Protein Science (1994), *3*:1582-1596. Cambridge University Press. Printed in the USA.

Derivation of rules for comparative protein modeling from a database of protein structure alignments

ANDREJ ŠALI[1] AND JOHN P. OVERINGTON[2]

[1] Department of Chemistry, Harvard University, Cambridge, Massachusetts 02138
[2] Pfizer Central Research, Ramsgate Road, Sandwich, Kent CT13 9NJ, United Kingdom

(RECEIVED March 8, 1994; ACCEPTED May 16, 1994)

Abstract

We describe a database of protein structure alignments as well as methods and tools that use this database to improve comparative protein modeling. The current version of the database contains 105 alignments of similar proteins or protein segments. The database comprises 416 entries, 78,495 residues, 1,233 equivalent entry pairs, and 230,396 pairs of equivalent alignment positions. At present, the main application of the database is to improve comparative modeling by satisfaction of spatial restraints implemented in the program MODELLER (Šali A, Blundell TL, 1993, *J Mol Biol 234*:779-815). To illustrate the usefulness of the database, the restraints on the conformation of a disulfide bridge provided by an equivalent disulfide bridge in a related structure are derived from the alignments; the prediction success of the disulfide dihedral angle classes is increased to approximately 80%, compared to approximately 55% for modeling that relies on the stereochemistry of disulfide bridges alone. The second example of the use of the database is the derivation of the probability density function for comparative modeling of the *cis/trans* isomerism of the proline residues; the prediction success is increased from 0% to 82.9% for *cis*-proline and from 93.3% to 96.2% for *trans*-proline. The database is available via electronic mail.

Keywords: comparative protein modeling; protein structure alignments; protein structure database

Once a sequence of a gene product has been determined, a search for related proteins can often provide important insights into its structure and function. This search is made possible by databases of protein sequences and structures. Sequence databases presently contain approximately 80,000 entries. The sequence databases include GenBank (Burks & Burks, 1988), Protein Information Resource (PIR) (George et al., 1986), and EMBO nucleotide sequences database (Hamm & Cameron, 1986).

The main database containing protein 3D structures is the Brookhaven Protein Data Bank (PDB) (Bernstein et al., 1977; Abola et al., 1987). The PDB contains almost 2,000 chains that represent approximately 112 unrelated folds (Chothia, 1992; Orengo et al., 1993). There are a number of protein structure databases that process and organize the atomic coordinates provided by the PDB to make it more useful for addressing particular problems. For example, ISIS (Akrigg et al., 1988; Islam & Sternberg, 1989; Thornton & Gardner, 1990) is a relational database containing protein structures, sequences, and analysis programs. It is accessed through a query system that can answer questions such as "List all examples of a positively charged side chain at the N-terminus of a helix." Two similar databases have also been described (Bryant, 1989; Huysmans et al., 1991). The power of the databases to address various questions is greatly enhanced when relationships between the entries are established. Several collections of alignments of protein structures have been published. Pascarella and Argos (1992) presented a database of alignments of protein structures as well as protein sequences, whereas Holm et al. (1992) described methods and programs for automatic derivation of structural alignments from the PDB coordinates. Recently, the available protein structures were systematically aligned and clustered into 112 protein fold families (Orengo et al., 1993).

Comparative protein modeling is based on the observation that proteins with similar sequences adopt similar 3D structures (Chothia & Lesk, 1986; Hubbard & Blundell, 1987). As a result, knowledge of 3D structure of one or more proteins is useful in modeling the 3D structure of a related sequence (Browne et al., 1969). Many of the methods that have been proposed for homology modeling were reviewed by Šali and Blundell (1993). Our particular approach to comparative modeling is based on satisfaction of spatial restraints that are obtained from an alignment of a target sequence with related 3D structures (Šali & Blundell, 1990, 1993; Šali et al., 1990). The method is not limited only to proteins related by divergent evolution, but also applies to protein engineering studies such as the humanization

Reprint requests to: Andrej Šali, Department of Chemistry, Harvard University, 12 Oxford Street, Cambridge, Massachusetts 02138; e-mail: sali@tammy.harvard.edu.

of monoclonal antibodies and enhancement of thermal stability through site-directed mutagenesis. The spatial restraints are expressed as probability density functions (pdf's) for the spatial feature that is restrained. For example, the probability of a certain C^{α}-C^{α} distance in the sequence of the unknown is a Gaussian function with the mean equal to the equivalent distance in a related structure. Such pdf's for several types of distances, main chain, and side chain dihedral angles were initially obtained from a small database of 17 family alignments (Šali, 1991) that was built by the protein structure comparison program COMPARER (Šali & Blundell, 1990; Zhu et al., 1992). This small database was then gradually extended and used to obtain environment-specific residue substitution tables (Overington et al., 1990, 1992), to improve homology modeling of loops (Topham et al., 1993), and to increase the sensitivity and accuracy of aligning sequences with structures (Johnson et al., 1993). Recently, 87 alignments were collected (Overington et al., 1993). This collection was used as the starting point for the present database with 105 groups of 416 aligned protein structures.

In this paper, we describe a database of multiple protein structure alignments and computer software that is designed to help improve comparative protein modeling but also has other applications. The main distinction of the present database is the inclusion of a procedure for derivation of spatial restraints useful in comparative modeling; i.e., various spatial features of a protein sequence can be correlated with many features of a protein structure aligned with that sequence and such relationships can then be used in our approach to comparative modeling (Šali & Blundell, 1993). In the Methods section, we describe the database of alignments,[3] its contents and organization, and the programs used to explore it. The programs calculate and correlate various properties of protein structure, such as solvent accessibility, residue type, and atomic distances, as well as analyze the correlations. In the Results section, we illustrate the usefulness of the database by deriving information for modeling a disulfide bridge in a target sequence based on the equivalent disulfide in the known template structure. We show that the prediction success of the disulfide dihedral angle classes is increased to approximately 80% compared to approximately 55% for modeling that relies on the stereochemistry of disulfide bridges alone. We also apply the database to derive restraints on the *cis/trans* isomerism of a proline main chain, given the knowledge of a related structure. In this case, the prediction success is increased from 0% to 82.9% for *cis*-proline and from 93.3% to 96.2% for *trans*-proline.

Results

Comparative modeling of disulfide bridges

The bond lengths, angles, and torsional preferences for S–S bonds in disulfide bridges are well established both from theoretical modeling studies (Qian & Krimm, 1993) and from analyses of small molecule crystal structures (Engh & Huber, 1991). Similarly, there have been a number of analyses of disulfide bridges in proteins that can be used as a basis for modeling their conformation (Richardson, 1981; Thornton, 1981; Pabo & Suchanek, 1986; Sowdhamini et al., 1989, 1993). However, none of these analyses describes the restraints on a conformation of a disulfide bridge that are provided by the information about the equivalent disulfide bridge in a related structure. Because disulfide bridges occur frequently and because their conformation sometimes significantly influences the 3D structure as a whole, especially for many small biologically active peptides, it is important to be able to model the disulfide bridges as well as possible. We use the database and the programs described in the Methods section to derive restraints on the disulfide conformation from the equivalent disulfides in related proteins. The restraints are expressed as pdf's suitable for comparative modeling by satisfaction of spatial restraints, implemented in the computer program MODELLER (Šali & Blundell, 1993). Many features of this analysis are similar to the derivation of the pdf's for modeling of all side chain conformations (Šali & Blundell, 1993).

Among the 78,495 residues in the 105 alignments, there are 936 half-cystine residues forming 468 disulfide bridges and 670 cysteine residues, which by definition are not in disulfide bridges. The 468 disulfides give 1,295 equivalent disulfide bridge pairs; 2 disulfides are equivalent when both of the half-cystine residues are equivalent. Half-cystine is by far the most conserved residue type in terms of evolutionary substitution with other residue types; in the present database, the probability that it is conserved is 0.873. By contrast, cysteine is significantly less conserved (0.430). For comparison, the other most conserved residue types are Trp (0.600), Gly (0.572), and Pro (0.484); the least conserved residue types are Met (0.229), Asn (0.254), and Lys (0.317). There are only 25 pairs of equivalent Cys residues where a residue in one protein is involved in a disulfide and the residue in the other protein is not. This number should be compared with the 1,295 pairs of equivalent half-cystines where both residues are involved in an equivalent disulfide bridge. This demonstrates that the presence or absence of a disulfide bond at a certain position in a family fold is a strongly conserved feature in evolution (Thornton, 1981). It remains to be seen how conserved the *conformation* of a disulfide bridge is; if both the conformation and presence of a bridge are conserved, they should make possible derivation of strong restraints for comparative modeling.

We first establish stereochemical preferences of a single disulfide bridge. The 5 dihedral angles of a disulfide bridge are defined in Figure 1. Due to the symmetry of the disulfide topology, dihedral angle χ_1 is equivalent to dihedral angle χ_5; similarly, χ_2 is equivalent to χ_4. The distributions of the dihedral angles in the 410 half-cystines from the 181 high-resolution crystallographic structures (resolution of 2 Å or less) in the database, $W(\chi_i)$ (Equation 3), are shown in Figure 2A–C. The distributions have typical trimodal (χ_1, χ_2) and bimodal (χ_3) shapes. The only exception is the χ_2 angle, which has a small peak at approximately 120°. This small peak is contributed largely by the high-resolution structures of the immunoglobulin variable domains. It is ignored in the present analysis because the subsequent results are very similar when the 2 alignments with immunoglobulin variable chains are omitted (data not shown). Each of the contiguous ranges centered on the peaks corresponds to a dihedral angle class. The distributions can be normalized and modeled by a weighted sum of Gaussian functions, each function corresponding to 1 class, as shown in Figure 2. The weights, means, and standard deviations of the Gaussians are given in Table 1. Since there are only 2 or 3 dihedral angle classes

[3] The alignment files (Fig. 8) are available upon request from the authors at the Internet e-mail address overingtonj@pfizer.com.

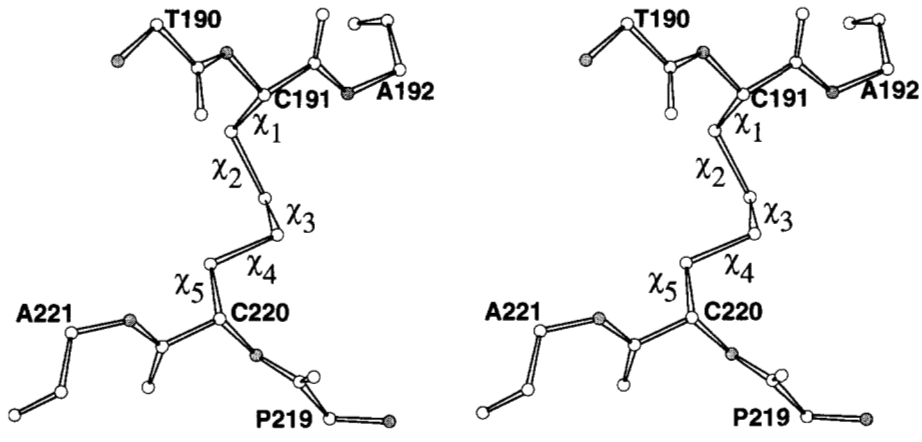

Fig. 1. Example of a disulfide bridge and definition of the disulfide dihedral angles. The disulfide bridge 191-220 from rat tonin is shown (PDB code 1TON). Labels are next to the C^{α} atoms. A disulfide bridge 220-191 has the χ_i angles labeled in the reverse order.

per dihedral angle type, the distribution of the actual dihedral angles among these classes can be estimated much more accurately than the distribution of the dihedral angles among the 36 10° intervals. In the subsequent derivations, we use all 416 structures in the database, not only the high-resolution entries.

When distinguishing only between the classes of the 5 dihedral angles in the disulfide bridge, there are 90 different possible conformations. However, only 10 of these are significantly populated (≥4% of the total): 11111 (10.9%), 11112 (8.8%), 11211 (7.3%), 11222 (5.8%), 11131 (5.6%), 13222 (5.6%), 22233 (4.5%), 22211 (4.6%), 11223 (4.3%), and 11221 (4.1%), where the 5-digit numbers are the classes of the 5 χ_i angles (Table 1) and the symmetry in the numbering of the dihedral angles has been taken into account (i.e., conformation 32221 is the same as conformation 12223). Together, these 10 conformations account for 57% of all disulfides. The top 20 conformations account for 81% of all disulfides.

Next, we examine the interdependence of the dihedral angles within the same disulfide bridge. To this end, we use the plots of $W(\chi_i, \chi_j)$ (Equation 2), $W'(\chi_i/\chi_j)$ (Equation 3) (Fig. 3), and their conditional entropies (Equation 6) (Table 2). The strongly correlated pairs of dihedral angles are (χ_1, χ_2), (χ_2, χ_3), (χ_3, χ_4), (χ_4, χ_5), and (χ_2, χ_4); with the exception of the last pair, they are all the neighboring dihedral angles. This suggests that, ideally, the individual dihedral angles should not be considered independently from each other but that there should be one 5-dimensional variable describing the conformational state of a disulfide bridge as a whole. However, the size of the present sample is not large enough to proceed in this way. Moreover, it is likely that in modeling, the interdependence of χ_1-χ_5 can be achieved by the use of the Lennard-Jones interaction terms (Dunbrack & Karplus, 1993). Thus, the 5 dihedral angles remain considered as independent variables.

We now establish restraints on the conformation of a given disulfide that are provided by an equivalent disulfide bridge in a related structure. The pdf's $p(\chi_i/\chi_i')$ are shown in Figure 4 together with their conditional entropies. These are larger than the conditional entropies for the interdependence of the dihedral angles in the same bridge, showing that the knowledge of an equivalent disulfide provides strong restraints on a given disulfide.

How much better is the modeling of a disulfide given an equivalent bridge compared to pure stereochemical modeling? A rigorous answer would be provided by testing the corresponding pdf's with MODELLER on a large number of cases. Because this is impractical, we make an estimate for the χ_3 dihedral an-

Table 1. *Definition of the dihedral angle classes for the dihedral angles in a disulfide bridge*[a]

Dihedral angle	Class	Weight	Mean (deg)	Standard deviation (deg)
χ_1	1	0.619	−64.73	12.09
	2	0.265	182.03	12.79
	3	0.116	60.34	9.27
χ_2	1	0.496	−73.69	20.92
	2	0.235	74.00	19.98
	3	0.269	181.10	35.68
χ_3	1	0.425	−85.77	7.56
	2	0.575	93.21	13.66

[a] χ_4 and χ_5 angle classes are equivalent to the χ_2 and χ_1 classes, respectively. The parameters of the dihedral angle classes (i.e., weights, means, and standard deviations) were obtained from least-squares fitting of a sum of Gaussian functions to the observed histograms, as shown in Figure 2.

Table 2. *Interdependence of the dihedral angle classes in the same disulfide bridge*[a]

	y				
x	χ_1	χ_2	χ_3	χ_4	χ_5
χ_1	1.000	0.146	0.000	0.013	0.011
χ_2	0.134	1.000	0.071	0.069	0.012
χ_3	0.001	0.109	1.000	0.109	0.001
χ_4	0.012	0.069	0.071	1.000	0.134
χ_5	0.011	0.013	0.001	0.146	1.000

[a] The dependence of the probabilities of a certain dihedral angle class on the actual value of another dihedral angle class is measured quantitatively by the conditional entropies $U(x/y)$ (Equation 6). Frequencies and conditional probabilities for three of the strongly correlated pairs of dihedral angles are shown in Figure 3.

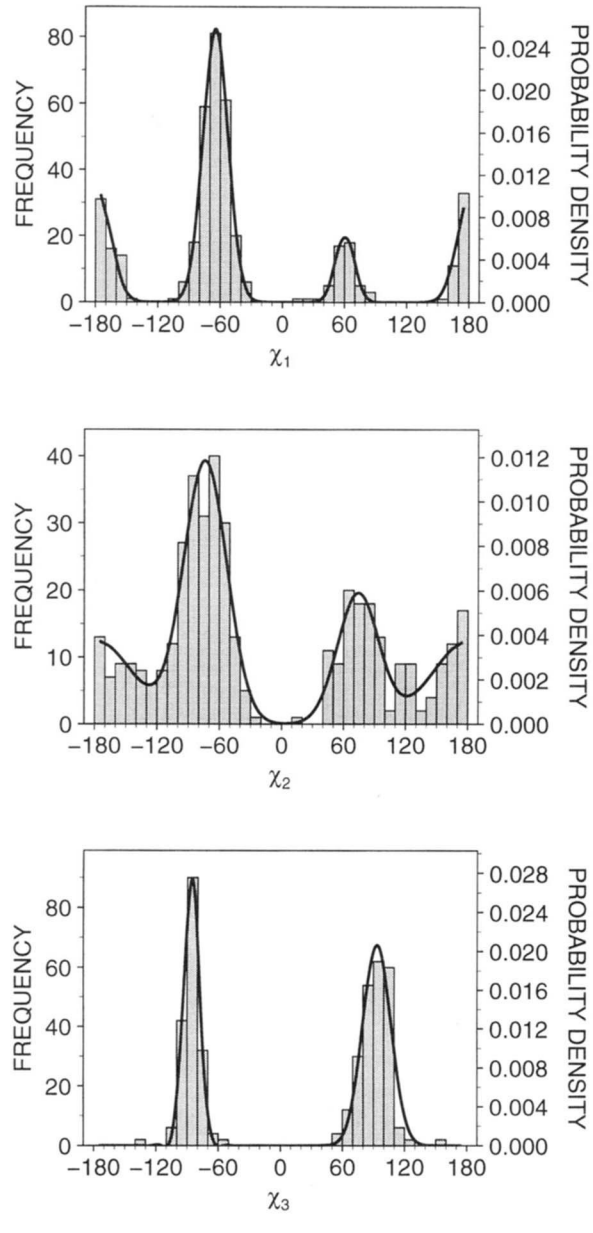

Fig. 2. Definition of the dihedral angle classes for the dihedral angles in a disulfide bridge. The histograms show the distribution of the corresponding dihedral angles in all 410 half-cystines from the 181 high-resolution structures (≤2.0 Å) in the alignments database. The curves show a fit of a weighted sum of 3 (χ_1 and χ_2) and of 2 (χ_3) Gaussian functions to the histograms. For example, for χ_3:

$$p(\chi) = \omega_1 \cdot \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\Delta(\chi, \bar{\chi}_1)}{\sigma_1}\right)^2\right] + \omega_2 \cdot \frac{1}{2\pi} \exp\left[-\frac{1}{2}\left(\frac{\Delta(\chi, \bar{\chi}_2)}{\sigma_2}\right)^2\right],$$

where $\bar{\chi}_i$ are means, σ_i are standard deviations, ω_i are weights, and $\omega_1 + \omega_2 = 1$. The function $\Delta(\alpha, \beta)$ returns the difference between angles α and β while allowing for the 360° periodicity of the angles: the difference is defined as the shortest path around the 360° circle from angle α to angle β (clockwise direction is positive). Program LSQ was used for least-squares fitting. The weights, means, and standard deviations of the Gaussians are listed in Table 1. The RMS deviations between the probability density models and the relative frequencies from the database are 0.5×10^{-3}, 0.9×10^{-3}, and 0.9×10^{-3} for χ_1, χ_2, and χ_3, respectively.

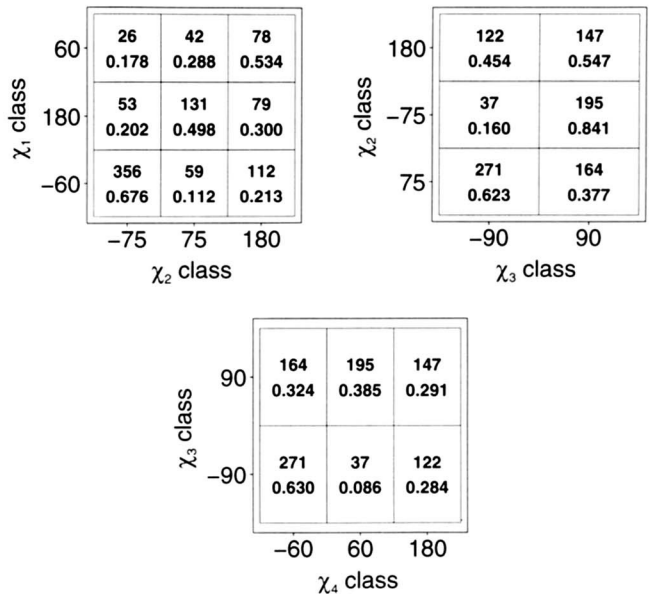

Fig. 3. Strength of association between 2 dihedral angle classes in the same disulfide bridge. Three strongly correlated pairs are shown. The dihedral angle classes are identified by their means. The top number in each cell is the number of disulfide bridges with a corresponding combination of the 2 dihedral angle classes in the database. In total, there are 468 disulfides in the database, resulting in 2 × 468 = 936 comparisons for each dihedral angle pair because of the bridge symmetry. The bottom number in each cell is a conditional probability $p(x/y)$; thus, the numbers in a row sum to 1. Note the differences between the rows. These differences are caused by the dependence of the x-axis dihedral angle class on the y-axis dihedral angle class. The magnitude of this dependence is quantified in Table 2.

gle class as follows. The predicted dihedral angle class is that class in a pdf that has the highest probability of occurrence. Using stereochemical preferences as reflected in $p(c_3)$ alone would correctly predict 58% of the disulfides in the database (Table 1). By contrast, using an equivalent disulfide as represented by $p(c_3/c_3')$ would correctly predict (1,394 + 2,698)/(1,394 + 2,698 + 544 + 544) = 79% of disulfides (Fig. 4C). The equivalent numbers for the χ_1 dihedral angle class are 62% and 83%, and for χ_2, 50% and 77%.

Comparative modeling of the cis/trans isomerism in proline residues

A number of analyses of the conformational properties of proline in proteins have been published (Stewart et al., 1990; MacArthur & Thornton, 1991). However, none of these analyses describes how much information on a state of a certain Pro residue in a modeled sequence is provided by a homologous structure. This is an important question because a significant fraction of proline residues are in the *cis* state (approximately 6.7%), and because a choice of a particular Pro isomer may have a significant impact on the 3D model (Fig. 5). We use the present database to attempt to improve comparative modeling of the *cis/trans* states of proline.

Among the 78,495 residues in the 105 alignments, there are 3,576 (4.5% of all residues) proline residues. Of these, 238 (6.7% of all prolines) are *cis*-prolines. The *cis* conformation is assigned

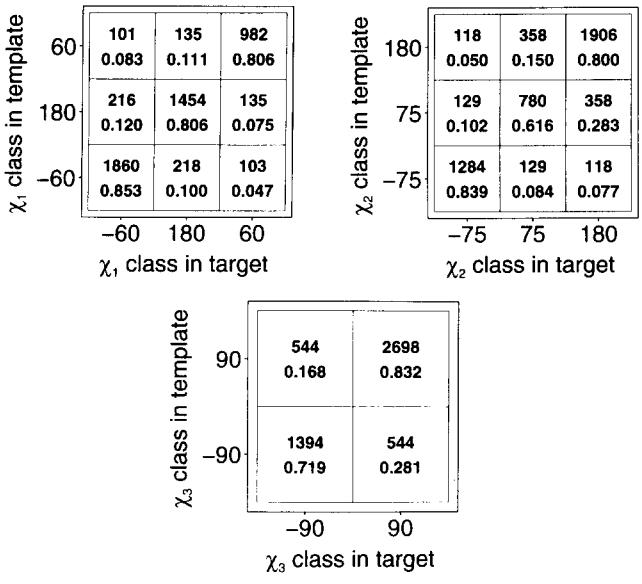

Fig. 4. Correlation between equivalent dihedral angles. This figure is similar to Figure 3 except that a dihedral angle class is correlated with the same dihedral angle class in an equivalent disulfide bridge, not with another dihedral angle class in the same disulfide bridge. There are 1,295 equivalent disulfide pairs that result in 5,180 (4 × 1,295) comparisons because of the disulfide bridge symmetry and because a comparison of disulfides from proteins A and B is different from a comparison of the same disulfides from proteins B and A. See the legend of Figure 3 for an interpretation of these plots. The entropies and conditional entropies are: $S(\chi_1/\chi_1') = 0.576$, $U(\chi_1/\chi_1') = 0.463$, $S(\chi_2/\chi_2') = 0.662$, $U(\chi_2/\chi_2') = 0.376$, $S(\chi_3/\chi_3') = 0.505$, and $U(\chi_3/\chi_3') = 0.236$. These conditional entropies can be compared with significantly smaller values for the intraresidue correlations (Table 2).

when the preceding ω dihedral angle is between $-90°$ and $+90°$; otherwise, the *trans* conformation is assigned. Seventy-one residues other than proline are also in the *cis* conformation; some of these may be errors in structure refinement. Out of the 211,556 equivalent residue pairs in the database, where a gap can be one of the "residues" in the pair, at least 1 proline occurs in 13,659 of the pairs (6.4%); in 4,365 out of 13,659 pairs, both residues are prolines. As mentioned above, proline is the fourth most conserved residue type in terms of substitution with other residue types. This is presumably caused by its unique lack of hydrogen bonding from the main chain amide and by conformational restrictions on the ϕ main chain dihedral angle. As a result, a hydrophobic Pro residue frequently occurs in turns, where the value of the ϕ angle favored by Pro and its ability to form *cis*-peptide bonds are often required. Pro also occurs at the termini of helices and β-sheets, and by implication on the surface, because it breaks the regular pattern of secondary structure hydrogen bonds (Richardson & Richardson, 1988). The conservation of proline increases to 0.654 for a *cis*-proline and decreases to 0.466 for *trans*-proline, compared to 0.484 for any proline. This makes *cis*-proline the second most conserved residue type, after the most conserved half-cystine (0.873); *trans*-proline remains one of the most conserved residue types. This indicates that the structural role of the *cis*-proline is more specific than that of *trans*-proline.

The first question that we ask is "Does the distribution of proline between the *cis* and *trans* states depend on the type of the equivalent residue when this equivalent residue is in the *trans* conformation?" To answer this question, the frequency table $W'(\omega_c, r = \text{Pro}, r', \omega_c' = trans)$ and the corresponding pdf $p(\omega_c/r = \text{Pro}, r', \omega_c' = trans)$ are shown in Figure 6; ω_c describes

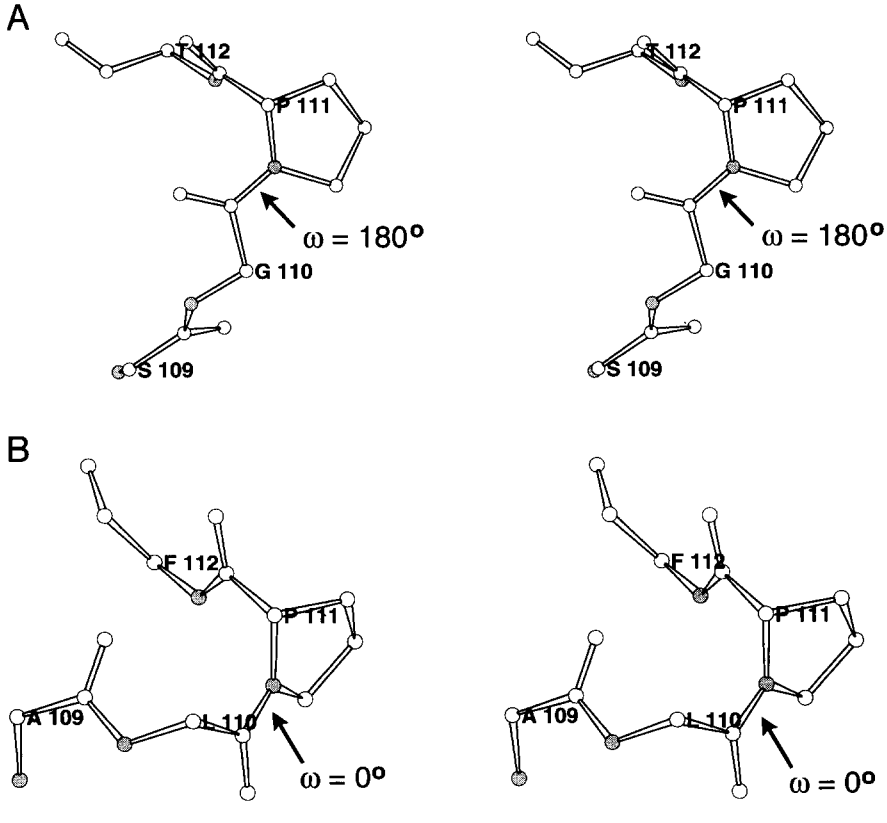

Fig. 5. Examples of the proline *cis* and *trans* conformations. Labels are next to the C^{α} atoms. **A:** *Mucor pusillus* pepsin (PDB code 1MMP) with *trans*-proline at position 111. **B:** Mouse renin (Dhanaraj et al., 1992) with *cis*-proline at the equivalent position 111.

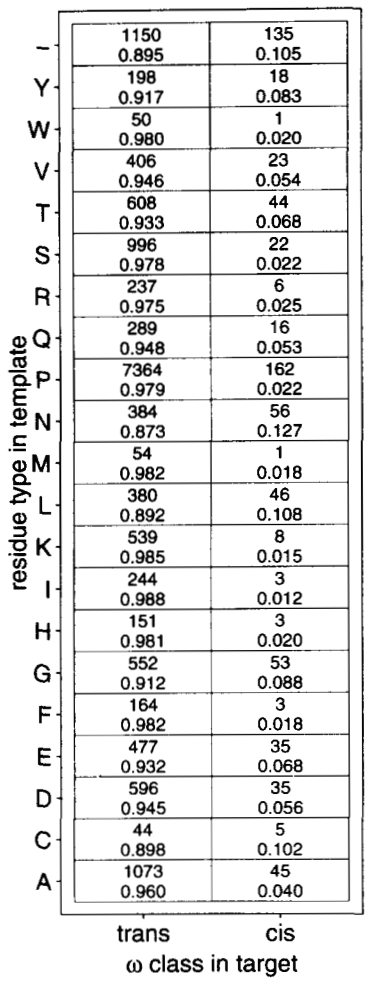

Fig. 6. Isomer propensity of proline as a function of the type of an equivalent residue. A gap residue type is indicated by a dash. The top number in each cell is the frequency $W(\omega_c, r = \text{Pro}, r', \omega'_c = \textit{trans})$. The bottom number is the pdf $p(\omega_c/r = \text{Pro}, r', \omega'_c = \textit{trans})$.

the main chain isomer of a given residue, r is the type of a given residue, r' is the type of an equivalent residue, and ω'_c is the isomer of an equivalent residue. Due to the relatively small size of the database, the differences among the estimated probabilities for all equivalent residue types r' are small compared to errors in these estimates. In other words, there is no reason to believe that any of the probabilities $p(\omega_c = \textit{cis}/r = \text{Pro}, r', \omega_c = \textit{trans})$ is significantly different from 6.7%; the only exceptions may be a substitution from *trans*-Pro to *cis*-Pro, which appears to be less likely than an average substitution to *cis*-Pro, and a substitution of a gap with a *cis*-Pro (i.e., an insertion of *cis*-Pro), which appears to be more likely.

This observation justifies combining all equivalent residue types when asking the following question: "What is the probability that a proline has a *cis*-peptide geometry given the state of an equivalent residue, regardless of its type?" To answer this question, the frequency table $W(\omega_c, r = \text{Pro}, \omega'_c)$ and the corresponding pdf $p(\omega_c/r = \text{Pro}, \omega'_c)$ are shown in Figure 7. The information about the isomeric state of an equivalent residue strongly restrains the conformation of a given proline: the restrained proline has a probability of 82.9% to be a *cis*-proline if the equivalent residue is *cis*, and a probability of 96.2% to be a *trans*-proline if the equivalent residue is *trans*.

It has been noted that residues close to proline in sequence may affect its probability to be in the *cis* state (MacArthur & Thornton, 1991), in particular, that a preceding tyrosine increases the likelihood of the *cis*-proline. Since such correlations could be used for homology modeling of proline, we derived 2 pdf's of the form $p(\omega_{c/r} = \text{Pro}, r_{i\pm1})$, where $r_{i\pm1}$ is the type of the preceding and subsequent residue, respectively. Thirty-one out of 156 Tyr–Pro pairs (20%) in the alignments database have proline in the *cis* conformation, 3 times more than expected by chance (6.7%). The second most biased pair is Phe–Pro; 17 of 150 pairs (11%) have *cis*-proline. The residue pair that is least likely to have *cis*-proline is Cys–Pro (only once out of 88 occurrences). There is also some influence on proline by the subsequent residue: Pro–His and Pro–Arg have 15 of 97 (15%) and 16 of 118 (13%) prolines in the *cis* state, respectively. The residues most successful in decreasing the probability of the *cis* state for the preceding proline are Asp (6/240, 2.5%) and Glu (8/270, 3%). Despite a small number of examples in the database, at least some of the preferences appear to be real. However, in homology modeling of *cis*-proline, we do not combine these preferences with $p(\omega_{c/r} = \text{Pro}, \omega'_c)$ into $p(\omega_{c/r} = \text{Pro}, \omega'_c, r_{i-1}, r_{i+1})$ because the database is too small to obtain a reliable estimate of the expanded pdf and because the correlation of proline conformation with the conformation of an equivalent residue is significantly stronger than the correlation with the preceding or subsequent residue. We also note that our database, which includes homologous structures, may not be as suitable for derivation of pdf's containing only features from a single protein (e.g., $p(\omega_{c/r} = \text{Pro}, r_{i\pm1})$) as the databases where special care was taken to minimize the similarities between the proteins in the database (MacArthur & Thornton, 1991).

The improvement in modeling the *cis* and *trans* states of proline can be estimated similarly to that of the disulfide modeling above. Since the vast majority of prolines in proteins are *trans* (93.3%), no *cis*-proline would be predicted correctly if only the overall stereochemical preference of proline (i.e., $p(\omega_{c/r} = \text{Pro})$) were taken into account. However, when knowledge of an equivalent conformation is used (i.e., $p(\omega_{c/r} = \text{Pro}, \omega'_c)$), 82.9% of all *cis*-prolines and 96.2% of *trans*-prolines are predicted correctly (Fig. 7). We can use pdf $p(\omega_{c/r} = \text{Pro}, \omega'_c, s)$ in the comparative modeling program MODELLER to improve modeling of the proline main chain.

Discussion

In the Methods section, we describe a database of alignments that contains 105 groups of 416 structurally defined proteins or their fragments. The alignments were obtained by the least-squares superposition of C^α backbones (Sutcliffe et al., 1987) and by a more flexible multifeature comparison method (Šali & Blundell, 1990). The database is used with programs for automated access to and processing of the information in it. This information includes the sequence of amino acid residues, positional coordinates, main chain and side chain dihedral angles, secondary structure assignments, residue solvent accessibilities, hydrogen bonds, neighboring residues, and many other features. We describe a systematic and quantitative approach to searching for significant associations between the features of protein sequence and structure. This involves expressing the association between selected features as a conditional pdf and quantifying the strength of the association by entropy, conditional entropy,

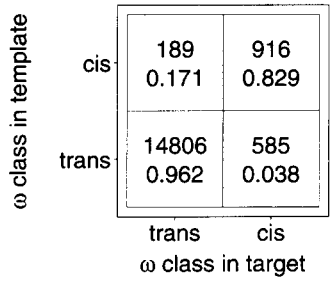

Fig. 7. Isomer propensity of proline as a function of conformation of the equivalent residue. The top numbers are frequency table $W(\omega_c, r = \text{Pro}, \omega_c')$. The bottom numbers are pdf $p(\omega_c/r = \text{Pro}, \omega_c')$.

and, where possible, by the prediction success of the tested pdf's. The features can be either of sequence or of 3D structure and they can come from 1, 2, or 3 related proteins. For example, a distribution of the differences between equivalent C^α–C^α distances from 2 aligned proteins can be easily prepared as a function of the overall sequence similarity of the 2 proteins. In a separate paper, the smoothing procedure of Sippl (1990) was extended to multidimensional pdf's to minimize the problem of a small database (Šali & Blundell, 1993); it was not necessary to apply this smoothing procedure to the pdf's derived here because the database was sufficiently large.

Several collections of protein structure alignments have been described (Šali, 1991; Holm et al., 1992; Pascarella & Argos, 1992; Orengo et al., 1993). The database of Pascarella and Argos (1992) contained 38 family alignments including 209 tertiary structures and 8 times as many related sequences for which no 3D structures were available. The Holm et al. (1992) database consisted of 1 data set for each of the 154 structures representative of the PDB; each data set contained an alignment of the corresponding 3D structure with related structures and sequences, including remotely related motifs. The Orengo et al. (1993) database includes alignments of pairs of related structures that were identified by comparison of all pairs of representative protein structures in the PDB; the structures were subsequently clustered into 112 different fold families. One of the main differences between these collections of alignments is that they use different comparison methods. The databases of Orengo et al. (1993) and Holm et al. (1992) are probably the most systematic and accurate in establishing the most remote structural relationships between the entries of the PDB. The main distinction of the database presented here is that it includes multiple alignments and a general mechanism for extracting rules directly applicable to protein modeling.

Frequency tables and related matrices, such as those used in this paper, are commonly applied to analyze or predict some aspects of protein structure. For example, multidimensional forms of the probability tables **W** and their transformations have been employed to search for combinations of protein features that are conserved in evolution (Overington et al., 1990, 1992); these features include residue type, its secondary structure state, solvent accessibility, and hydrogen bonding properties. Similar matrices were used to detect distantly related sequences (Lüthy et al., 1991; Johnson et al., 1993), to identify sequences that fold into a known 3D structure (Bowie et al., 1991; Johnson et al., 1993), and to assess protein 3D models (Overington et al., 1990; Lüthy et al., 1992). Other examples of frequency tables and closely related matrices include Dayhoff's MDM250 mutation matrix (Dayhoff et al., 1978), the Ramachandran plot obtained from a database of known protein structures (Wilmot & Thornton, 1990), various parameter sets for secondary structure prediction (Chou & Fasman, 1974), side chain rotamer libraries (Janin et al., 1978; Ponder & Richards, 1987; Dunbrack & Karplus, 1993), and hydrophobicity scales found by analyzing known protein structures (Manavalan & Ponnuswamy, 1978). Finally, there is also close correspondence between the pdf's and the potentials of mean force as derived from a database of known protein structures (Miyazawa & Jernigan, 1985; Sippl, 1990). It is likely that future studies like those mentioned above will be facilitated by the alignments database, programs, and methods described in this paper.

In the Results section, we illustrate the usefulness of the alignments database by applying it to comparative modeling of disulfide bridges and *cis*-prolines. We show that the homology-derived restraints on the disulfide dihedral angles are strong relative to the stereochemical restraints alone and are thus useful in comparative modeling of disulfide bridges. This is a result of conservation of disulfide bridge conformation in a family of related proteins. When supplementing stereochemical preferences with information about the conformation of an equivalent disulfide bridge, the prediction success is estimated to improve from 62% to 83%, from 50% to 77%, and from 58% to 79% for χ_1, χ_2, and χ_3 dihedral angle classes, respectively. Similar conclusions are also valid for proline main chain conformation. In this case, the prediction success is estimated to increase from 0% to 82.9% for *cis*-proline and from 93.3% to 96.2% for *trans*-proline.

In this study, we used dihedral angles both to describe the conformation of a disulfide bridge and to evaluate the prediction accuracy of a given pdf. While it is clear that the present pdf's significantly improve the accuracy of disulfide bridge modeling, it may be necessary for further improvement to use restraints on the disulfide bridge atoms that specify their positions relative to the rest of the molecule (Schrauber et al., 1993); such restraints could include distance restraints relative to the neighboring amino acid residues.

The present pdf's do not take into account the dependence of a change in a dihedral angle or in proline conformation on sequence similarity between the proteins compared; i.e., the df's are an average over the whole alignments database, which spans the range of pairwise sequence identities from 3% to 98% with a mean of 43%. Further small improvement in the prediction success may result from expanding the current pdf's of the form $p(x/a)$ to $p(x/a, s)$, where s is some measure of structural similarity between the 2 proteins that can be calculated from their sequences and the structure of the template, e.g., solvent accessibility of the residue in the template, overall sequence identity, or similarity of the sequence segments folded around the residues compared (Šali & Blundell, 1993). Such a measure would result in pdf's that predict greater conservation of the dependent variable when similarity is high and converge to stereochemical preference when similarity is low.

The pdf's derived from the alignments are generally not very sensitive to small random mistakes in the alignments because each pdf consists of a large number of points resulting in cancellation of systematic errors in the position of the maxima. However, the entropy of the pdf's estimated from suboptimal

alignments may be increased compared to that of the true pdf's, similar to the increase in the spread of the distribution of observed side chain dihedral angles when the protein structures determined at a lower resolution are included in the derivation of these distributions (Ponder & Richards, 1987). In the present work, the problems resulting from suboptimal alignments are minimized because the alignments were obtained by comparing 3D structures, not amino acid sequences. Since 3D structure is more conserved in evolution than sequence, such alignments are more reliable.

Another point to bear in mind when judging the suitability of an alignments database is that there are different best alignments for different purposes. For example, it has been recently shown that an insertion or a deletion of a single residue within a helix may not disrupt the helix (Heinz et al., 1993). As a consequence, the best structural alignment is shifted for 1 residue relative to the best sequence alignment. Thus, the ultimate criterion of the alignment quality is the quality of the final results obtained on the basis of the alignments, such as the increase in the prediction success in modeling disulfide bridges and proline main chains reported here.

The present alignment files will be extended to include sequences aligned with structures as well as single sequence or structure entries. This will allow the use of the current software for other purposes, in addition to improving comparative modeling, as described in this paper. For example, the best pdf's for matching sequences with sequences, sequences with structures, and structures with structures may be found. The applications of such pdf's could include sequence profile methods (Gribskov et al., 1987), structure profile methods (Lüthy et al., 1991; Johnson et al., 1993), structural comparison (Šali & Blundell, 1990), and sequence threading (Finkelstein & Reva, 1991; Godzik et al., 1992; Jones et al., 1992; Sippl & Weitckus, 1992). The alignment database may also be a good starting point for deriving the rules for combinatorial modeling (Cohen & Kuntz, 1989; Taylor, 1991) and for deriving the pseudopotentials for ab initio prediction of protein structure.

Methods

Organization of the database

The database of alignments consists of 1 alignment file for each group of aligned protein structures. The contents of these alignments are described below and in Table 3. There is also computer software for creating and exploring the database (see below). For most applications, the coordinate sets in the PDB are also needed (Bernstein et al., 1977; Abola et al., 1987). Some types of data, such as solvent accessibilities, hydrogen bonds, residue neighbors, secondary structure assignments, and dihedral angles, are calculated on demand and stored in separate files for faster availability the next time they are required.

The format of the alignment file is reminiscent of that of the PIR sequence database (George et al., 1986). A sample alignment file is shown in Figure 8. The corresponding alignment displayed by the formatting program JOY (Overington et al., 1990, 1992) is shown in Figure 9. In order to increase the usefulness of the database and the programs, each alignment file may contain any number of sequences or structures; this will allow the database eventually to contain entries for most of the structures in the PDB, even if without the structurally defined homologues, as well as the sequences from the sequence databases aligned with the structures. However, at present, the database contains only the alignment files with 2 or more aligned 3D structures.

Selection and comparison of protein structures

The current alignments database was built by reorganizing and extending a collection of alignments (Overington et al., 1993) to allow a number of fully automated operations, such as database processing and scanning; for example, the list of all the entries and the distribution of the selected combinations of many protein features can be easily obtained. The structures were selected from the PDB release of January 1994 (Bernstein et al., 1977; Abola et al., 1987), from both the full release and prerelease entries. Additional data sets were added if these were available from the authors (see legend to Table 3). Structural alignments were performed either with the program COMPARER (Šali & Blundell, 1990) or with a modified version of the multiple structure superposition program MNYFIT (Sutcliffe et al., 1987). Most of the alignments include structures where the rigid-body superposition implemented in MNYFIT gives high-quality structural alignments. As in all protein comparisons, the best alignment in the gap regions is most difficult to determine. However, the present applications of the database rely on the comparison of topologically equivalent regions and should not be too sensitive on the small number of ambiguously aligned positions close to the gap regions. Where possible, we use the native structures in all comparisons and generally keep key prosthetic groups in the structures (e.g., the heme rings in the cytochromes and globins). Where multiple copies of a structure are available, we use the data set at the highest resolution. Similarly, when given the choice of an X-ray or NMR-derived structure, we use the X-ray structure. When NMR-derived structures are included, we use either the minimized average coordinate set or the first structure listed in the PDB file.

As new structures appear in the PDB they are screened for similarity on the basis of sequence against all other PDB structures. Occasionally, similarities reported in literature are also used as a basis for the database alignments. In general, a structure is added to the alignments database if it can be reasonably aligned with the programs COMPARER or MNYFIT and differs by more than a few residues from the existing members of the database.

Some of the alignment families in the database are themselves related, e.g., the various immunoglobulin fold families. We have chosen to keep these groups separate because, although the structures can be aligned, the differences between them are substantial. Thus, the alignments would be full of uncertainties and would therefore be less useful for deriving reliable restraints for comparative modeling.

Contents and composition of the alignments

The members of 105 groups of related proteins and protein segments extracted from the PDB are listed in Table 3. The 105 alignments are classified into the following groups: small proteins (12 alignments), small proteins dominated by disulfide bonds (10), all-α (19), $\alpha+\beta$ (14), α/β (21), α/β-barrel (6), all-β (21), membrane-bound all-α (1), and membrane-bound all-β (1), although the distinction between some groups is blurred. Thus,

Table 3. *List of 105 alignments in the database*

Family		$N_{str.}$	$N_{ave.}$	$\%ID_{ave.}$	PDB codes					
Small										
1	Zinc finger—CCHC-type	2	17	47.06	**1ncpN**	**1ncpC**				
2	Zinc finger—CCHH-type	8	28	36.60	**5znf**	**3znf**	**1ard**	1bbo	**1znf**	1zaa1
					1zaa	1zaa3				
3	Metallothionein—β-domain	3	30	83.33	**2mhu**	**2mrb**	**2mrt**			
4	Metallothionein—α-domain	3	31	93.55	**1mrb**	**1mrt**	**1mhu**			
5	E3-binding domain	2	35	33.33	**1bbl**	**1pde**				
6	Pancreatic hormone	2	36	41.67	**1bba**	1ppt				
7	Rubredoxin	5	51	63.00	6rxn	1rdg	7rxn	4rxn	**1zrp**	
8	Serine proteinase inhibitor—potato I-type	2	64	35.48	2ci2	1cseI				
9	SH3 domain	4	68	30.37	**1hsp**	1shf	1shg	**1pnj**		
10	Ferredoxin (4Fe-4S)	3	72	30.53	4fd1	1fdx	1fxd			
11	High potential iron protein	2	78	23.19	2hipA	1hip				
12	Ferredoxin (2Fe-2S)	3	97	72.04	1fxiA	3fxc	1fxaA			
Small—Disulfide										
13	Serine proteinase inhibitor—squash-type	2	28	71.43	**2eti**	**1cti**				
14	Sea anemone toxin	2	45	27.03	**1bds**	**1sh1**				
15	EGF-like domain	4	46	36.26	**1ixa**	**4tgf**	**1apo**	**1epi**		
16	Insulin	3	50	52.16	4ins	2ins	6rlx			
17	Serine proteinase inhibitor—Bowman-Birk-type	3	56	74.89	1tabI	**1bbi**	1pi2			
18	Serine proteinase inhibitor—Kazal-type	5	56	44.31	1ovo	2ovo	**2bus**	1tgsI	1cgiI	
19	Serine proteinase inhibitor—Kunitz-type	4	56	39.70	5pti	1aap	**1shp**	1dtx		
20	C-module domain	2	60	37.93	**1hfh1**	**1hfh2**				
21	Snake toxin	8	64	47.12	2ctx	2abx	**1nbt**	**1nea**	**1nor**	**1ntx**
					1nxb	1cdt				
22	Kringle domain	4	86	39.98	1pk4	1tpk	**1kdu**	2pf1		
All										
23	DNA-binding homeodomain	2	62	50.88	1hddC	**1hom**				
24	DNA-binding repressor	3	71	32.59	2cro	1r69	1lrd3			
25	Steroid-binding protein	2	73	55.71	2utg	1ccd				
26	Cytochrome-c_5	3	82	39.36	**1cor**	351c	1cc5			
27	Cytochrome-*b*	2	88	29.41	3b5c	1fcbA				
28	Calcium-binding protein—parvalbumin-like	4	107	52.10	5pal	5cpv	1omd	1pal		
29	Cytochrome-*c*	7	111	44.56	1yea	1ycc	1ccr	5cyt	2c2c	1c2r
					155cA					
30	Cytochrome-c_3	2	112	35.42	1cy3	2cdv				
31	Hemerythrin	2	116	46.02	2mhr	2hmq				
32	Phospholipase A_2	6	122	46.75	1bp2	1p2p	1bbc	1pp2	1ppa	1pob
33	Cytochrome-*c'*	2	129	21.60	2ccyA	1bbhA				
34	Globin	16	146	27.12	2mm1	1pmbA	1mbs	4mbn	2hhbA	2mhbA
					1pbxA	2hhbB	2mhbB	2lhb	1mba	1sdhA
					1lh1	1ithA	1ecd	2hbg		
35	Cytokine—granulocyte colony-stimulating factor	3	153	81.85	1bgc	1bgd	1rhg			
36	Calcium-binding protein—calmodulin-like	5	162	33.44	3cln	4cln	5tnc	2scpA	1sas	
37	Fe/Mn superoxide dismutase	2	192	36.41	1abm	3sdp				
38	Glutathione *S*-transferase	4	213	37.73	1gss	1gsr	5gst	1guh		
39	Annexin	2	317	77.85	1ala	1avh				
40	Peroxidase	3	324	23.81	1arp	1lga	2cyp			
41	Cytochrome p450	2	431	17.34	2cpp	2hpd				
Membrane-bound all-α										
42	Photosynthetic reaction center	2	826	48.49	1prc	4rcr				
$\alpha+\beta$										
43	Protein G domain	2	63	87.50	1pgx	**2gb1**				
44	Histidine carrier protein	2	87	59.30	**1hid**	1ptf				
45	Ribonuclease—bacterial	3	104	61.44	1fus	1rds	1rnt			
46	FK506-binding protein	2	110	57.01	1fkb	1yat				
47	Ribonuclease—mammalian	2	124	81.45	1rbb	1bsr				
48	Lysozyme	6	128	63.90	1ghl	1hhl	1lzt	1lz3	1lz1	1alc
49	Ribonuclease H	3	141	34.01	1rnh	1ril	1hrh			

(*continued*)

Table 3. *Continued*

Family		$N_{str.}$	$N_{ave.}$	$\%ID_{ave.}$	PDB codes					
$\alpha+\beta$ (*continued*)										
50	Class 1 histocompatibility antigen binding domain	4	178	79.49	2hlaA	3hlaA	1hsaA	1vabA		
51	Cysteine proteinase	3	215	55.35	9pap	2act	1ppo			
52	Carbonic anhydrase	2	256	61.57	1ca2	2cab				
53	Thymidylate synthase	2	290	59.85	3tms	4tms				
54	Zinc metalloproteinase	3	310	44.73	3tln	1npc	1ezm			
55	Serine proteinase inhibitor—serpin-type	4	378	31.41	Xpai	1hle	1ovaA	9api		
56	Amylase	3	485	37.52	1cdg	2aaa	6taa			
α/β										
57	Thioredoxin	4	96	14.53	1aaz	**1ego**	**3trx**	2trx		
58	Flavodoxin	5	159	33.16	3fxn	1fx1	1flv	1ofv	2fcr	
59	GTP-binding protein	2	171	15.13	1etu	5p21				
60	Dihydrofolate reductase	4	172	35.96	3dfr	4dfrA	8dfr	1dhfA		
61	Nucleotide kinase	4	202	24.95	1akeA	1ak3A	3adk	1gky		
62	β-Lactamase	2	256	43.14	3blm	4blmA				
63	Ricin-like protein	2	264	28.85	1fmp	1paf				
64	Subtilase	7	274	51.88	Xesp	1st3	1sbt	1meeA	1sbc	1thm
					2prk					
65	Periplasmic binding protein—sugar	3	295	21.33	1abp	2gbp	1dri			
66	Phosphofructokinase	2	319	55.35	1pfk	4pfk				
67	Lactate/malate dehydrogenase	9	321	36.86	6ldh	1lld	9ldb	5ldh	1ldb	2ldx
					1llc	4mdh	2cmd			
68	Glyceraldehyde phosphate dehydrogenase	4	339	56.64	3gpdR	1gpdG	1ggaO	1gd1O		
69	Periplasmic binding protein—amino acid	2	345	79.07	2lbp	2liv				
70	Alcohol dehydrogenase	2	374	87.17	3hud	8adh				
71	Actin/heat-shock cognate	2	377	14.10	1atnA	Xhsc				
72	Isocitrate dehydrogenase	2	379	28.23	1ipd	3icd				
73	Aspartate aminotransferase	2	398	40.40	3aat	1ama				
74	Disulfide oxidoreductase	5	466	29.96	2tprA	3grsA	31adA	11pfA	1npx	
75	α/β-Hydrolase	2	534	27.14	1ace	1thg				
76	Cholesterol oxidase	2	541	16.40	1cox	1gal				
77	Hemocyanin	2	617	34.36	1hc1	1lla				
α/β-Barrel										
78	Tryptophan biosynthesis enzyme	2	226	10.22	1pii1	1pii2				
79	Triose phosphate isomerase	4	249	45.48	1tim	5tim	1ypi	1tre		
80	Fructose-1,6-biphosphatase aldolase	2	361	70.56	1ald	1fbaa				
81	Flavin-binding β-barrel	2	376	41.67	1gox	1fcba				
82	Xylose isomerase	3	390	66.95	4xia	6xia	1xim			
83	Ribulose-1,5-biphosphate carboxylase/oxygenase	3	537	49.57	4rub	8rub	5rubA			
All-β										
84	Immunoglobulin—cell surface—type 2	2	75	17.91	2cd42	1cid2				
85	Immunoglobulin—constant domain	11	98	34.31	2hfl	4fab	1mam	2fb4	2fbj	2fbj
					1dfb	2fb4	1fc1	1fc1	1pfc	
86	Immunoglobulin—cell surface—type 1	5	103	16.95	1cd8	2cd41	1cid1	3hlaB	1vabB	
87	Retroviral proteinase	3	104	32.65	3phv	1ivp	2rsp			
88	Azurin/plastocyanin	8	109	35.10	2azaA	1azu	1pcy	2plt	**9pcy**	7pcy
					1paz	1mdaE				
89	Antibacterial protein	3	111	43.07	2mcm	1noa	1acx			
90	Immunoglobulin—variable domain, light chain	23	112	55.95	1hil	1bbd	1mcp	1nca	1igf	4fab
					1mam	1igm	6fab	1rei	1dfb	1fdl
					3hfm	1baf	2hfl	1jhl	2fbj	1bjl
					2fb4	2rhe	2mcg	7fab	8fab	
91	Avidin	2	120	32.46	1pts	2avi				
92	Immunoglobulin—variable domain, heavy chain	20	123	52.13	1igm	1baf	3hfm	7fab	1fai	1jhl
					1fdl	1igf	1hil	2fbj	1mcp	1mam
					4fab	1dfb	2fb4	8fab	1bbd	2hfl
					6fab	1nca				
93	Interleukin 1-β-like growth factor	4	141	29.50	1i1b	1mib	2fgf	1barB		
94	Lipocalin	8	144	18.07	1mup	1rbp	1bbp	1ifb	1alb	1mdc
					2hmb	1opa				

(*continued*)

Table 3. *Continued*

Family		$N_{str.}$	$N_{ave.}$	$\%ID_{ave.}$	PDB codes					
All β (*continued*)										
95	Cu/Zn superoxide dismutase	3	152	56.19	1srd	1sdy	2sod			
96	Glucose permease	2	154	42.28	1f3g	1gpr				
97	Crystallin	4	175	57.96	4gcr	3gcrA	2gcr	2bb2		
98	Plant virus coat protein	2	186	23.64	2tbv	4sbvA				
99	Serine proteinase—bacterial	3	188	45.36	2alp	2sga	3sgbE			
100	Plant lectin	4	236	40.10	2ltn	1lte	1lec	4cna		
101	Serine proteinase—mammalian	12	238	37.56	1hneE	3est	1tbs	2ptn	1trm	2gch
					2pka	1ton	1ppb	1bbr	3rp2	1sgt
102	Aspartic proteinase	10	331	35.67	Xypa	Xren	1bbs	1lya	5pep	4cms
					1mpp	4ape	3app	2apr		
103	Neuraminidase	3	389	35.68	1nsb	1nca	2bat			
104	Picornavirus coat proteins	6	780	33.05	4rhv	1r1a	2plv	2mev	1tme	1bbt
Membrane-bound all-β										
105	Porin	2	335	64.53	1pho	1omf				

[a] The alignments are segregated into 9 groups on the basis of the structural type of member proteins. For each alignment, we show: the number of structures in it ($N_{str.}$), the average sequence length ($N_{ave.}$), the average pairwise sequence identity ($\%ID_{ave.}$), and the PDB codes of the member proteins. The fifth character in the PDB code is sometimes a PDB chain identifier, sometimes an arbitrary identifier to distinguish between different segments and domains in the same PDB data set. The code is printed in bold if the structure was determined by NMR. X as the first character in the PDB code indicates that the structure was obtained directly from the authors: Xypa, proteinase A from *Saccharomyces cerevisiae* (Carlos Aguilar & Tom Blundell); Xren, mouse renin (Dhanaraj et al., 1992); Xpai, plasminogen activator inhibitor type-1 from human (Mottonen et al., 1992); Xesp, esperase from *Bacillus lentus* (Unilever); Xhsc, heat-shock cognate from *Bos taurus* (Flaherty et al., 1990). The large 43-member family of immunoglobulin variable chains is divided into 2 groups for the purpose of statistics collection: a group of 23 light chains and a group of 20 heavy chains. In this way, the sample of all related protein pairs in the database is not dominated by the immunoglobulin family.

the database includes representatives of all the major structural classes of proteins. The 105 alignments contain 416 entries that come from 375 different PDB coordinate sets. There are 1,233 aligned entry pairs, 78,495 residues, and 230,396 pairs of equivalent alignment positions.

Of the 416 entries in the database, structures of 373 were determined by X-ray crystallography, and structures of 43 by NMR; NMR structures occur in 23 families, 8 of which consist entirely of NMR entries. The resolution and R-factor for the crystallographic analyses are stored in the alignment files so that

```
C; family:  DNA-binding repressor

>P1;2cro
structureX:2cro:  -1 :  :  63 :  :cro repressor:phage 434:  2.30:19.50
-----MQTLSERLKKRRIALK----MTQTELATKAGVKQQSIQLIEAGVTKR-PRFLFEIAMALNC-----DPVW
LQYGT-------*

>P1;1r69
structureX:1r69:  1 :  :  63 :  :repressor:phage 434:  2.00:19.30
-------SISSRVKSKRIQLG----LNQAELAQKVGTTQQSIEQLENGKTKR-PRFLPELASALGV-----SVDW
LLNGT-------*

>P1;1lrd3
structureX:1lrd:  6 :3:  92 :3:1 repressor:bacteriophage l:  2.50:24.20
PLTQEQLEDARRLKAIYEKKKNELGLSQESVADKMGMGQSGVGALFNGINALNAYNAALLAKILKVSVEEFSPSI
AREIYEMYEAVS*
```

Fig. 8. Sample alignment file in the database. An alignment of 3 DNA-binding repressors is stored in this file. The format for each entry is similar to that of the PIR sequence database. The second line of the entry contains all the information necessary to extract the atomic coordinates of the segment from the original PDB coordinate set. The fields in this line are separated by the columns and indicate the type of the method used to obtain the structure (X-ray, NMR, model, or sequence), the PDB code, the residue numbers and chain identifiers for the first and last residues in the segment, protein name, source of the protein, resolution, and R-factor of the crystallographic analysis.

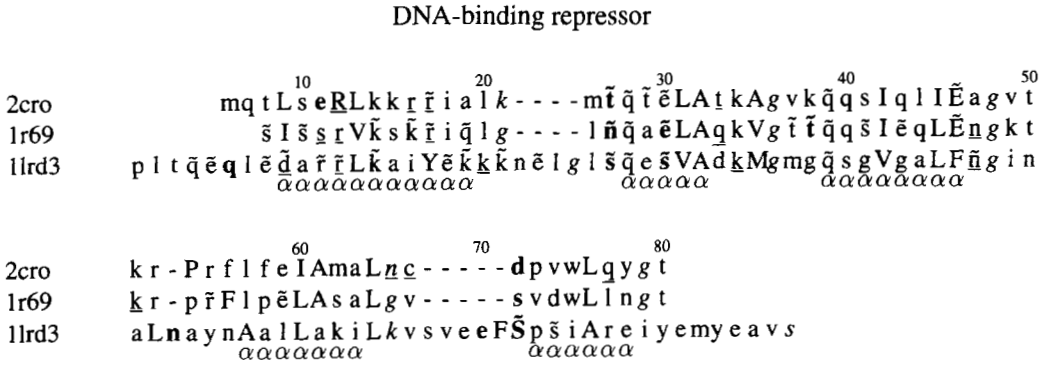

Fig. 9. Sample alignment as displayed by the program JOY. The formatting convention of JOY (Overington et al., 1990) is: uppercase letters, solvent-inaccessible amino acid residues; lowercase letters, solvent-accessible amino acid residues; bold type, hydrogen bonds to main chain amide; underline, hydrogen bonds to main chain carbonyl; ~, side chain–side chain hydrogen bonds; italic, positive main chain dihedral angle ϕ. Standard 1-letter amino acid residue symbols are used. Gaps are indicated by a hyphen. The top line shows alignment positions; the bottom line gives consensus secondary structure. Note the amphipathic nature of some helices: e.g., the buried hydrophobic residues at alignment positions 42, 45 and 60, 64. Similarly, N- and C-terminal caps of the helices (Richardson & Richardson, 1988) are readily apparent in this representation: e.g., hydrogen bonds from side chains at 27, 38, and 72 to the main chain amide at the N-terminus of a helix. In addition, several of the helices are capped at the C-terminus by a residue with the positive ϕ main chain dihedral angle, usually glycine (e.g., positions 21, 48, 65, and 79). Thus, the representation emphasizes physicochemical and conformational restraints on residue substitution in protein evolution, and can be a valuable guide in modeling and structure-based alignment studies.

only high-quality structures can be selected when so desired. The distribution of the resolution is bimodal, with peaks at 1.8 Å and 2.4 Å (Fig. 10A). There are 181 high-resolution entries (≤2 Å) and 235 medium-resolution entries (>2 Å). The alignment files do not contain any indicator of the quality of the NMR-derived structures.

The entry lengths vary widely from 17 to 852 amino acid residues, with a mean of 187 (Fig. 10B). The relative frequency of the residues in the β-strand, β-bulge, and extended conformations combined (46%) is somewhat larger than the relative frequency of the residues in any of the helical conformations (29%) (Fig. 10C). The percentage sequence identity for the 1,233 sequence pairs varies from 3% to 98% (Fig. 10D), with the average and peak at 43%. Figure 10 indicates that the database of alignments contains a representative sample of the globular proteins in the PDB and that it may be suitable for uncovering the general relationships between features of protein structure.

Multidimensional tables of protein features

We give here a brief overview of the way we explored the alignments of known protein structures to improve comparative modeling (Šali & Blundell, 1993). Central to this is an extraction of correlations among different sequence and structural features of proteins. A convenient representation of these correlations, both for their analysis and for their use in modeling, is in the form of frequency tables and conditional probability tables. We emphasize that this component of the present database is the main distinction from the other databases of protein alignments (Holm et al., 1992; Pascarella & Argos, 1992; Orengo et al., 1993), which do not include a general mechanism for extracting rules directly applicable to protein modeling.

Our comparative modeling program MODELLER relies on the satisfaction of spatial restraints on a large number of spatial features x, such as distances and dihedral angles (Šali & Blundell, 1993). These restraints may be conveniently expressed as conditional pdf's of the form

$$p(x/a, b, \ldots, c). \tag{1}$$

This conditional pdf gives a probability density for feature x when $a, b, \ldots, c$ are specified. For example, $p(\chi_1$/residue type, $\phi, \psi)$ could be used to predict the side chain dihedral angle χ_1 from the type of a residue and its main chain dihedral angles ϕ and ψ.

In reality, it is not possible to obtain the true function p, but only its approximation:

$$p(x/a, b, \ldots, c) \approx W'_{x,a,b,\ldots,c}, \tag{2}$$

where $W'_{x,a,b,\ldots,c}$ is a multidimensional table spanned by x, a, $b, \ldots, c$ that contains as its elements the observed relative frequencies for the occurrence of x given $a, b, \ldots, c$. The multidimensional table of relative frequencies $\mathbf{W}'$ is calculated from the absolute frequencies $\mathbf{W}$ using

$$W'_{x,a,b,\ldots,c} = \frac{W_{x,a,b,\ldots,c}}{\sum_x W_{x,a,b,\ldots,c}}. \tag{3}$$

The absolute frequencies, $\mathbf{W}$, are obtained directly by the program MDT, which counts the number of occurrences of each combination of $(x, a, b, \ldots, c)$ in the sample. The sample is derived from a database of known protein structures and their alignments, depending on the type of features that are correlated among themselves. For example, if only a distribution of residue types is wanted, then the sample consists of all amino acid residues in the database; if C^α–C^α distances in one protein are correlated with C^α–C^α distances in another protein, then the sample includes all pairs of equivalent C^α–C^α distances in all homologous pairs of proteins in the database. Associations

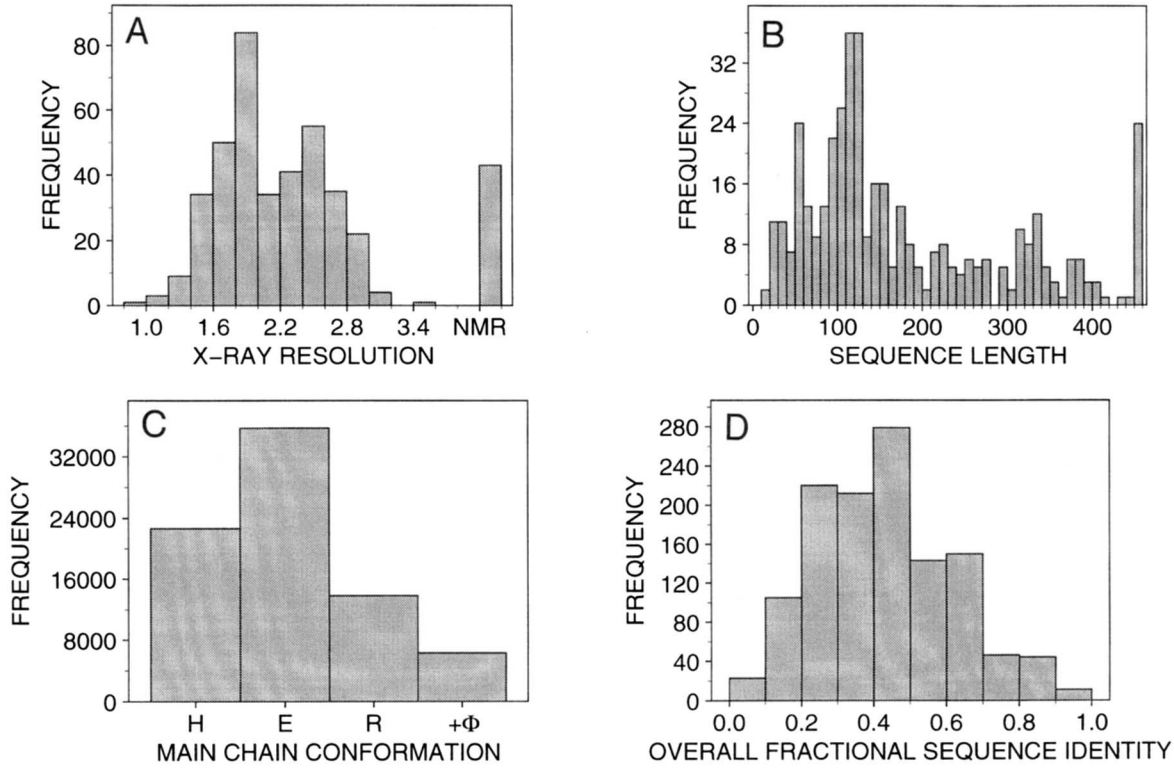

Fig. 10. Composition of the database. Distributions of various features in the alignments database are shown. **A:** Resolution of X-ray analysis of an entry for all 416 entries. The bar labeled NMR indicates the 43 structures determined by NMR. **B:** Number of amino acid residues in an entry. The last bar combines all entries with more than 450 residues. **C:** Numbers of residues in different secondary structure states. There are 78,495 residues in the whole database. If the ϕ angle is positive, the main chain conformation is assigned to class $+\phi$; otherwise the secondary structure assignments from the PROCHECK program (Laskowski et al., 1993), which implements the algorithm in the DSSP program (Kabsch & Sander, 1983), are used to select 1 of the 3 remaining classes: H, helical (Kabsch & Sander codes: H, α-helix; G, 3_{10}-helix; I, π-helix); E, extended (E, strand in a β-sheet; B, β-bulge; a blank, extended chain); and R, other (T, turn; S, bend). **D:** Fractional sequence identity of a pair of related entries for all 1,233 such pairs. Fractional sequence identity is calculated as the number of identical amino acid residues divided by the length of the shorter sequence.

among features of 2 related proteins are crucial for comparative modeling. Consequently, for each feature type, the MDT program distinguishes at least 2 "values"; the first value is a feature associated with the first protein in a pairwise alignment and the second value is the same feature associated with the second protein in the alignment. These 2 proteins would be treated as the template and the target in prediction, but at this stage both structures are known. To correlate features from 3 proteins, MDT can also use all triple alignments that can be obtained from the multiple alignment of 3 or more structures (see Šali & Blundell, 1993, for an application). For a summary of the protein features and their symbols that can be selected in MDT, see Table 4. Several features are defined in Figure 10. For detailed definitions, see Šali (1991) and Šali and Blundell (1993).

The main supporting programs for the database of alignments include protein structure comparison programs COMPARER (Šali & Blundell, 1990) and MNYFIT (Sutcliffe et al., 1987); the KITSCH program for clustering of protein sequences and structures (Felsenstein, 1985); program HBOND for calculating hydrogen bonds (Overington et al., 1990); program PSA for solvent accessibility (Richmond & Richards, 1978; Šali & Blundell, 1990); program DIH for main chain and side chain dihedral angles (Šali & Blundell, 1993); program NGH for residue neighbors (Šali & Blundell, 1993); program PROCHECK for secondary structure assignments (Laskowski et al., 1993) using the algorithm of Kabsch and Sander (1983); the JOY program for displaying alignments (Overington et al., 1990); programs MDT and PLOT for scanning the database and processing the pdf's (Šali & Blundell, 1993); and program LSQ for nonlinear least-squares fitting (Press et al., 1992; Šali & Blundell, 1993). These programs can be extended to a large number of different analyses. All that is needed to explore a new feature in relation to other features is to add a function that defines the new feature.

Strength of associations among the features of protein structure

The most useful pdf for modeling is that which predicts the unknown feature most accurately. Provided that pdf's are not constructed from a sparse and nonrepresentative database, the most precise pdf is on the average also the most accurate pdf; therefore, the most accurate pdf is the pdf with the sharpest shape. A quantitative measure of sharpness of any distribution is its entropy

Table 4. *Features that may be selected in MDT to span multidimensional frequency table* ***W***

Variable	Feature
r	Amino acid residue type
Φ, $\Delta\Phi$	Main chain dihedral angle Φ
Φ_c	Main chain dihedral angle Φ class
Ψ, $\Delta\Psi$	Main chain dihedral angle Ψ
Ψ_c	Main chain dihedral angle Ψ class
ω, $\Delta\omega$	Main chain dihedral angle ω
ω_c	Main chain dihedral angle ω class
β_i	Side chain dihedral angle χ_i, $i = 1, 2, 3, 4, 5$
c_i	Side chain dihedral angle χ_i class, $i = 1, 2, 3, 4, 5$
t	Secondary structure class of a residue (positive Φ, α, β, other)
M	Main chain conformation class of a residue (Wilmot & Thornton, 1990)
α	Fractional content of residues in the main chain conformation class A
S	Side chain conformation class (χ_1, χ_2)
a, $\bar{a}$	(Fractional) contact solvent area of a residue
s, $\bar{s}$	Residue neighborhood difference between 2 proteins
i	Fractional sequence identity between 2 proteins
d, Δd	Distance between 2 specified atom types
b	Average residue isotropic temperature factor
R	Resolution of X-ray analysis
n	Number of atomic contacts with nonprotein nonwater atoms per residue
g, $\bar{g}$	Distance of a residue from a gap in the alignment
l	Number of residues in the protein
G	Several residue type groups (e.g., hydrophobic/hydrophilic)

[a] The first column lists the variable names that are used for these features. It also indicates whether an intramolecular average or intermolecular difference can be calculated. The overbar indicates an average of the feature at 2 residue positions in the same protein, such as an average accessibility of a certain residue pair. Features that are not associated with 2 proteins can be used independently for 2 related proteins in a pairwise alignment or for 3 related proteins in a triple alignment. For example, a 2D table can be constructed that is spanned by a residue type r in one protein and a residue type r' at the equivalent position in a related protein; the prime is generally used to designate that the feature is from the second protein and 2 primes that it is from the third protein. The Δ symbol refers to the difference between features f and f': $\Delta f = f - f'$.

$$S[p(x)] = -\sum_i p(x_i)\ln p(x_i). \quad (4)$$

For a discrete conditional probability distribution, entropy is defined similarly as

$$S[p(x/a, b, \ldots, c)] = \sum_{a,b,\ldots,c} p(a, b, \ldots, c)S[p(x/a, b, \ldots, c)]. \quad (5)$$

Thus, to find the known features $(a, b, \ldots, c)$ that are best for the prediction of the unknown feature x, we search for the features that minimize entropy S of a corresponding conditional pdf. A convenient measure of how much the independent features determine the dependent feature is given by the uncertainty coefficient of x (Press et al., 1992):

$$U(x/a, b, \ldots, c) = \frac{S[p(x)] - S[p(x/a, b, \ldots, c)]}{S[p(x)]}. \quad (6)$$

This measure lies between 0 and 1. The value 0 means that x is not associated with $(a, b, \ldots, c)$, and the value 1 implies that $(a, b, \ldots, c)$ completely determine x.

Acknowledgments

We thank Martin Karplus, Tom L. Blundell, and Mark Johnson for discussing the work described in this paper. We are also grateful to Roman Laskowski and Janet Thornton for giving us the PROCHECK program, and to crystallographers Carlos Aguilar, Chris DeAlwis, Elizabeth Goldsmith, and David McKay for making the protein structures available before their release to the PDB. We thank Dasa Šali for making useful comments on the manuscript. A.Š. is a Fellow of The Jane Coffin Childs Memorial Fund for Medical Research. This investigation has been aided by a grant from The Jane Coffin Childs Memorial Fund for Medical Research (A.Š.). The computations were performed on a NeXTstation workstation.

References

Abola EE, Bernstein FC, Bryant SH, Koetzle TF, Weng J. 1987. Protein Data Bank. In: Allen FH, Bergerhoff G, Sievers R, eds. *Crystallographic databases—Information, content, software systems, scientific applications*. Bonn/Cambridge/Chester: Data Commission of the International Union of Crystallography. pp 107–132.

Akrigg D, Bleasby AJ, Dix NIM, Findlay JBC, North ACT, Parry-Smith D, Wooton JC, Blundell TL, Gardner SP, Hayes F, Islam S, Sternberg MJE, Thornton JM, Tickle IJ. 1988. A protein sequence/structure database. *Nature 335*:745–746.

Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. 1977. The Protein Data Bank: A computer-based archival file for macromolecular structures. *J Mol Biol 112*:535–542.

Bowie JU, Lüthy R, Eisenberg D. 1991. A method to identify protein sequences that fold into a known three-dimensional structure. *Science 253*:164–170.

Browne WJ, North ACT, Phillips DC, Brew K, Vanaman TC, Hill RC. 1969. A possible three-dimensional structure of bovine α-lactalbumin based on that of hen's egg-white lysozyme. *J Mol Biol 42*:65–86.

Bryant SH. 1989. PKB: A program system and data base for analysis of protein structure. *Proteins Struct Funct Genet 5*:233–247.

Burks HS, Burks C. 1988. The Genbank sequence data bank. *Nucleic Acids Res 15*:1861–1864.

Chothia C. 1992. One thousand families for the molecular biologist. *Nature 360*:543–544.

Chothia C, Lesk AM. 1986. The relation between the divergence of sequence and structure in proteins. *EMBO J 5*:823–826.

Chou PY, Fasman GD. 1974. Prediction of protein conformation. *Biochemistry 13*:222–245.

Cohen FE, Kuntz ID. 1989. Tertiary structure prediction. In: Fasman GD, ed. *Prediction of protein structure and the principles of protein conformation*. New York: Plenum Press. pp 647–705.

Dayhoff MO, Schwartz RM, Orcutt BC. 1978. In: Dayhoff MO, ed. *Atlas of protein sequence and structure, vol 5, suppl 3*. Washington, D.C.: National Biomedical Research Foundation, pp 345–352.

Dhanaraj R, DeAlwis C, Frazao C, Badasso M, Sibanda BL, Tickle IJ, Cooper JB, Driessen HPC, Newman M, Aguilar C, Wood SP, Blundell TL, Hobart PM, Geoghegan KF, Ammirati MJ, Danley DE, O'Connor BA, Hoover DJ. 1992. X-ray analyses of peptide-inhibitor complexes define the structural basis of specificity for human and mouse renins. *Nature 357*:466–472.

Dunbrack RL, Karplus M. 1993. Prediction of protein sidechain conformations from a backbone conformation dependent rotamer library. *J Mol Biol 230*:543–571.

Engh RA, Huber R. 1991. Accurate bond and angle parameters for X-ray protein structure refinement. *Acta Crystallogr A 47*:392–400.

Felsenstein J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution 39*:783–791.

Finkelstein AV, Reva BA. 1991. A search for the most stable folds of protein chains. *Nature 351*:497–499.

Flaherty KM, DeLuca-Flaherty C, McKay DB. 1990. Three-dimensional

structure of the ATPase fragment of a 70K heat-shock cognate protein. *Nature 346*:623–628.
George DG, Barker WC, Hunt LT. 1986. The protein identification resource. *Nucleic Acids Res 14*:11–15.
Godzik A, Kolinski A, Skolnick J. 1992. Topology fingerprint approach to the inverse protein folding problem. *J Mol Biol 227*:227–238.
Gribskov M, McLachlen AD, Eisenberg D. 1987. Profile analysis: Detection of distantly related proteins. *Proc Natl Acad Sci USA 84*:4355–4358.
Hamm G, Cameron G. 1986. The EMBL data library. *Nucleic Acids Res 14*:5–9.
Heinz DW, Baase WA, Dahlquist FW, Matthews BW. 1993. How amino-acid insertions are allowed in an alpha-helix of T4 lysozyme. *Nature 361*: 561–564.
Holm L, Ouzounis C, Sander C, Tuparev G, Vriend G. 1992. A database of protein structure families with common folding motifs. *Protein Sci 1*:1691–1698.
Hubbard TJP, Blundell TL. 1987. Comparison of solvent inaccessible cores of homologous proteins: Definitions useful for protein modelling. *Protein Eng 1*:159–171.
Huysmans M, Richelle J, Wodak SJ. 1991. SESAM: A relational database for structure and sequence of macromolecules. *Proteins Struct Funct Genet 11*:59–76.
Islam SA, Sternberg MJE. 1989. A relational database of protein structures designed for flexible enquires about conformation. *Protein Eng 2*:431–442.
Janin J, Wodak S, Levitt M, Maigret B. 1978. Conformation of amino acid side-chains in proteins. *J Mol Biol 125*:357–386.
Johnson MS, Overington JP, Blundell TL. 1993. Alignment and searching for common protein folds using a data bank of structural templates. *J Mol Biol 231*:735–752.
Jones DT, Taylor WR, Thornton JM. 1992. A new approach to protein fold recognition. *Nature 358*:86–89.
Kabsch W, Sander C. 1983. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers 22*:2577–2637.
Laskowski RA, McArthur MW, Moss DS, Thornton JM. 1993. PROCHECK: A program to check the stereochemical quality of protein structures. *J Appl Crystallogr 26*:283–291.
Lüthy R, Bowie JU, Eisenberg D. 1992. Assessment of protein models with three-dimensional profiles. *Nature 356*:83–85.
Lüthy R, McLachlan AD, Eisenberg D. 1991. Secondary structure-based profiles: Use of structure-conserving scoring tables in searching protein sequence databases for structural similarities. *Proteins Struct Funct Genet 10*:229–239.
MacArthur MW, Thornton JM. 1991. Influence of proline residues on protein conformation. *J Mol Biol 218*:397–412.
Manavalan P, Ponnuswamy PK. 1978. Hydrophobic character of amino acid residues in globular proteins. *Nature 275*:673–674.
Miyazawa S, Jernigan RL. 1985. Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation. *Macromolecules 18*:534–552.
Mottonen J, Strand A, Symersky J, Sweet RM, Danley DE, Geoghegan KF, Gerard RD, Goldsmith EJ. 1992. Structural basis of latency in plasminogen activator inhibitor-1. *Nature 355*:270–273.
Orengo CA, Flores TP, Taylor WR, Thornton JM. 1993. Identification and classification of protein fold families. *Protein Eng 6*:485–500.
Overington J, Donnelly D, Johnson MS, Šali A, Blundell TL. 1992. Environment-specific amino acid substitution tables: Tertiary templates and prediction of protein folds. *Protein Sci 1*:216–226.
Overington J, Johnson MS, Šali A, Blundell TL. 1990. Tertiary structural constraints on protein evolutionary diversity: Templates, key residues and structure prediction. *Proc R Soc Lond B 241*:132–145.
Overington JP, Zhu ZY, Šali A, Johnson MS, Sowdhamini R, Louie GV, Blundell TL. 1993. Molecular recognition in protein families: A database of aligned three-dimensional structures of related proteins. *Biochem Soc Trans 21*:597–604.
Pabo CO, Suchanek EG. 1986. Computer-aided model-building strategies for protein design. *Biochemistry 25*:5987–5991.
Pascarella S, Argos P. 1992. A data bank merging related protein structures and sequences. *Protein Eng 5*:121–137.
Ponder JW, Richards FM. 1987. Tertiary templates for proteins: Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J Mol Biol 193*:775–791.
Press WH, Teukolsky SA, Vetterling WT, Flannery BP. 1992. *Numerical recipes, 2nd ed.* Cambridge, UK: Cambridge University Press.
Qian W, Krimm S. 1993. Energetics of the disulphide bridge: An ab initio study. *Biopolymers 33*:1591–1603.
Richardson JS. 1981. The anatomy and taxonomy of protein structure. *Adv Protein Chem 34*:167–339.
Richardson JS, Richardson DC. 1988. Amino acid preferences for specific locations at the ends of α-helices. *Science 240*:1648–1652.
Richmond TJ, Richards FM. 1978. Packing of α-helices: Geometrical constraints and contact areas. *J Mol Biol 119*:537–555.
Šali A. 1991. Modelling three-dimensional structure of proteins from their sequence of amino acid residues [thesis]. London: University of London.
Šali A, Blundell TL. 1990. Definition of general topological equivalence in protein structures: A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. *J Mol Biol 212*:403–428.
Šali A, Blundell TL. 1993. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol 234*:779–815.
Šali A, Overington JP, Johnson MS, Blundell TL. 1990. From comparisons of protein sequences and structures to protein modelling and design. *Trends Biochem Sci 15*:235–240.
Schrauber H, Eisenhaber F, Argos P. 1993. Rotamers: To be or not to be? An analysis of amino acid side-chain conformations in globular proteins. *J Mol Biol 230*:592–612.
Sippl MJ. 1990. Calculation 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.
Sippl MJ, Weitckus S. 1992. Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a data base of known protein conformations. *Proteins Struct Funct Genet 13*:258–271.
Sowdhamini R, Ramakrishnan C, Balaram P. 1993. Modelling multiple disulphide loop containing polypeptides by random conformation generation. The test cases of α-conotoxin GI and endothelin. *Protein Eng 6*:873–882.
Sowdhamini R, Srinivasan N, Shoichet B, Santi DV, Ramakrishnan C, Balaram P. 1989. Stereochemical modeling of disulphide bridges. Criteria for introduction into proteins by site-directed mutagenesis. *Protein Eng 3*:95–103.
Stewart DE, Sarkar A, Wampler JE. 1990. Occurrence and role of *cis* peptide bonds in protein structures. *J Mol Biol 214*:253–260.
Sutcliffe MJ, Haneef I, Carney D, Blundell TL. 1987. Knowledge based modelling of homologous proteins, Part 1: Three dimensional frameworks derived from the simultaneous superposition of multiple structures. *Protein Eng 1*:377–384.
Taylor WR. 1991. Towards protein tertiary fold prediction using distance and motif constraints. *Protein Eng 4*:853–870.
Thornton JM. 1981. Disulphide bridges in globular proteins. *J Mol Biol 151*:261–287.
Thornton JM, Gardner SP. 1990. Protein motifs and data-base searching. In: Bradshaw RA, Purton M, eds. *Proteins: Form and function*. Cambridge, UK: Elsevier Science Publishers. pp 153–161.
Topham CM, McLeod A, Eisenmenger F, Overington JP, Johnson MS, Blundell TL. 1993. Fragment ranking in modelling of protein structure. Conformationally constrained environmental amino acid substitution tables. *J Mol Biol 229*:194–220.
Wilmot CM, Thornton JM. 1990. β-Turns and their distortions: A proposed new nomenclature. *Protein Eng 3*:479–493.
Zhu ZY, Šali A, Blundell TL. 1992. A variable gap penalty function and feature weights for protein 3-D structure comparisons. *Protein Eng 5*:43–51.