance for Cross-Strand Interactions

Interaction	x^{ca} (Å)[a]	x^{sc} (Å)[b]
Antiparallel		
i → j − 2 (nonhbonded–nonhbonded)	6.9	3.0
i → j − 2 (hbonded–hbonded)	8.3	6.9
i → j − 1 (hbonded–nonhbonded)	5.7	5.3
i → j − 1 (nonhbonded–hbonded)	5.7	5.3
i → j (nonhbonded–nonhbonded)	4.5	2.4
i → j (hbonded–hbonded)	5.3	2.8
i → j + 1 (hbonded–nonhbonded)	6.4	6.0
i → j + 1 (nonhbonded–hbonded)	6.4	5.9
i → j + 2 (nonhbonded–nonhbonded)	8.8	4.5
i → j + 2 (hbonded–hbonded)	8.3	7.8
Parallel		
i → j − 2 (hbonded–nonhbonded)	7.4	3.4
i → j − 2 (nonhbonded–hbonded)	8.6	7.3
i → j − 1 (hbonded–hbonded)	6.3	5.8
i → j − 1 (nonhbonded–nonhbonded)	5.9	5.6
i → j (hbonded–nonhbonded)	4.9	2.7
i → j (nonhbonded–hbonded)	4.8	2.4
i → j + 1 (hbonded–hbonded)	6.3	5.6
i → j + 1 (nonhbonded–nonhbonded)	5.8	5.6
i → j + 2 (hbonded–nonhbonded)	8.5	7.2
i → j + 2 (nonhbonded–hbonded)	7.5	3.6

[a] x^{ca} is the distance between α-carbons and [b] x^{sc} is the distance between sidechain shells, as defined in Materials and Methods.

APPENDIX C. Insignificant Classifications of Interactions. $\overline{D_{123}}$ is the Measure of Significance of the Classification

Parent category	Child subcategories	$\overline{D_{123}}$
X-strand $i \rightarrow j^* \pm 2$ (antiparallel)	$n = 2/n = -2$	1.41
X-strand $i \rightarrow j^* \pm 2$ (antiparallel, hbonded–hbonded)	$n = 2/n = -2$	1.38
X-strand $i \rightarrow j^* \pm 1$ (parallel, $n = 1$)	hbonded–hbonded/nonhbonded–nonhbonded	1.37
X-strand $i \rightarrow j^* \pm 1$ (antiparallel, $n = -1$)	edge–edge/edge–internal/internal–edge/internal–internal	1.35
X-strand $i \rightarrow j^* \pm 2$ (antiparallel)	edge–edge/edge–internal/internal–edge/internal–internal	1.34
Contact (turn − Y)	turn-α-helix/turn-β-strand/turn–loop/turn–turn/turn–other	1.33
X-strand $i \rightarrow j^* \pm 2$	antiparallel/parallel	1.32
X-strand $i \rightarrow j^* \pm 2$ (antiparallel, $n = -2$)	edge–edge/edge–internal/internal–edge/internal–internal	1.32
X-strand $i \rightarrow j^* \pm 1$ (antiparallel)	edge–edge/edge–internal/internal–edge/internal–internal	1.32
X-strand $i \rightarrow j^* \pm 2$ (parallel, nonhbonded–hbonded)	$n = 2/n = -2$	1.30
X-strand $i \rightarrow j^* \pm 2$ (antiparallel, $n = 2$)	edge–edge/edge–internal/internal–edge/internal–internal	1.30
X-strand $i \rightarrow j^*$	antiparallel/parallel	1.29
X-strand $i \rightarrow j^* \pm 1$ (antiparallel, $n = 1$)	edge–edge–internal/internal–edge/internal–internal	1.29
X-strand $i \rightarrow j^* \pm 2$ (parallel)	$n = 2/n = -2$	1.29
X-strand $i \rightarrow j^* \pm 2$	$n = 2/n = -2$	1.29
X-strand $i \rightarrow j^*$ (antiparallel, edge–internal)	hbonded–hbonded/nonhbonded–nonhbonded	1.29
X-strand $i \rightarrow j^* \pm 2$ (antiparallel, internal–edge)	$n = 2/n = -2$	1.28
Local $i \rightarrow i + 3$	α-helix/loop/β-strand	1.25
X-strand $i \rightarrow j^* \pm 1$ (parallel, nonhbonded–nonhbonded)	$n = 1/n = -1$	1.23
Local $i \rightarrow i + 5$	α-helix/loop/β-strand	1.23
Local $i \rightarrow i + 2$ (loop)	high–high/high–low/low–high/low–low	1.22
Local $i \rightarrow i + 5$ (loop)	high–high/high–low/low–high/low–low	1.22
Local $i \rightarrow i + 3$ (loop)	high–high/high–low/low–high/low–low	1.20
X-strand $i \rightarrow j^* \pm 2$ (antiparallel, internal–internal)	$n = 2/n = -2$	1.18
Local $i \rightarrow i + 1$ (loop)	high–high/high–low/low–high/low–low	1.18
X-strand $i \rightarrow j^* \pm 1$ (parallel, $n = -1$)	hbonded–hbonded/nonhbonded–nonhbonded	1.17
Local $i \rightarrow i + 4$ (β-strand)	edge/internal	1.16
X-strand $i \rightarrow j^* \pm 1$ (parallel)	$n = 1/n = -1$	1.16
Local $i \rightarrow i + 4$ (α-helix)	edge–edge/edge–internal/internal–edge/internal–internal	1.16
Local $i \rightarrow i + 2$ (α-helix)	edge–edge/edge–internal/internal–edge/internal–internal	1.15
X-strand $i \rightarrow j^* \pm 1$ (antiparallel, edge–internal)	$n = 1/n = -1$	1.15
Local $i \rightarrow i + 4$ (loop)	high–high/high–low/low–high/low–low	1.15
X-strand $i \rightarrow j^* \pm 2$ (antiparallel, edge–edge)	$n = 2/n = -2$	1.15
Local $i \rightarrow i + 2$ (β-strand)	edge/internal	1.15
Local $i \rightarrow i + 1$ (β-strand)	edge/internal	1.14
X-strand $i \rightarrow j^*$ (antiparallel, internal–edge)	hbonded–hbonded/nonhbonded–nonhbonded	1.14
X-strand $i \rightarrow j^* \pm 2$ (parallel, $n = 2$)	hbonded–nonhbonded/hbonded–nonhbonded	1.14
Local $i \rightarrow i + 1$ (α-helix)	edge–edge/edge–internal/internal–edge/internal–internal	1.14
Local $i \rightarrow i + 3$ (α-helix)	edge–edge/edge–internal/internal–edge/internal–internal	1.13
Contact (other − Y)	other–α–helix/other–β–strand/other–loop/other–turn/other–other	1.13
X-strand $i \rightarrow j^*$ (antiparallel, nonhbonded–nonhbonded)	edge–edge/edge–internal/internal–edge/internal–internal	1.12
X-strand $i \rightarrow j^* \pm 2$ (antiparallel)	hbonded–hbonded/nonhbonded–nonhbonded	1.12
X-strand $i \rightarrow j^*$ (antiparallel, internal–internal)	hbonded–hbonded/nonhbonded–nonhbonded	1.11
X-strand $i \rightarrow j^* \pm 1$ (antiparallel)	hbonded–nonhbonded/nonhbonded–hbonded	1.11
X-strand $i \rightarrow j^* \pm 1$ (antiparallel, internal–edge)	hbonded–nonhbonded/nonhbonded–hbonded	1.11
Local $i \rightarrow i + 5$ (α-helix)	edge–edge/edge–internal/internal–edge/internal–internal	1.10
X-strand $i \rightarrow j^* \pm 1$	$n = 1/n = -1$	1.10
X-strand $i \rightarrow j^* \pm 2$ (parallel, hbonded–nonhbonded)	$n = 2/n = -2$	1.10
X-strand $i \rightarrow j^* \pm 2$ (antiparallel, edge–internal)	$n = 2/n = -2$	1.10
X-strand $i \rightarrow j^*$ (antiparallel, hbonded–hbonded)	edge–edge/edge–internal/internal–edge/internal–internal	1.09
X-strand $i \rightarrow j^* \pm 2$ (antiparallel, edge–edge)	hbonded–hbonded/nonhbonded–nonhbonded	1.08
X-strand $i \rightarrow j^* \pm 1$ (antiparallel, hbonded–nonhbonded)	$n = 1/n = -1$	1.08
X-strand $i \rightarrow j^* \pm 1$ (antiparallel)	$n = 1/n = -1$	1.08
X-strand $i \rightarrow j^* \pm 2$ (antiparallel, edge–internal)	hbonded–hbonded/nonhbonded–nonhbonded	1.07
X-strand $i \rightarrow j^* \pm 2$ (antiparallel, nonhbonded–nonhbonded)	edge–edge/edge–internal/internal–edge/internal–internal	1.06

APPENDIX C. (Continued)

Parent category	Child subcategories	$\overline{D}_{123}$
X-strand i → j* (antiparallel, edge–edge)	hbonded–hbonded/nonhbonded–nonhbonded	1.06
X-strand i → j* ± 2 (antiparallel, internal–edge)	hbonded–hbonded/nonhbonded–nonhbonded	1.06
X-strand i → j* ± 1	antiparallel/parallel	1.05
Local i → i + 3 (β-strand)	edge/internal	1.04
X-strand i → j* ± 2 (parallel)	hbonded–nonhbonded/hbonded–nonhbonded	1.03
Local i → i + 1 (β-strand, edge)	parallel/antiparallel/nobridge	1.02
X-strand i → j* ± 1 (antiparallel, nonhbonded–hbonded)	edge–edge/edge–internal/internal–edge/internal–internal	1.02
X-strand i → j* ± 1 (antiparallel, edge–edge)	n = 1/n = −1	1.01
X-strand i → j* ± 2 (parallel, n = −2)	hbonded–nonhbonded/hbonded–nonhbonded	1.01
X-strand i → j* ± 1 (antiparallel, hbonded–nonhbonded)	edge–edge/edge–internal/internal–edge/internal–internal	1.00
X-strand i → j* ± 2 (antiparallel, hbonded–hbonded)	edge–edge/edge–internal/internal–edge/internal–internal	1.00
X-strand i → j* ± 1 (antiparallel, edge–edge)	hbonded–nonhbonded/nonhbonded–hbonded	0.99
X-strand i → j* ± 1 (antiparallel, nonhbonded–hbonded)	n = 1/n = −1	0.99
X-strand i → j* ± 2 (antiparallel, internal–internal)	hbonded–hbonded/nonhbonded–nonhbonded	0.98
X-strand i → j* ± 1 (antiparallel, edge–internal)	hbonded–nonhbonded/nonhbonded–hbonded	0.97
X-strand i → j* ± 1 (antiparallel, internal–internal)	hbonded–nonhbonded/nonhbonded–hbonded	0.96
X-strand i → j* ± 1 (antiparallel, internal–edge)	n = 1/n = −1	0.96
X-strand i → j* ± 1 (parallel, hbonded–hbonded)	n = 1/n = −1	0.94
Local i → i + 2 (β-strand, internal)	both parallel/both antiparallel/mixed	0.94
X-strand i → j* ± 1 (parallel)	hbonded–hbonded/nonhbonded–nonhbonded	0.93
X-strand i → j* ± 1 (antiparallel, internal–internal)	n = 1/n = −1	0.93
Local i → i + 1 (β-strand, internal)	both parallel/both antiparallel/mixed	0.87
Local i → i + 2 (β-strand, edge)	parallel/antiparallel/nobridge	0.86

*j is the β-bridge partner of i.

APPENDIX D. The Most Significant Residue Pairs for the Antiparallel i → j − 2 (nonhbonded–nonhbonded) Interaction. Pearson's χ^2_{ij} Is a Measure of the Statistical Significance of the Deviation of the Pair Count N_{ij} From the Expected Count E_{ij} If There Were No Pairwise Dependence.

Residue pair ij	Pearson's χ^2_{ij}	Count N_{ij}	Expected count E_{ij}	$\frac{N_{ij}}{E_{ij}}$
WY	62.4	22	4.8	4.6
KE	46.1	21	5.3	3.9
ER	32.2	20	6.0	3.3
RE	31.9	21	6.6	3.2