Protein Engineering vol. 16 no. 6 pp. 407–414, 2003
DOI: 10.1093/protein/gzg052

Have we seen all structures corresponding to short protein fragments in the Protein Data Bank? An update

Peicheng Du, Michael Andrec and Ronald M.Levy[1]

Department of Chemistry and Chemical Biology and BioMaPS Institute, Rutgers, the State University of New Jersey, 610 Taylor Road, Piscataway, NJ 08854-8087, USA

[1]To whom correspondence should be addressed.
E-mail: ronlevy@lutece.rutgers.edu

Assembling short fragments from known structures has been a widely used approach to construct novel protein structures. To what extent there exist structurally similar fragments in the database of known structures for short fragments of a novel protein is a question that is fundamental to this approach. This work addresses that question for seven-, nine- and 15-residue fragments. For each fragment size, two databases, a query database and a template database of fragments from high-quality protein structures in SCOP20 and SCOP90, respectively, were constructed. For each fragment in the query database, the template database was scanned to find the lowest r.m.s.d. fragment among non-homologous structures. For seven-residue fragments, there is a 99% probability that there exists such a fragment within 0.7 Å r.m.s.d. for each loop fragment. For nine-residue fragments there is a 96% probability of a fragment within 1 Å r.m.s.d., while for 15-residue fragments there is a 91% probability of a fragment within 2 Å r.m.s.d.. These results, which update previous studies, show that there exists sufficient coverage to model even a novel fold using fragments from the Protein Data Bank, as the current database of known structures has increased enormously in the last few years. We have also explored the use of a grid search method for loop homology modeling and make some observations about the use of a grid search compared with a database search for the loop modeling problem.

Keywords: database/fold/loop modeling/protein fragment/ protein structure

Introduction

Modeling of protein structure based on sequence and structural homology or limited experimental data can make use of either systematic search of conformational space (Deane and Blundell, 2000), use of spatial restraints (Fiser *et al.*, 2002) or databases of fragments of known proteins. The fragment database approach dates back to 1986, when retinal binding protein was reconstructed by choosing fragments from only three other proteins (Jones and Thirup, 1986). Since then, protein fragment databases have been used to build complete protein backbone structures (Reid and Thornton, 1989; Correa, 1990; Summers and Karplus, 1990; Holm and Sander, 1991; Levitt, 1992) or serve as candidates for loop modeling [e.g. (van Vlijman and Karplus, 1997; Wojcik *et al.*, 1999)]. The protein structure prediction program ROSETTA (Simons *et al.*, 1999), which was used successfully in the CASP competition, also built structures by assembling short fragments.

As an example of an application of fragment-based modeling, we recently determined the backbone structure of ubiquitin using limited NMR residual dipolar coupling data (Andrec *et al.*, 2001; 2002). In our method, we selected a small number of seven-residue fragments which fit well to the experimental data from a library of nearly 200 000 fragments drawn from the SCOP40 database (Murzin *et al.*, 1995; Brenner *et al.*, 2000; Chandonia *et al.*, 2002). This was done for all overlapping seven-residue data windows. The chosen fragments were subjected to a filtering procedure to maximize their structural similarity over overlapping regions of sequence and the complete protein backbone structure was built by superimposing the selected fragments (Figure 1).

To construct the structure of a novel protein from a database of protein fragments, it is assumed that there exists at least one fragment from a known protein structure which is structurally similar to each fragment of the novel protein. The validity of this assumption is the focus of this study. Previous work based on an all-against-all comparison of a database of 2743 seven-residue fragments from 57 high-resolution non-homologous proteins concluded that there was a 96% probability of finding a fragment from a non-homologous structure with a Cα root mean square deviation (r.m.s.d.) <1 Å and that this probability decreased significantly for longer fragments (Fidelis *et al.*, 1994). Lessel and Schomburg later studied fragments of length 3–12 residues in the Protein Data Bank (PDB) using a clustering approach: two fragments were grouped to the same cluster if the difference between the distance of the first and last Cα atoms was <1.6 Å and if the Cα r.m.s.d. between two fragments was <0.8 Å (Lessel and Schomburg, 1997). Based on the number of members in the clusters they concluded that the database was complete only for three and four residue fragments and was incomplete for longer fragments. Since the number of structures in the PDB (Berman *et al.*, 2000) has increased rapidly in recent years, the previous conclusions need to be updated. In this work, we examine the degree to which a protein fragment can be found in a database of non-homologous protein structures using current, much larger databases. In addition, we have performed conformational searching in the context of loop modeling using a grid search method and explore the merits of this strategy relative to a database approach for loop homology modeling.

Methods

For seven-residue fragments, two protein structure databases were constructed for purposes of this study: a query database, which is used to represent fragments of unknown novel structures, and the template database, which is used to represent fragments of known structures. The query and template databases are derived from the SCOP20 and SCOP90 databases, version 1.53 (Brenner *et al.*, 2000;

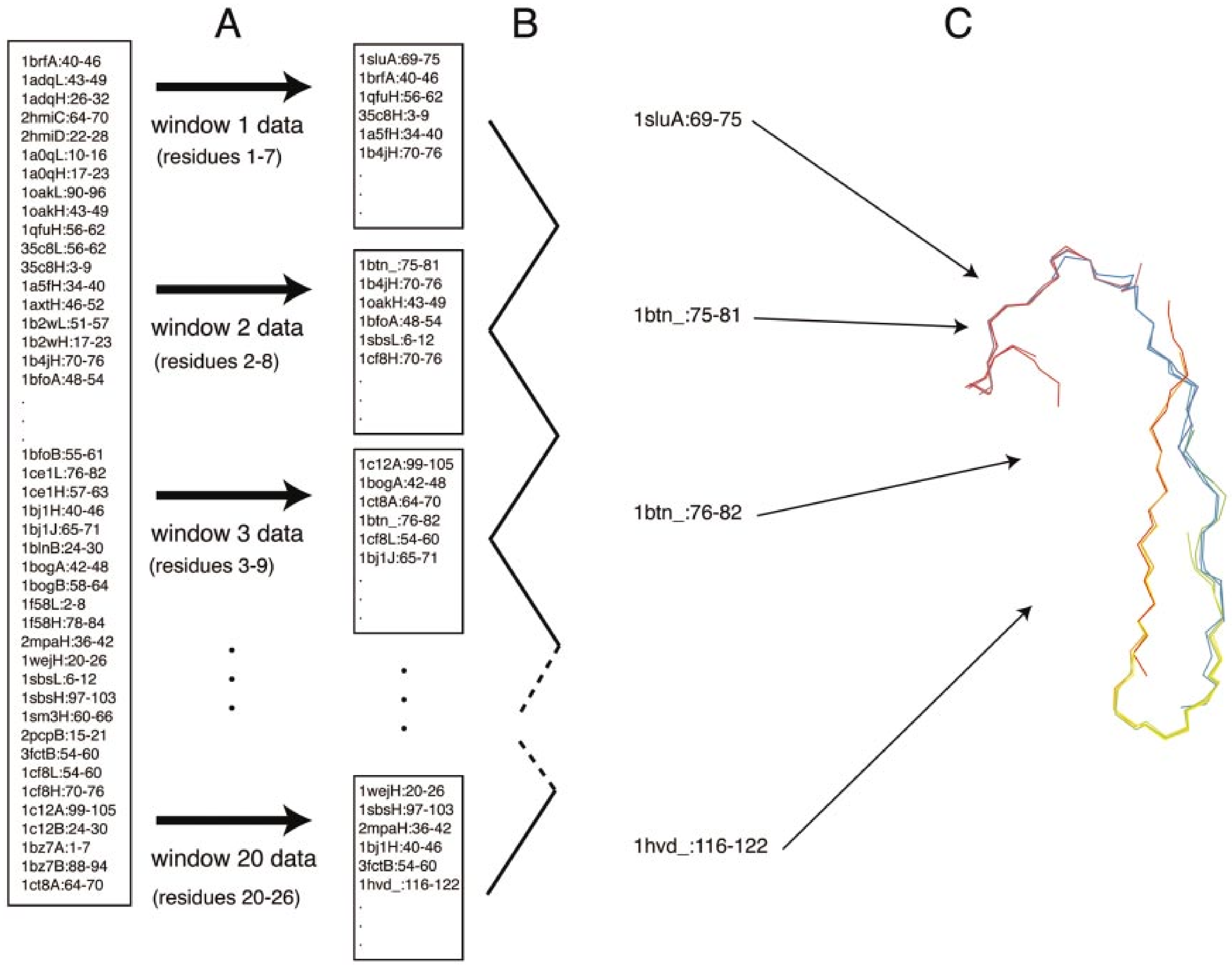

Fig. 1. An example of the use of a protein fragment database in the construction of a backbone model based on limited experimental data (Andrec *et al.*, 2001, 2002). The complete database consisting of nearly 200 000 seven-residue fragments (represented by the left-most box) was first filtered in step A using experimental NMR residual dipolar coupling data. These were used to select the 15 fragments from the complete database which best agreed with the experimental data for each seven-residue window in the protein sequence. One best representative from each of these sets was then chosen in step B in such a way that maximized the structural similarity in overlapping regions of the sequence. These representatives were then combined in step C by rotating each into a common reference frame to minimize the r.m.s.d.. In subsequent steps (not shown), the overlapping fragments were converted into a consensus backbone model and the resulting model was further refined with respect to the NMR data.

Chandonia *et al.*, 2002). These databases consist of domains from the complete SCOP database (Murzin *et al.*, 1995) selected so that no pair of domains has more than 20 or 90% sequence identity, respectively. Our databases are constructed using only those structures with an *R*-factor of <20% and resolution better than 2 Å. The query database contains 34 205 seven-residue fragments from 172 domains selected from SCOP20 so that no pair of domains belong to the same fold, while the much larger template database contains 174 914 seven-residue fragments from 955 domains selected from SCOP90. The fragments in both the template and query databases can be overlapping, e.g. residues 1–7, 2–8 and 3–9 on the same peptide chain can all be in the database.

The Fidelis *et al.* study (Fidelis *et al.*, 1994) considered only loop structures, which they defined as seven-residue fragments with fewer than four continuous α-helical or β-strand residues as defined by DSSP (Kabsch and Sander, 1983). In this study, a loop fragment is defined in exactly the same way. Furthermore, we define an α fragment to be a seven-residue fragment with four or more continuous α-helical residues and a β fragment to be a seven-residue fragment with four or more continuous β-strand residues. The percentage of α, β or loop fragments in our query database is 33.7, 18.8 and 47.6%, respectively. For the template database, the corresponding percentages are 32.3, 19.5 and 48.2%, respectively.

The goal of this work was to study the distribution of the r.m.s.d.s of the most similar fragment in the template database (the 'nearest neighbor fragment') for every fragment in the query database. The structures of protein fragments were compared by calculating the r.m.s.d. after optimal superimposition of the Cα atoms (Kabsch, 1976; McLachlan 1979). Since we wish our results to be relevant even in the case where the unknown structure corresponds to a previously unobserved fold, for each fragment in the query database we eliminate from

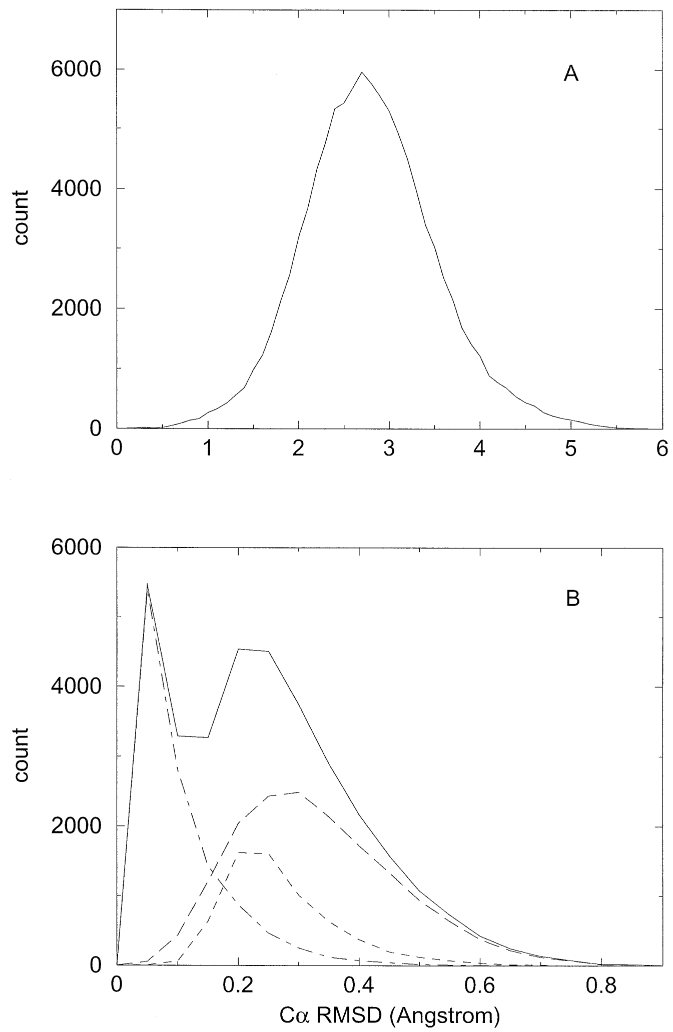

Fig. 2. (**A**) Histogram of the distribution of the Cα r.m.s.d. of 99 995 pairs of seven-residue loop fragments randomly chosen from the template database. The bin width is 0.1 Å. (**B**) Histogram of the distribution of nearest neighbor Cα r.m.s.d.s in the template database for seven-residue fragments in the query database. The solid, dot-dashed, short dashed and long dashed curves are the distributions for all fragments, α, β and loop fragments, respectively. The bin width is 0.05 Å. The height for each bin is the number of fragments in the query database whose nearest neighbor Cα r.m.s.d.s in the template database fall in that bin. The total number of fragments in the query database is 34 205. Among them, 11 513 are α fragments, 6424 are β fragments and 16 268 are loop fragments. The total number of fragments in the template database is 174 914. Among them, 56 535 are α fragments, 34 024 are β fragments and 84 355 are loop fragments.

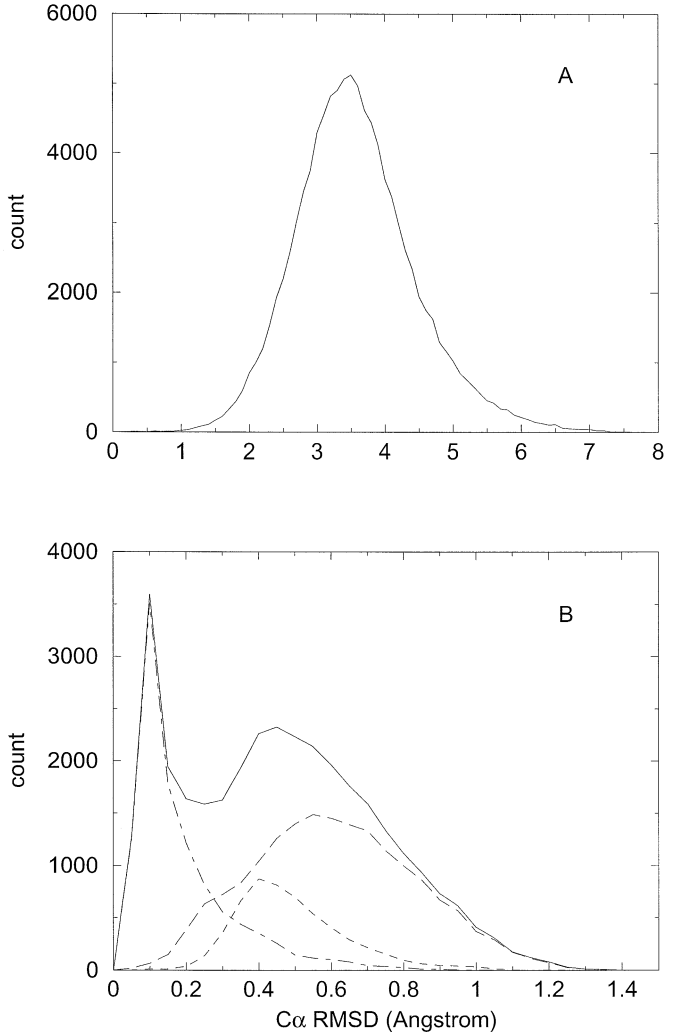

Fig. 3. (**A**) Histogram of the distribution of the Cα r.m.s.d. of 99 994 pairs of nine-residue loop fragments randomly chosen from the template database. The bin width is 0.1 Å. (**B**) Histogram of the distribution of nearest neighbor Cα r.m.s.d.s in the template database for nine-residue fragments in the query database. The solid, dot-dashed, short dashed and long dashed curves are the distributions for all fragments, α, β and loop fragments, respectively. The bin width is 0.05 Å. The height for each bin is the number of fragments in the query database whose nearest neighbor Cα r.m.s.d. in the template database fall in that bin. The total number of fragments in the query database is 33 775. Among them, 10 775 are α fragments, 5513 are β fragments and 17 487 are loop fragments. The total number of fragments in the template database is 172 345. Among them, 52 804 are α fragments, 28 783 are β fragments and 90 758 are loop fragments.

consideration all fragments in the template database which come from domains belonging to the same fold according to the SCOP classification (Murzin *et al.*, 1995) as the domain from which the query fragment is derived.

Fragments of nine and 15 residues were studied in the same manner as seven-residue fragments, except that the definitions of loop, α and β fragments were adjusted for the longer size. In particular, an α fragment was defined to be a nine-residue fragment with five or more continuous α-helical residues or a 15-residue fragment with eight or more continuous α-helical residues. A β fragment was defined to be a nine-residue fragment with five or more continuous β-strand residues or a 15-residue fragment with eight or more continuous β-strand residues. Loop fragments were defined to be all fragments that are neither α fragments nor β fragments. These definitions serve the purpose of roughly separating the contributions from fragments of different secondary structures.

Results

For seven residues, Figure 2A shows the distribution of the Cα r.m.s.d. for 99 995 pairs of loop fragments randomly chosen from the template database such that the two fragments do not come from domains of the same SCOP fold. These random pair r.m.s.d.s have a bell-shaped distribution ranging from 0 to almost 6 Å, with the peak at 2.8 Å. This distribution is almost identical with Fidelis *et al.*'s result of an all-against-all comparison of loop fragments [their figure 1A (Fidelis *et al.*, 1994)]. Figure 2B shows the distribution of the r.m.s.d.s of the nearest neighbor fragments in the template database for each fragment in the query database (solid curve). This distribution has a sharp peak at 0.05 Å, a broader peak at ~0.25 Å and a long tail extending to 1.0 Å. When that distribution is decomposed according to the relative contributions from loop, α and β query fragments (long dashed, dot-dashed and short dashed curves, respectively), it becomes obvious that the sharp peak at 0.05 Å is due to the α query fragments, as the distribution of nearest neighbor r.m.s.d.s of the α fragments shows a peak that overlaps with the 0.05 Å peak of the solid curve. The distribution of nearest neighbor r.m.s.d.s for loop fragments is a broad curve with a peak at 0.3 Å. Beyond 0.6 Å, the tail of this distribution almost overlaps the tail of the solid curve, which is consistent with the commonly held view that loops are irregular structures and have fewer close neighbors in the structural database. In comparison, Fidelis *et al.*'s distribution of nearest neighbor r.m.s.d.s for loop fragments has a peak at 0.5–0.6 Å and has a tail extending to 1.4 Å (Fidelis *et al.*, 1994). This significant improvement in the peak position and the upper tail over the results obtained by Fidelis *et al.* in 1994 is a reflection of the vastly increased structural diversity of today's PDB and suggests that there is sufficient coverage in the present database to construct models for novel structures based on existing protein fragments. The distribution of nearest neighbor r.m.s.d.s for β query fragments is also broadly distributed, with a peak at 0.2–0.25 Å and a long tail extending to 0.8 Å. The solid curve in Figure 5 shows the cumulative probability of nearest neighbor r.m.s.d.s for all loop fragments in the query database. The height of the curves in Figure 5 at an r.m.s.d. value of x is equal to the percentage of the query fragments whose nearest neighbor r.m.s.d. is $\leqslant x$. If the threshold for structural similarity is set to 1 Å, then we are virtually guaranteed to find a fragment in the template database which has a similar structure to a given query database fragment, even if we insist that the two fragments come from domains with different folds. With a stricter similarity threshold of 0.7 Å, the probability is still 99.3%.

For nine-residue fragments, the distribution of Cα r.m.s.d. of fragment pairs randomly chosen from the database is also bell-shaped and is centered at 3.5 Å (Figure 3A). The centers of distribution of nearest neighbor r.m.s.d.s for α fragments, β fragments and loop fragments are 0.1, 0.4 and 0.5–0.6 Å, respectively (Figure 3B). The loop fragments account for almost all the upper tail beyond 0.9 Å, which extends to 1.5 Å. From the cumulative probability plot (dashed curve in Figure 5), we see that for 96% of the loop fragments in the query database, a fragment exists in the template database which is within 1 Å Cα r.m.s.d. of the query fragment, which happens to equal Fidelis *et al.*'s results for seven-residue fragments (Fidelis *et al.*, 1994). Compared with the seven-residue results, it seems the additional structural complexity introduced by two more residues is compensated by the expansion of the structure database since 1994.

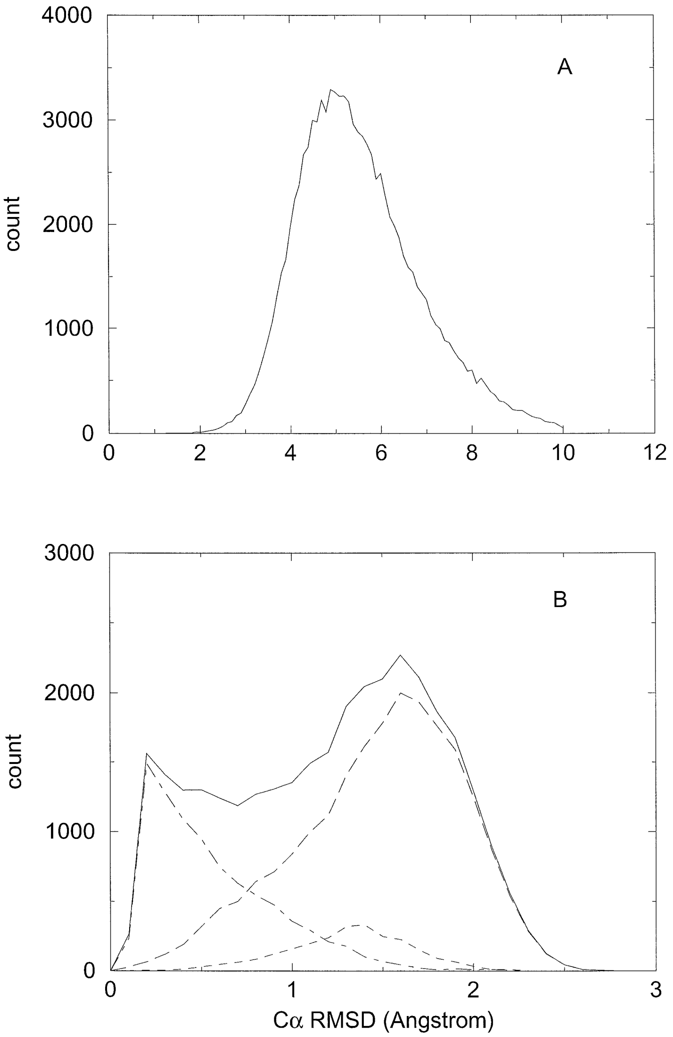

Fig. 4. (**A**) Histogram of the distribution of the Cα r.m.s.d. of 99 385 pairs of 15-residue loop fragments randomly chosen from the template database. The bin width is 0.1 Å. (**B**) Histogram of the distribution of nearest neighbor Cα r.m.s.d.s in the template database for 15-residue fragments in the query database. The solid, dot-dashed, short dashed and long dashed curves are the distributions for all fragments, α, β and loop fragments, respectively. The bin width is 0.1 Å. The height for each bin is the number of fragments in the query database whose nearest neighbor Cα r.m.s.d. in the template database fall in that bin. The total number of fragments in the query database is 32 499. Among them, 8781 are α fragments, 2483 are β fragments and 21 235 are loop fragments. The total number of fragments in the template database is 164 842. Among them, 43 411 are α fragments, 11 806 are β fragments and 109 625 are loop fragments.

The chance that there exists a similar structure in the database for each fragment decreases for 15-residue fragments. The distribution of Cα r.m.s.d. of fragment pairs randomly chosen from the database is still bell-shaped (Figure 4A) and is centered at 5.0 Å. The centers of the distributions of nearest neighbor r.m.s.d.s for α, β and loop fragments are 0.2, 1.3–1.4 and 1.6 Å, respectively (Figure 4B). Since a β fragment must have eight or more continuous β residues and the typical length of a β strand is approximately six residues, the number of 15-residue β fragments is small relative to the seven- or nine-residue case. The upper tail beyond 1.6 Å, which extends to 3 Å, is again overwhelmingly dominated by the loop fragments. However, this is still much smaller than the center of the distribution of randomly chosen fragments (Figure 4A). If the structural similarity for 15-residue fragments is defined as a Cα r.m.s.d. $\leqslant 2$ Å, then the dot-dashed curve in Figure 5 shows that 91% of the loop fragments in the query database have a similar structure in the template database. While this similarity criterion is not as stringent as for the smaller loops above, fragments with such structural similarities are likely to be useful for some homology modeling applications.

Our results for all three sizes were verified by repeating the above procedure using a smaller query database consisting of

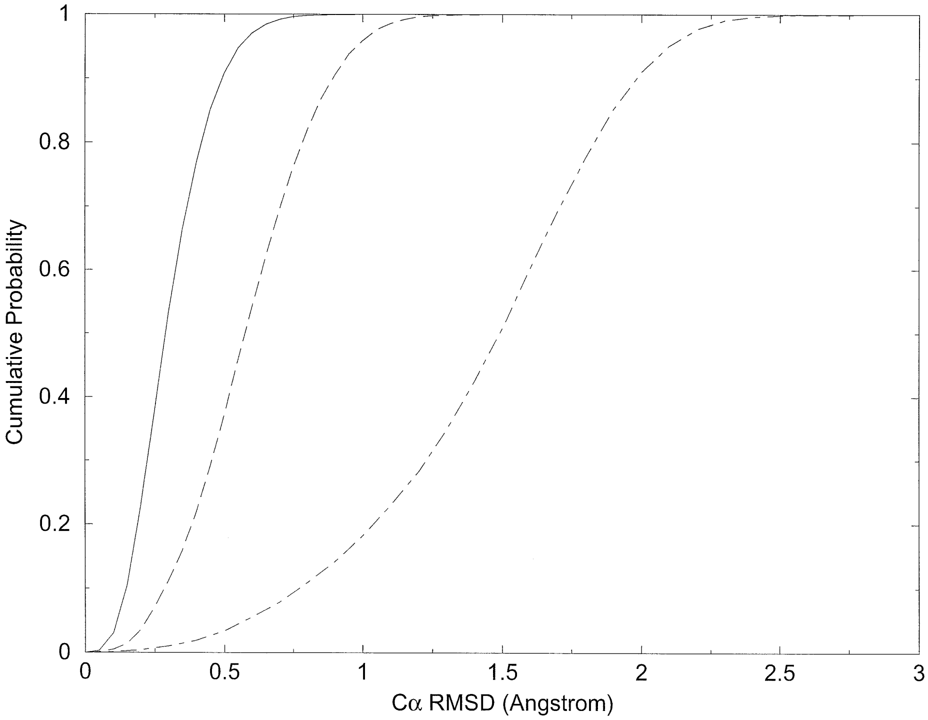

Fig. 5. The cumulative probability of the nearest neighbor Cα r.m.s.d.s in the template database for loop fragments in the query database. The height at an r.m.s.d. value of *x* is equal to the percentage of the query fragments whose nearest neighbor Cα r.m.s.d. is ≤*x*. The solid, dashed and dot-dashed curves represent seven-, nine- and 15-residue fragments, respectively.

loop fragment from a set of four domains from CASP4 which were classified as novel folds (1FW9, 1EWQ, 1FU1 and 1JAD) (Sippl *et al.*, 2001) (the 'CASP4 new fold set') (Figure 6). The resulting distributions of nearest neighbor r.m.s.d.s shown as the solid, dashed and dot-dashed curves in Figure 6 are comparable to the long dashed curves in Figures 2, 3 and 4, respectively. Despite the fact that the CASP4 new fold set is much smaller than the query database, each curve in Figure 6 has an overall shape similar to its counterpart generated from the query database. For example, the distribution of nearest neighbor r.m.s.d.s for seven-residue loop fragments in the CASP4 new fold set is also centered at 0.2–0.3 Å and no fragment has a nearest neighbor r.m.s.d. >0.8 Å in the template database. Results with the CASP4 new fold set are consistent with the results generated from the template and query database and confirm that our results are applicable even for a protein with an entirely new fold, since none of the domains in the CASP4 new fold set have a similar fold in the version of SCOP from which the template database was constructed.

Discussion

The results described above are considerably more promising than those reported by Fidelis *et al.* in 1994 (Fidelis *et al.*, 1994) and are due to the much larger database of protein structures currently available. The center of the distribution of nearest neighbor r.m.s.d.s for a seven-residue loop fragment is improved significantly from ~0.6 to 0.3 Å. The fact that 99.3% of the time one can find a seven-residue fragment in the template database that is within 0.7 Å of a given fragment in the query database implies that a database approach is applicable for the construction of complete protein folds from short fragments, when combined with sparse experimental data, such as given by Andrec *et al.* (Andrec *et al.*, 2001, 2002). Consider the case where the threshold of similarity is realistically set to 0.7 Å and a protein of unknown structure is 200 residues long. The probability that all 194 overlapping seven-residue fragments in the target protein have a similar structure in the template database is 0.256 (0.993^{194}), while the probability of encountering one 'rare' fragment for which no similar structure exists in the database is 0.350 ($194\times0.993^{193}\times0.007$). The latter probability decreases rapidly for more than one rare fragment. Since the fragments can be overlapping, the presence of a small number of rare fragments is not fatal, since there may be sufficient structural information in the neighboring fragments to allow for the construction of the protein structure.

As expected, α fragments are structurally very similar to each other for all three sizes. It is surprising, however, that for seven-residue fragments, the average nearest neighbor r.m.s.d. for β fragments is 0.2–0.25 Å and is broadly distributed. In fact, the distribution of nearest neighbor r.m.s.d. for β fragments is actually close to that for loop fragments (Figure 2B). Although it is known that β strands are structurally diverse owing to twisting and other distortions (Chothia, 1973), the observed variability is somewhat greater than might have been expected.

Enforcing the condition that the query and template database fragments belong to domains having different SCOP folds might be too strict to use as a criterion to remove structural homology, since proteins in the same fold class but different superfamilies are not homologous to each other. The fact that structurally similar fragments can be found even under this very strong condition implies that the results are applicable even when the prediction target is of an entirely new fold and is confirmed by the results with the CASP4 new fold set (Figure 6).

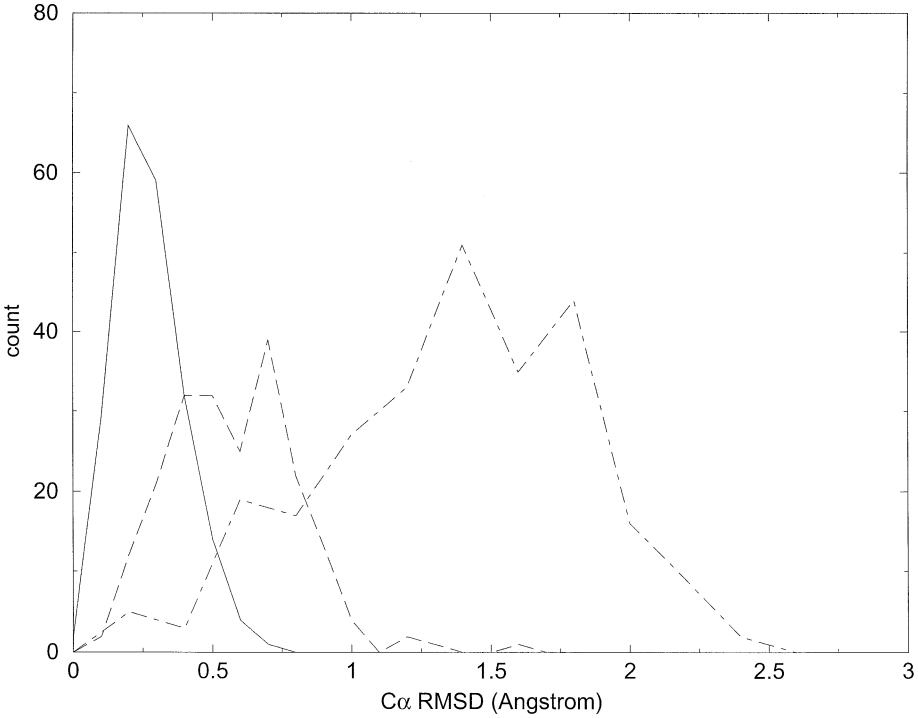

Fig. 6. Histogram of the distribution of nearest neighbor Cα r.m.s.d.s in the template database for all the loop fragments in the CASP4 new fold set for lengths seven, nine and 15. The solid, dashed and dot-dashed curves are the distributions for seven-, nine- and 15-residue fragments, respectively. The bin width is 0.1 for lengths seven and nine, and 0.2 for length 15. The height for each bin is the number of fragments in the CASP4 new fold set database whose nearest neighbor Cα r.m.s.d. in the template database fall in that bin. The target number, PDB code and residues in the CASP4 new fold set are as follows: T0086, 1FW9, A:1–164; T0116, 1EWQ, A:250–542; T0120, 1FU1, A:1–116; and T0124, 1JAD, A:3–244. The number of loop fragments in the CASP4 new fold set for lengths seven, nine and 15 is 207, 204 and 261, respectively.

For longer fragments, it becomes more difficult to find a similar fragment in the template database for each fragment in the query database. However, how similar a fragment must be to the native structure for it to be useful clearly depends on the application. For example, the existence of fragments in the database which are within ~0.7 Å r.m.s.d. over seven residues is necessary for building models of protein structure using an NMR residual dipolar coupling based approach (Andrec *et al.*, 2001). As the fragment size becomes longer, the structure becomes more fold-specific, i.e. fragments from a different fold are less likely to share similar structure. For example, an average 15-residue fragment typically has a nearest neighbor r.m.s.d. about 1.5 Å (e.g. the peak of the distribution in Figure 4B or the dot-dashed curve in Figure 5). Such fragments are not structurally similar enough to fit NMR dipolar coupling data, though they will be useful in building homology models. When compared with the distribution of the r.m.s.d. of pairs of randomly chosen fragments (Figure 4A), which has a mean of 5.5 Å and a standard deviation of 1.6 Å, the 1.5 Å r.m.s.d. between nearest neighbor non-homologous 15 residue fragments is 2.8 standard deviations from the mean and is therefore highly statistically significant in the sense of structural similarity.

Knowing that the right structural fragment is in the database is only the first step: one must also be able to pick it out. For the loop modeling problem, one can search for fragments in the PDB whose residues adjacent to the loop can be superimposed to those of the target (e.g. Greer, 1980; Summers and Karplus, 1990; van Vlijman and Karplus, 1997; Wojcik *et al.*, 1999; Deane and Blundell 2000). This approach has already been shown to be effective for loops up to nine residues long (van Vlijman and Karplus, 1997), but becomes less reliable for loops of longer size. A second way to pick out the correct fragments is to use information in the amino acid sequence and composition, which are used in the loop modeling programs by Kwasigroch *et al.* (Kwasigroch *et al.*, 1996) and Wojcik *et al.* (Wojcik *et al.*, 1999). Short fragments whose structures correlate strongly with sequence profiles can also be predicted using sequence to structure clustering according to the method of Bystroff and Baker (Bystroff and Baker, 1998).

An increasingly important way to pick out the correct fragments from the database is the use of sparse experimental data [e.g. (Jones and Thirup, 1986; Cornilescu *et al.*, 1999; Delaglio *et al.*, 2000; Andrec *et al.*, 2001, 2002)], which very effectively reduces the number of feasible fragments. An example of such a strategy based on NMR residual dipolar couplings is shown in Figure 1. Residual dipolar coupling data can be highly sensitive to small changes in conformation, resulting in a relatively high rate of false negatives, that is, fragments that are similar in structure to the target but which do not score sufficiently well to pass the filter in step A of Figure 1 (Andrec *et al.*, 2001, 2002). For such a strategy to be sucessful, it is critical that there be a sufficient number of structures close to the target structure. The growth of the structural database described above has been essential for the feasiblity of model building based on fragment libraries using NMR data.

The database-oriented approach to protein structure ‘construction’ is complementary to the grid search approach, which systematically searches the conformational space. The former is efficient and the fragments found are guaranteed to be

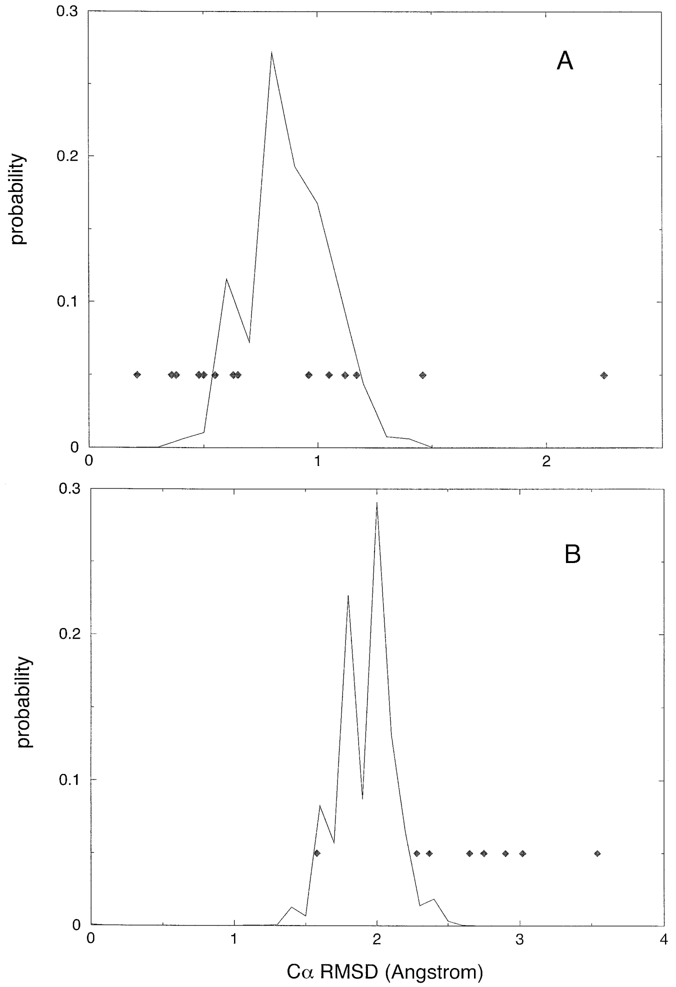

Fig. 7. Histograms of the distributions of nearest neighbor r.m.s.d.s in the template database to loops generated using the PLOP software of M.P.Jacobson and R.A.Friesner (personal communication). For each PLOP-generated loop, the nearest neighbor r.m.s.d. in the template database was found (as described above) and the resulting histograms were summed after normalizing by the number of loop conformations. The diamonds indicate the r.m.s.d. to the native loop conformation for the best PLOP-generated conformation for each of the 14 target loops. (**A**) Results for 14 nine-residue loops. The loops used as targets were 1BUE:A 86–94 (1297), 1BUE:A 156–164 (10 000), 1BUE:A 158–166 (10 000), 1BUE:A 160–168 (10 000), 1BUE:A 213–221 (4203), 1BUE:A 267–275 (2978), 2ACT 141–149 (2612), 2ACT 198–206 (6765), 2APR 76–84 (4509), 2APR 202–210 (3164), 2PTN 69–77 (1547), 2PTN 71–79 (1547), 3APP 129–137 (1158) and 8GCH 95–103 (3049) [PDB accession code:chain, residue range (number of loop conformations generated)]. (**B**) The corresponding results for nine 15-residue loops. The targets were 1PLC 41–56 (119), 2ACT 89–103 (155), 2ACT 141–155 (3095), 3APP 42–56 (370), 8TLN:E 55–69 (164), 8TLN:E 221–235 (1176), 1BGC 56–71 (10 477), 1BUE:A 155–169 (436) and 1OIS 217–231 (7008).

physically reasonable, but is limited by the completeness of the database for longer fragments. The latter is not limited by completeness, but rather by the exponential increase of the conformational space that must be searched. Based on results from this study, the database is essentially complete for seven- and nine-residue fragments. Furthermore, the efficiency of database methods can be greatly increased by making use of clustering methods to reduce the structural redundancy of the database (Lessel and Schomburg, 1997; Kolodny *et al*., 2002). In particular, the recent work of Kolodny *et al.* has demonstrated that the database size can be reduced to under 500 fragments for fragment lengths of four to seven residues and still result in adequate modeling accuracy. In addition, fragments selected by the database-based approach can be successfully used as initial conformation for optimization (van Vlijman and Karplus, 1997; Simons *et al*., 1999). The usefulness of database search for longer fragments will depend on the 'radius of convergence' necessary for the particular homology modeling application and the ability to 'anneal' database loop fragments onto template frameworks.

In order to compare the database search results with a grid search method, we performed both types of searches on a set of 23 loops, nine and 15 residues in length. The conformational search was performed using an early implementation of the PLOP (Protein Local Optimization Program) loop homology modeling software of M.P.Jacobson and R.A.Friesner (personal communication), which has been partially described in other publications (Jacobson *et al*., 2002a,b). Loop conformations are generated in PLOP by sampling the backbone dihedral angles using a discretized version of the Ramachandran plot for the N- and C-terminal halves of the loop independently and then applying a loop closure algorithm in the middle of the loop. The primary mechanism for screening loop conformations is identification of steric clashes. However, other criteria are also employed to eliminate rapidly unlikely conformations, including screens to ensure that the side chains on the loop can fit properly. Because the accessible backbone conformational space of a loop can vary widely, the number of conformations in the backbone dihedral angle library is not set in advance. Rather, the PLOP algorithm commences with coarse sampling and gradually samples more finely until a prescribed number of loop conformations have been generated.

One question of interest is whether nature makes use of all sterically satisfactory loops of a given size. For example, one could imagine that not every sterically feasible loop has a close neighbor in the PDB. If that were the case, then the use of database search would have a distinct advantage over systematic search, since one would avoid those sterically feasible loops that nature (for whatever reasons) does not use. To determine if this is the case, we used PLOP to generate sterically feasible loop models for each of the target loops as described in the caption of Figure 7. For all of these feasible loop models, we found the nearest neighbor in the template database as described above and generated overall histograms of the resulting distributions of r.m.s.d.s, which are shown in Figure 7. For both nine- and 15-residue loops, these distributions are qualitatively very similar to the distributions in Figures 3 and 4, particularly in the location and thickness of the upper tail. This indicates that, for the most part, nature does in fact use all sterically feasible loops, since loops systematically generated by grid search have nearest neighbors in the database with the same distribution as loops from actual proteins.

Since we have not yet developed a complete database-oriented loop modeling method, we cannot directly compare the efficiencies of the two approaches. However, in order to compare the distributions in Figures 3 and 4 with the PLOP-generated loops, we did examine the distributions of the best loops generated by PLOP relative to the native conformation. These are shown as the diamonds in Figure 7. For the grid spacings used in these PLOP runs, the distribution for nine-residue loops is approximately similar to the distribution of nearest database neighbors of both Figures 7A and 3 (except for one outlier at 2.25 Å), whereas for 15 residues the database search appears to find nearest neighbors with significantly smaller r.m.s.d.s to the native than are generated by the grid search (compare Figures 4B and 7B). It should be noted that the latter result is not surprising, since PLOP uses a coarser grid for 15-residue loops because of the exponential increase in the search space. Since PLOP uses a systematic search, we expect the distribution of the best r.m.s.d. relative to native to go to zero as the fineness of the grid (and the CPU time used) increases. For the runs shown here, PLOP required several hours to 1 day of CPU time per loop, while calculation of the r.m.s.d. of a given loop against every fragment in the database can be done in ~15 s. Of course, these times are not comparable, since no filtering of the database loops for clashes with the rest of the protein are being performed and no scoring or ranking is being performed. However, these results suggest that a database search approach for loop homology modeling may have an efficiency advantage relative to conformational grid search for long loops and the problem merits further comparative study.

Overall, we have shown that there exist structurally very similar fragments in the PDB from a non-homologous protein for short loops; for seven-residue fragments there is a 99% probability that there exists a non-homologous structure within 0.7 Å, whereas for nine-residue fragments there is a 96% probability that there exists a non-homologous structure within 1.0 Å. For longer loops (15 residues) we observe a >90% probability that there exists a non-homologous structure within 2 Å r.m.s.d.. Compared with systematic search in conformational space, the use of a database of known structures has the potential advantage of more efficient sampling and guarantees that all backbone conformations are physically reasonable. Results from this study are far more optimistic than those from previous studies of a similar nature and should encourage the use of fragment databases for protein structure determination, prediction and loop modeling, either alone or in combination with the conformational search approach to building protein structures.

Acknowledgements

We thank Matthew Jacobson and Richard Friesner for providing us with an early version of their homology modeling program and for helpful discussions. This work was supported in part by grants from the National Institutes of Health (GM 30580 and GM 06899).

References

Andrec,M., Du,P. and Levy,R.M. (2001) *J. Biomol. NMR*, **21**, 335–347.

Andrec,M., Harano Y., Jacobson,M.P., Friesner,R.A. and Levy,R.M. (2002) *J. Struct. Funct. Genomics*, **2**, 103–111.

Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) *Nucleic Acids Res.*, **28**, 235–242.

Brenner,S.E., Koehl,P. and Levitt,M. (2000) *Nucleic Acids Res.* **28**, 254–256.

Bystroff,C. and Baker,D. (1998) *J. Mol. Biol.*, **281**, 565–577.

Chandonia,J.M., Walker,N.S., Lo Conte,L., Koehl,P., Levitt,M. and Brenner,S.E. (2002) *Nucleic Acids Res.*, **30**, 260–263.

Chothia,C. (1973) *J. Mol. Biol.*, **75**, 295–302.

Cornilescu,G., Delaglio,F. and Bax,A., (1999) *J. Biomol. NMR*, **13**, 289–302.

Correa,P.E. (1990) *Proteins*, **7**, 366–377.

Deane,C.M. and Blundell,T.L. (2000) *Proteins*, **40**, 135–144.

Delaglio,F., Kontaxis,G. and Bax,A., (2000) *J. Am. Chem. Soc.*, **122**, 2142–2143.

Fidelis,K., Stern,P.S., Bacon,D. and Moult,J. (1994) *Protein Eng.*, **7**, 953–960.

Fiser,A., Feig,M., Brooks,C.L.,III and Sali,A. (2002) *Acc. Chem. Res.*, **35**, 413–421.

Greer,J. (1980) *Proc. Natl Acad. Sci. USA*, **77**, 3393–3397.

Holm,L. and Sander,C. (1991) *J. Mol. Biol.*, **218**, 183–194.

Jacobson,M.P., Friesner,R.A., Xiang,Z. and Honig,B., (2002a) *J. Mol. Biol.*, **320**, 597–608.

Jacobson,M.P., Kaminski,G.A., Friesner,R.A. and Rapp,C.S., (2002b) *J. Phys. Chem. B*, **106**, 11673–11680.

Jones,T.A. and Thirup,S. (1986) *EMBO J.*, **5**, 819–822.

Kabsch,W. (1976) *Acta Crystallogr.*, **A32**, 922–923.

Kabsch,W. and Sander,C. (1983) *Biopolymers*, **22**, 2577–2637.

Kolodny,R., Koehl,P., Guibas,L. and Levitt,M. (2002) *J. Mol. Biol.*, **323**, 297–307.

Kwasigroch,J., Chromilier,J. and Mornon J. (1996) *J. Mol. Biol.*, **259**, 855–872.

Lessel,U. and Schomburg,D. (1997) *Protein Eng.*, **10**, 659–664.

Levitt,M. (1992) *J. Mol. Biol.*, **226**, 507–533.

McLachlan,A.D. (1979) *J. Mol. Biol.*, **128**, 49–79.

Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995). *J. Mol. Biol.*, **247**, 536–540.

Reid,L.S. and Thornton,J.M. (1989) *Proteins*, **5**, 170–182.

Simons,K.T., Bonneau,R., Ruczinski,I. and Baker,D. (1999) *Proteins*, Suppl. **3**, 171–176.

Sippl,M.J., Lackner,P., Domingues,F.S., Prlic,A., Malik,R., Andreeva,A. and Wiederstein,M. (2001) *Proteins*, Suppl. **5**, 55–67.

Summers,N.L. and Karplus,M. (1990) *J. Mol. Biol.*, **216**, 991–1016.

van Vlijmen,H.W.T. and Karplus,M. (1997) *J. Mol. Biol.*, **267**, 975–1001.

Wojcik,J., Mornon,J. and Chomilier,J. (1999) *J. Mol. Biol.*, **289**, 1469–1490.

Received October 23, 2002; revised March 3, 2003; accepted March 30, 2003