doi:10.1016/j.jmb.2003.10.027 J. Mol. Biol. (2003) 334, 793–802

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

The PDB is a Covering Set of Small Protein Structures

Daisuke Kihara and Jeffrey Skolnick*

Center of Excellence in Bioinformatics, University at Buffalo, 901 Washington St. Suite 300, Buffalo, NY 14203 USA

Structure comparisons of all representative proteins have been done. Employing the relative root mean square deviation (RMSD) from native enables the assessment of the statistical significance of structure alignments of different lengths in terms of a *Z*-score. Two conclusions emerge: first, proteins with their native fold can be distinguished by their *Z*-score. Second and somewhat surprising, all small proteins up to 100 residues in length have significant structure alignments to other proteins in a different secondary structure and fold class; i.e. 24.0% of them have 60% coverage by a template protein with a RMSD below 3.5 Å and 6.0% have 70% coverage. If the restriction that we align proteins only having different secondary structure types is removed, then in a representative benchmark set of proteins of 200 residues or smaller, 93% can be aligned to a single template structure (with average sequence identity of 9.8%), with a RMSD less than 4 Å, and 79% average coverage. In this sense, the current Protein Data Bank (PDB) is almost a covering set of small protein structures. The length of the aligned region (relative to the whole protein length) does not differ among the top hit proteins, indicating that protein structure space is highly dense. For larger proteins, non-related proteins can cover a significant portion of the structure. Moreover, these top hit proteins are aligned to different parts of the target protein, so that almost the entire molecule can be covered when combined. The number of proteins required to cover a target protein is very small, e.g. the top ten hit proteins can give 90% coverage below a RMSD of 3.5 Å for proteins up to 320 residues long. These results give a new view of the nature of protein structure space, and its implications for protein structure prediction are discussed.

Keywords: PDB; protein structure comparison; protein structure space; relative RMSD; fragments

*Corresponding author

Introduction

The expansion of the number of solved protein structures[1] over the last decade has opened up a new horizon in protein research. Through the classification of protein structures,[2–4] we have obtained a clearer perspective of the universe of known protein folds. Based on the argument that the number of protein folds in nature is limited,[5–7] threading-based approaches to protein structure prediction[8–11] have been developed. The increasing library of solved protein structures has greatly benefited the entire field of protein structure prediction. This is true not only for comparative modeling methods,[12] which use a homologous protein structure as a template, but for various methods which use parameters extracted from the structure database, the Protein Data Bank (PDB)[1] as well. These include secondary structure prediction methods using a neural network[13,14] or hidden Markov models,[15,16] and knowledge-based potentials which are frequently used in threading, fold refinement, selection of protein models[17,18] and *ab initio* protein structure prediction.

The most successful *ab initio* methods incorporate information extracted from the PDB. For example, ROSETTA[19] assembles protein structures using structure fragments excised from known

Present address: D. Kihara, Biological Sciences/Computer Sciences, Purdue University, Lilly Hall, 915 West State St., West Lafayette, IN 47907, USA.

Abbreviations used: PDB, Protein Data Bank; RMSD, relative root mean square deviation; DP, dynamic programming; SAL, structure alignment; CE, combinatorial extension.

E-mail address of the corresponding author: skolnick@buffalo.edu

proteins. TOUCHSTONE[20,21] spans the region from homology modeling to *ab initio* folding, and employs inter-residue contact predictions derived from at least weakly hit threading template proteins together with protein-specific short-range[22] and long-range[23] knowledge-based potentials.

The success of protein structure prediction relies on the method to identify the similarity of protein structures. If two proteins are of the same length, then their root mean square deviation (RMSD)[24] is a good similarity measure. When a protein pair starts to structurally diverge and is of different length, another strategy is needed. One class of structural alignment algorithms employs dynamic programming (DP).[25–27] The advantages of DP are that it is easy to take local features into account in the scoring function and the requisite computational time is rather short. The drawback is that global optimality is not guaranteed. DALI[28] focuses on the contacts in proteins by comparing a portion of their distance maps. Grindley *et al*.[29] and Mizuguchi & Go[30] compare spatial arrangements of secondary structure elements represented as vectors. Nussinov *et al*.[31] employ geometric hashing, while an incremental combinatorial extension (CE) method, which combines structurally similar fragments, was employed by Shindyalov & Bourne.[32] Different methods capture different aspects of protein structure and differ in how they search for optimal structure alignments.[33]

Here, using a new structure alignment method, SAL, on a representative database of protein structures, we found that almost all small proteins can be structurally aligned to other proteins in a different fold class, with an alignment that covers almost the entire molecule with a low RMSD from native. For larger proteins, a significant portion can be covered by non-related proteins. Interestingly, different proteins cover different parts of the target protein, so that the whole protein is covered if all are combined.

Results

Database of protein structures

Target proteins

To carry out our analysis of protein structure space, we prepared two protein structure sets: the target protein structures are a set of protein chains each having a known fold. The template proteins are sets of protein chains to which each target protein is compared. All protein structures are extracted from the PDB.

The target protein set, containing 411 chains, represents all non-homologous single-domain protein chains of different topologies which fall into classes 1–4 in the CATH database (version 2.4).[4] The first three digits in the CATH hierarchy define the topology. All share less than 35% pairwise sequence identity. Domains are also assigned by CATH. We discarded proteins classified into classes 5–9 of CATH, because these are preliminary.

Template proteins

For each of the 411 target proteins, six template protein sets of different structure similarity level are prepared against which each target protein will be structurally aligned. The structure similarity level is based on CATH, namely, secondary structure class, architecture, and topology are defined by the first one, two three digit(s) in the CATH hierarchy, respectively. The architecture means the spatial arrangement of the secondary structures in a protein.

For a target protein, template proteins having different secondary structure classes are the "−.−.−.−" set. Those proteins having the same first digit (i.e. secondary structure class), but not the same CATH second digit (i.e. architecture) as the target protein constitute the same secondary structure class template ("C.−.−.−") set. Similarly, proteins sharing the first and second digits with the target protein but not the third digit constitute the same architecture template ("C.A.−.−") set. Proteins sharing the same first to third digits but not the fourth digit constitute the same topology ("C.A.T.−") set. Proteins in these template sets have less than 35% sequence identity with each other and the target proteins.

We divided the homologous superfamily level of the CATH hierarchy into two levels, the close and distant homolog template set, labeled the "C.A.T.Hc" and "C.A.T.Hd" sets, respectively. The C.A.T.Hc and the C.A.T.Hd set consist of proteins with 35% or more sequence identity and those closer than an *E*-value of 0.01 as evaluated by FASTA[34] to the target protein, respectively. These proteins share less than 95% sequence identity to the target protein and to each other. All six template sets are exclusive.

Due to the elimination of too similar proteins (by sequence) and also by the sparseness of the PDB, there are no proteins in certain template sets for some of the target proteins. C.A.THc, C.A.T.Hd, C.A.T.−, C.A.−.−, C.−.−.−, and −.−.−.− sets were constructed for 329 (17.8), 204 (5.7), 231 (9.4), 400 (258.0), 360 (704.6), 411 (1602.7) target (average number of template) proteins, respectively.

Identification of the native fold

The first question is does SAL reproduce standard fold assignments? For all 411 target proteins, SAL was run against the six template protein sets. Figure 1 shows an example of the distribution of the relative RMSD *Z*-scores (Figure 1(a)) and the conventional RMSD (Figure 1(b)) obtained for a target protein, 1tlk (an immunoglobulin fold). Proteins with different fold similarity are shown separately (see legend). Proteins homologous to the target have the lowest *Z*-scores (a good *Z*-score has a negative value), followed by same

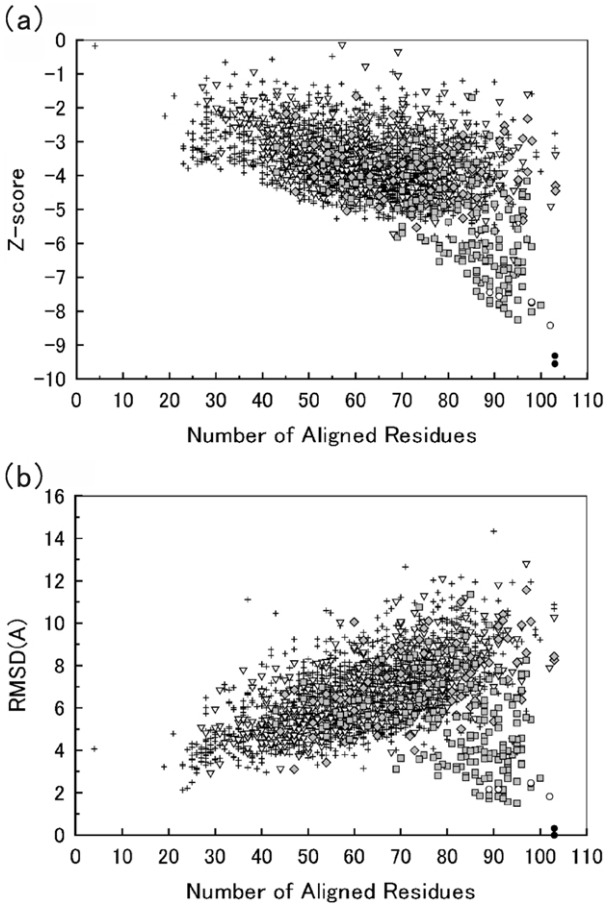


Figure 1. For 1tlk, the distribution of the relative RMSD *Z*-score and the RMSD. The *X*-axis is the number of the aligned residues. (a) The distribution of the relative RMSD *Z*-score; (b) the RMSD distribution. Black circles, template proteins in C.A.T.Hc set; open circles, C.A.T.Hd set; gray squares, C.A.T.– set; gray diamonds, C.A.–.– set; triangles, C.–.–.– set; crosses, –.–.–.– set.

topology proteins. Comparing Figure 1(a) and (b), the *Z*-score of the relative RMSD reflects structure similarity better, making it possible to detect protein structures of the same fold.

The distribution of the relative RMSD *Z*-score of the six template sets of all 411 representative target proteins is shown in Figure 2. The average *Z*-score of each template set is shown in Table 1. The C.A.T.Hc, C.A.T.Hd and C.A.T.– sets have a distinctively lower *Z*-score than the rest, although there are also overlaps. The average alignment length with a *Z*-score > -5 obtained by comparison with the C.A.T.Hc, C.A.T.Hd and C.A.T.– sets

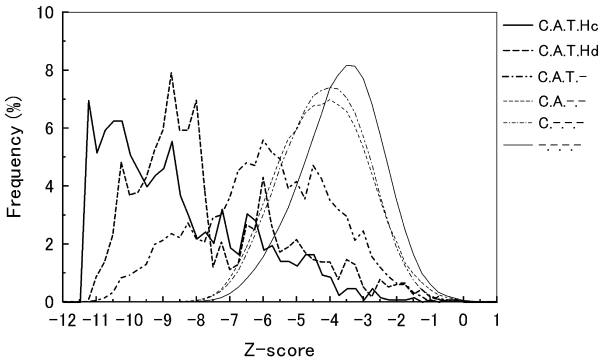


Figure 2. Histogram of the relative RMSD *Z*-score. Continuous line, target proteins are compared to template proteins in the C.A.T.Hc set; broken line, C.A.T.Hd set; dot-dash line, C.A.T.– set; thin broken line, C.A.–.– set; thin dot–dash line, C.–.–.– set; thin continuous line, –.–.–.– set. Since the number of pairs in these sets differs significantly, the frequency (*y* axis) lies within each set. The bin width is 0.25.

is 70.3, 85.0, and 108.3 residues, respectively. In contrast, 149.9, 142.6, and 195.2 residues are the average alignment length with *Z*-scores less than -5 for the same three sets, respectively.

Note that the *Z*-score of the C.A.–.– sets and C.–.–.– sets are not differentiated.

Since the relative RMSD *Z*-scores from proteins having the same fold as the target protein are generally lower than those having a different fold, we can identify the native fold by its *Z*-score alone. For brevity, we term the union of C.A.T.Hc, C.A.T.Hd and C.A.T.– sets as the "correct fold" set and the rest as the "different fold" set. Some proteins in the correct fold set have a high *Z*-score (close to 0) in Figure 2 and considerably different global architecture. Therefore, describing the entire C.A.T.Hc, C.A.T.Hd and C.A.T.– sets as the correct fold is not entirely appropriate. But these exceptional proteins do not greatly affect the analysis, because they are buried in the bulk of the *Z*-score distribution.

Figure 3 shows the cumulative fraction of proteins in each template set detected below a given *Z*-score and also (in the continuous line with filled circles) the fraction of correct fold proteins detected relative to the total number of proteins detected at or below the specified *Z*-score averaged over all 411 proteins ("accuracy" of correct fold detection). For all 411 targets, only correct fold proteins are detected below a *Z*-score of -9.5 (continuous line with filled circles). With a *Z*-score of -7, on average 79% are correct fold proteins. Almost no

Table 1. Average and standard deviation of the relative RMSD *Z*-score for each template set

Template set	C.A.T.Hc	C.A.T.Hd	C.A.T.–	C.A.–.–	C.–.–.–	–.–.–.–
Average	−8.44	−7.47	−5.78	−4.06	−4.03	−3.53
Standard deviation	2.15	2.20	1.99	1.33	1.25	1.23

The distribution of the *Z*-score is shown in Figure 2.

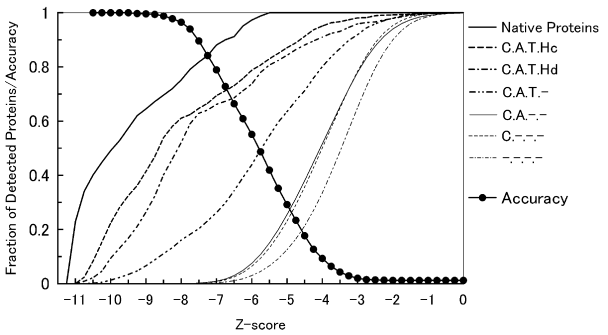


Figure 3. Identification of the correct fold on the basis of the relative RMSD *Z*-score. The ratio of the target proteins (among the 411 proteins), for which a given fraction of correct fold proteins (C.A.T.Hc set + C.A.T.Hd set + C.A.T.− set) are detected at or below a given *Z*-score threshold: in the continuous line with filled circles, the fraction of correct fold proteins detected relative to the total number of proteins detected at or below the specified *Z*-score averaged over all 411 proteins. Continuous line, the fraction of native (target) proteins found below a given *Z*-score threshold; broken line, the fraction of the detected proteins in C.A.T.Hc sets under a given *Z*-score; dot-dash line, the fraction for C.A.T.Hd sets; dot-dot-dash line, the fraction for C.A.T.− sets; thin continuous line, the fraction for C.A.−.− sets; thin broken line, the fraction in C.−.−.− sets; thin dot-dash line, the fraction in −.−.−.− sets.

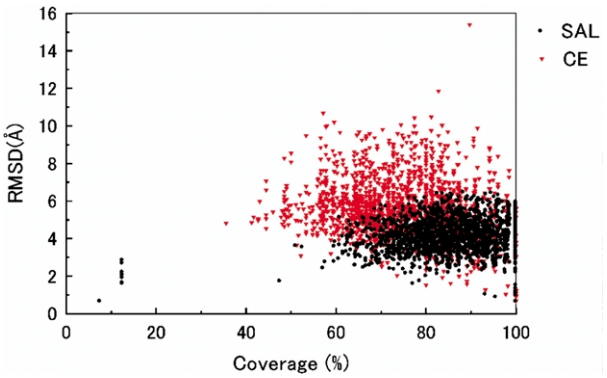


Figure 4. The coverage and RMSD of the top ten hit proteins of the different fold type (i.e. in the −.−.−.− set) for small target proteins up to 100 residues. The best hit proteins whose RMSD is larger than 6.5 Å are excluded. Black circles, SAL; red triangles, CE algorithm by Shindyalov & Bourne is used for the same protein pairs for comparison.

proteins with different folds are found with a *Z*-score of −8 or less (0.06%, 0.03%, 0.01% of the template proteins in the C.A.−.−, C.−.−.−, and −.−.−.− sets, respectively). The fraction of detected correct folds rapidly drops as the *Z*-score threshold is raised due to the *Z*-score gap between the correct and different folds that lies around −7. Thus, SAL as other structural alignment tools, can reproduce the conventional fold classification scheme.

Structural similarity to template proteins in −.−.−.− set

Returning to Figure 1 as an example, one remarkable feature is that there are some template proteins in −.−.−.− set aligned with a low RMSD to 1tlk over a significantly large region. For example, 1bmuA (214 residues) is aligned with a RMSD of 5.8 Å over 97 residues (94.2% of 1tlk).

More generally, for the small target proteins below 100 residues the coverage and RMSD of the top ten template proteins in the −.−.−.− set selected by their *Z*-score are shown in Figure 4. For comparison, structure alignments of the same protein pairs are calculated using CE.[36] The plot clearly shows that alignments by SAL tend to have a larger coverage with a smaller RMSD compared to those generated by the CE method. An extreme case is the alignment of 1aonO (97 residues long, β roll) to 1di1A (290 residues long, orthogonal α bundle), where CE gives a RMSD of 15.4 Å with 89.7% coverage, while SAL gives a RMSD of 3.7 Å and 60.8% coverage, inserting 22 gaps. Figure 4 supports the idea that the PDB is a covering set of protein structures at low resolution.

Examples of aligned structures to proteins in the −.−.−.− set are shown in Figure 5(a)–(d). Figure 5(a) shows the structure alignment between the Rossman fold, a three-layer αβα fold (1aoxA), and the immunoglobulin fold (1tlk), which is a two-layer β-sandwich fold. Despite differences in overall topology and secondary structure elements, both are well aligned with a RMSD of 5.3 Å for 85.4% of the structure of 1tlk. Figure 5(b) shows that a small protein 1aoo (40 residues) has a similar topology to the bigger α-helical protein 1csgA (120 residues) in the region from residues 46 to 115. Figure 5(c) shows that the spatial arrangement and the connectivity of the β strands in 1g8jB is similar to a part of 1dqyA, although the aligned part in 1dqyA contains several α helices. Figure 5(d) shows that the fold of 1bnkA fits to the C-terminal part (273–500 residues) of the bigger protein 1eeeA.

In Figure 4, for 100 residue proteins, we addressed the distribution of the top scoring templates in the −.−.−.− set, their coverage and RMSD from native. Now, we examine this question more generally. In Figure 6(a), the coverage (with a 3.5 Å RMSD threshold) for target proteins of a given length by the top 1, 3, 5, 10, 30 or 50 hit template proteins in the −.−.−.− set is shown. The average sequence identity between the pairs is 11.1% (standard deviation: 8.3%). The coverage by the best covering template protein does not differ much from that of the average, if the top 3, 5, 10, 30 or 50 proteins are used. Thus, there are not only a few but literally dozens of template proteins that have the similar coverage to a target protein; i.e. protein structure space is strikingly dense.

Figure 6(a) and (b) also shows that almost the entire target protein can be covered if all top hit template proteins are combined at a given RMSD threshold. If the top ten hit proteins are combined,

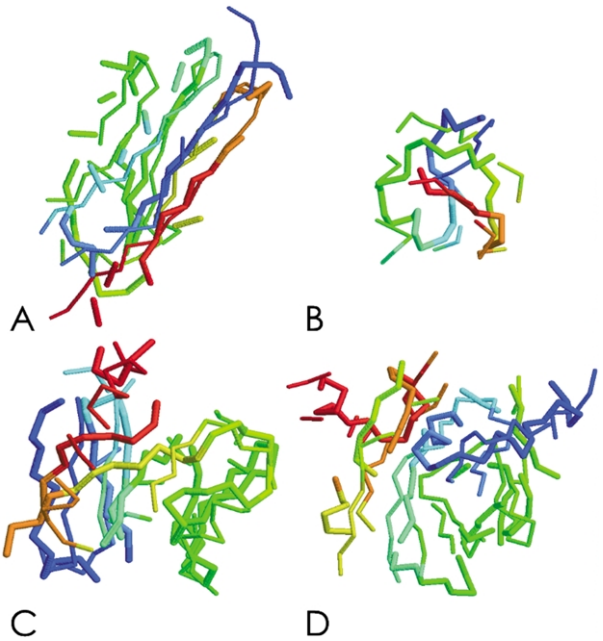


Figure 5. Examples of structure alignments of protein pairs between proteins in the −.−.−.− set. The CATH classification code is shown in the parentheses together with the topology name. The first digit of the CATH code shows the fold class type: 1, α type; 2, β; 3, α,β; 4, proteins with little secondary structure. (A) 1tlk (103 residues; immunoglobulin-like: 2.60.40, shown in a thick tube) and 1aoxA (201 residues; Rossmann fold: 3.40.50 shown in a thin tube). The aligned region is 88 residues long, i.e. 85.4% coverage for 1tlk, with a RMSD of 5.3 Å and a *Z*-score = −5.57. (B) 1aoo (40 res.; Ag-metallothionein: 4.10.650, thick tube) and 1csgA (120 residues; four helix bundle: 1.20.120). The aligned region is 40 residues long with a RMSD of 4.6 Å and a *Z*-score of −2.53. (C) 1g8jB (128 res.; Rieske iron-sulfur protein: 2.102.10, thick tube) and 1dqyA (283 res.; Rossmann fold: 3.40.50). The aligned region is 98 residues long with a RMSD of 6.1 Å and a *Z*-score of −5.24. (D) 1bnkA (200 residues; 3-methyladenine DNA glycosylase: 3.10.300, thick tube) and 1eeeA (582 residues; methanol dehydrogenase chain A: 2.140.10). The aligned region is 112 residues long with a RMSD of 6.1 Å and a *Z*-score of −6.08. The structures are colored in blue to red from the N to C terminus.

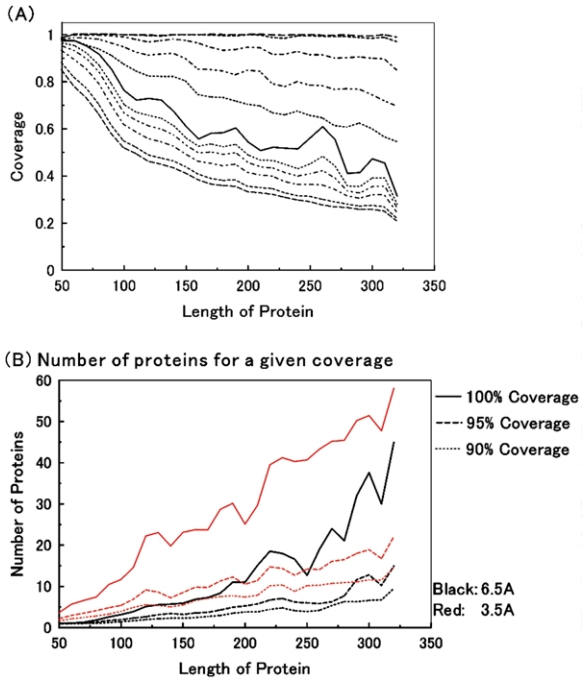


Figure 6. The coverage by top hit template proteins in −.−.−.− set by the relative RMSD *Z*-score (a) which have a RMSD from native less than 3.5 Å. Two types of data are plotted, one is the average coverage and the other is the combined coverage by all top hit proteins. Continuous line, the coverage by the top protein; thin dot, thin dot-dot-dash, thin dot-dash, thin dash line, the average coverage by the top 3, 5, 10, 50 hit proteins, respectively. Dot, dot-dot-dash, dot-dash, dash line, the combined coverage from the top 3, 5, 10, 50 proteins, respectively. B, The combined number of proteins required covering a given fraction of the target protein with a given RMSD threshold. The proteins used to cover a target protein are taken in the order of their covering length; therefore their combination is not necessarily optimal (i.e. the combination which uses the minimum number of proteins). Continuous line, the number of proteins required for the 100% coverage; broken line, 95% coverage; dotted line, 90% coverage. Two different RMSD thresholds are used: black, 6.5 Å; red, 3.5 Å.

on average, more than 85% of the entire molecule is covered, regardless of length and the RMSD threshold. Figure 6(b) shows the number of template proteins required to cover a specified fraction of a target protein as a function of target protein length. When 3.5 Å is the RMSD threshold, 24 target proteins do not achieve 100% coverage.

Until this juncture, we have been considering the hardest of all possible cases: structural alignments between pairs of proteins of entirely different secondary structure type, that is the −.−.−.− set. This insured that the evolutionary relationship between such protein pairs, if any, is very distant. A more practical question is to examine whether the PDB is a covering set of all pairs of proteins whose sequence identity is in the twilight zone. Thus, we considered a representative set of all single domain proteins (no two of which have more than 35% pairwise sequence identity to each other) ranging from 41 to 200 residues in length; there are 1491 such proteins†. These are compared against a benchmark PDB template library comprised of 3575 templates‡. All templates with greater than 22% sequence identity to the target are excluded. For each target structure, the template with highest relative RMSD *Z*-score is selected. A total of 1470 targets can be matched with a RMSD <5 Å. The average RMSD is 2.6 Å with 9.8% sequence identity and 78% alignment coverage. Of this set,

† http://bioinformatics.buffalo.edu/threading/LIST.benchmark

‡ http://bioinformatics.buffalo.edu/threading/LIST.templates

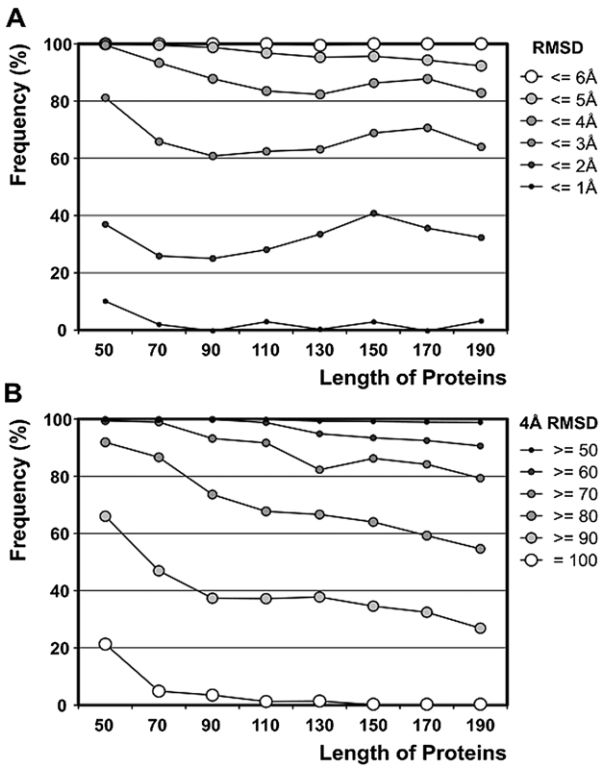


Figure 7. Plotted along the *X* axis is the target protein's length (in 20 residue intervals). For the representative set of 1491 target proteins: (a) the fraction of proteins at a given length that have a RMSD less than or equal to the specified threshold; (b) where for a RMSD 4 Å, the fraction of proteins at a given length having different extents of coverage specified in the legend.

1336/1491 (93%) have an RMSD <4 Å with 79% coverage.

These results are expanded on in Figure 7(a) and (b), where we show the fraction of proteins at a given length that have an RMSD less than or equal to the specified threshold. Even for 200-residue target proteins, greater than 90% have a structural alignment 5 Å, and more than 60% have a structural alignment 3 Å. Now, it could be argued that the coverage is small at this RMSD threshold, so the results are not significant. That this is not the case is shown in Figure 7(b), where for a RMSD threshold 4 Å, respectively, we plot the fraction of proteins at a given length having different extents of coverage. More than 80% of the proteins have greater than 70% coverage. Thus, most proteins can be covered over a significant fraction of their length by a single template structure whose average sequence identity is about 10%.

Discussion

Here, we discussed the nature of protein structure space based on structural comparisons that use a dynamic programming-based method. Two major conclusions are evident: the first is that the native fold can be detected with a distinct Z-score value. In *ab initio* protein structure prediction, some authors have used structural similarity to known proteins as a measure of prediction confidence for a structure model.[19,37] Using SAL, it is also possible to give a strong indication that a predicted model is correct if it has a sufficiently low relative RMSD *Z*-score. On the other hand, for any kind of protein structure, SAL finds dozens of known structures in the *Z*-score range around -3 to -5 which cover a significant part of the target with a reasonable RMSD; this leads to the most important conclusion.

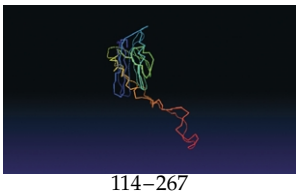
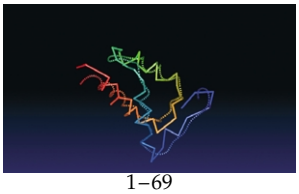
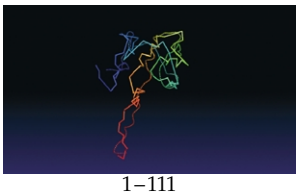
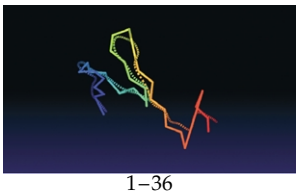
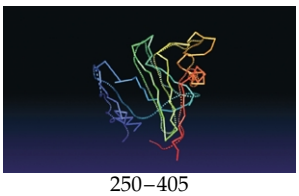
Protein structure space provided by the current PDB is already very dense; i.e. for an arbitrary target protein, there are many proteins that have a significant large covering structure to the target protein if gaps are allowed in the alignment. In addition to the significant coverage by a single protein, combining several proteins will cover the entire structure of the target protein. The RMSD threshold used to assess the coverage does not affect this general trend.

Several authors have addressed the nature of protein structure space by comparing all (or representative) structures in the PDB.[4,38] These studies have emphasized the discreteness of space on the domain level of protein structures. On the other hand, recently Shindyalov & Bourne[39] pointed out that substructures obtained from an all-against-all structure comparison using their CE method sometimes distribute among protein domains transgressing their respective fold types. Their substructures are around 130 residue long continuous chains, longer than the conventional concept of supersecondary structure.[40] Harrison *et al*. also revisited the protein structure similarities using a graph-based algorithm, and concluded that fold space is a continuum for some topology types in the β or α,β secondary structure class.[41] What both found was that there are local structure motifs that occur in many other folds. Thus, some regions of protein fold space are not as distinct as was thought. Yang & Honig[42] also showed that their structure comparison program detects structure similarity between different folds in the SCOP database.[3]

Compared to their results, our conclusion is more far reaching: we claim that the PDB is already a dense covering set of known protein structures. Why is our result so different? The reason is because SAL is permissive in generating structure alignment in the sense that it allows for gaps in alignments, and fragments of different secondary structures can also be aligned if necessary (see for example Figure 5, the alignment of 1aonO to 1di1A). And last, but not least, the statistical significance of alignments with gaps are properly estimated by the relative RMSD *Z*-score, making it possible to select template structures with significant coverage to a target protein. Therefore, SAL detects the occupation of space by contiguous (not necessarily continuous, i.e. gaps are permitted) geometric objects. In contrast, the graph

theory-based program by Harrison *et al*. treats continuous secondary structure fragments as nodes in a graph representation of protein structure, while CE combines corresponding fragments to obtain continuous alignments. Needless to say, SAL still does not lose the ability to detect very close structure pairs with significant scores (see Figure 1), the minimum requirement of any acceptable

Table 2. Covering structures for the CASP5 new fold protein domains

ID	PDB code[a]	CASP category[b]	Length	Coverage -by best template[c]	Structural superposition of the target /best template pair
T162_3	1IZN, 114–281	NF	168	55% 4.88 Å 1dt9A	 114–267
T170	1H40	FR(A)/NF	69	83% 2.81 Å 1ku1A	 1–69
T172_2	1M6Y, 116–216	FR(A)/NF	101	71.2% 3.31 Å 1dq8A	 116–216
T181	1NYN	NF	90 (111)[d]	93.3% 5.16 Å 1al6_	 1–111
T186_3	1O12, 257–292	FR(A)/NF	36	94.4% 2.16 Å 1yprA	 1–36
T187_1	1O0U, (4-22) +250–417	FR(A)/NF	168	56% 3.61 Å 1dmuA	 250–405

Superposition of the target–template pair. The target (template) protein is shown in springs (solid) tube. Blue to red runs from the N to C terminus. The numbers below the structural superposition are the residue numbers at the beginning and end of the structural alignment.

[a] The region of a domain is specified with a PDB code. T187_1 consists of two separate parts of 1O0U and the N-terminal helix (residues 4–22) is omitted from the calculation.

[b] The fold category of the targets was determined by the CASP organizers. NF, new fold; FR(A), fold recognition target which has analogous structure in PDB.

[c] Coverage (%), RMSD (Å), and PDB entry which gives the coverage to a given target protein is shown; template protein is selected on the basis that it has the best relative RMSD *Z*-score.

[d] This structure is determined by NMR, which has 20 models in the PDB file. Only the N-terminal consistent region among the models (1–90 residues) is used in the calculation.

structure comparison program. The point is that we focused our attention on the structure similarity of non-related structures and demonstrated that such template structures can cover a target protein with gaps; this is of great importance in the context of protein structure prediction.

By way of further illustration, SAL also detects structures for CASP5 protein domains (Table 2) classified as (at least partly) "New Folds" by the CASP5 assessors. Each of the new fold CASP5 targets can have a significant fraction of their chain covered by a single (unrelated) protein structure. The superposition of the best target-template pair is shown in Table 2, column 6. Even when the alignment only covers about 55% (T162_3) of the residues, it stretches from almost the N to C terminus of the target protein and thus is buildable. It is in this sense that the PDB is a covering set.

It is very important to understand the universe of protein folds. Moreover, this can have a strong impact on protein structure prediction, especially of the *ab initio* type. Recent *ab initio* structure prediction methods have been improved by using information taken from the PDB. Our current observation leads to the possibility of developing a new, more aggressive way to extract structure information from databases, which is beyond contact prediction[20,21] or the use of continuous fragments.[19] Covering template structures for a target protein can be used as a scaffold in structure prediction. From Figures 4 and 5, it is evident that all proteins below 100 residues have template structures (of different secondary structure type) sufficient for predicting the topology of the target proteins.

When the more realistic (but still very difficult situation) is considered (see Figure 7). Targets with less than 22% sequence identity to their closest template (on average for the representative benchmark considered here, 9.8% sequence identity), then 88% of all representative PDB structures have a structural alignment with an RMSD below 4 Å, an average RMSD of 2.4 Å and 86% average coverage. This is certainly of the requisite quality to build a low-to-moderate resolution structure and in work to be published elsewhere we demonstrate that this can in fact be done. Based on the current rapid growth of the structure and sequence databases, there is no doubt that this is the fastest way to develop more powerful structure prediction approaches.

Materials and Methods

Structure comparison algorithm

In the protein structure alignment algorithm, SAL, DP was iteratively applied. The main differences between our and existing DP-based structure alignment programs are the use of the relative RMSD *Z*-score to assess the statistical significance of structure alignments, and the gradual lowering of the gap penalty in the DP calculation to identify more distant structural similarities. The *Z*-score of the relative RMSD is explained in the Appendix.

The program starts with an initial guess for the superimposition of two protein structures constructed as follows: from the longer protein (of length m), a continuous fragment of the same length as the shorter protein (of length n) is chosen and their RMSD calculated. The rotation matrix $\boldsymbol{Q}$ and the translation vector $\boldsymbol{T}$ that give the smallest RMSD among the $m - n + 1$ different fragments are stored. This provides the initial guess for the best $\boldsymbol{Q}$ and $\boldsymbol{T}$, which are then applied to the entire longer protein.

Next, the score assigned to residue i in protein α and residue j in protein β is calculated from:[25]

$$\mathrm{Score}(\alpha, i; \beta, j) = A/\{1 + ((\overrightarrow{r_{\alpha,i}} - Q(\overrightarrow{r_{\beta,j}} - T))/B)^2\} \quad (1)$$

where A and B are parameters; $(A, B) = (20.0, 2.24)$ gives the best results. Using this scoring matrix, DP is performed to give a structure alignment. The relative RMSD *Z*-score of the aligned residues is calculated. If it is better (smaller) than previously, the $\boldsymbol{Q}$ and $\boldsymbol{T}$ from the RMSD calculation are again applied, and the score table for DP updated. Since the algorithm employs DP, the selected structure alignment can have gaps. The gap penalty increases from -10 to 0. For a given gap penalty, this iteration was performed at most 200 times or until the structure alignment converged. Finally, for a given pair of proteins, the structure alignment with the best relative RMSD *Z*-score was selected.

Acknowledgements

This research was supported in part by NIH Grant GM-48835 of the Division of General Medical Sciences and by the Oishei Foundation.

References

1. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H. *et al.* (2000). The Protein Data Bank. *Nucl. Acids Res.* **28**, 235–242.
2. Holm, L. & Sander, C. (1998). Touring protein fold space with Dali/FSSP. *Nucl. Acids Res.* **26**, 316–319.
3. Lo Conte, L., Brenner, S. E., Hubbard, T. J., Chothia, C. & Murzin, A. G. (2002). SCOP database in 2002: refinements accommodate structural genomics. *Nucl. Acids Res.* **30**, 264–267.
4. Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. M. (1997). CATH—a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.
5. Chothia, C. (1992). Proteins. One thousand families for the molecular biologist. *Nature*, **357**, 543–544.
6. Orengo, C. A., Jones, D. T. & Thornton, J. M. (1994). Protein superfamilies and domain superfolds. *Nature*, **372**, 631–634.
7. Wang, Z. X. (1996). How many fold types of protein are there in nature? *Proteins: Struct. Funct. Genet.* **26**, 186–191.
8. Skolnick, J. & Kihara, D. (2001). Defrosting the frozen approximation: PROSPECTOR—a new approach to threading. *Proteins: Struct. Funct. Genet.* **42**, 319–331.
9. Bowie, J. U., Luthy, R. & Eisenberg, D. (1991). A method to identify protein sequences that fold into

a known three-dimensional structure. *Science*, **253**, 164–170.
10. Jones, D. T. (1999). GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.* **287**, 797–815.
11. Madej, T., Gibrat, J. F. & Bryant, S. H. (1995). Threading a database of protein cores. *Proteins: Struct. Funct. Genet.* **23**, 356–369.
12. Sali, A. & Blundell, T. L. (1993). Comparative protein modeling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779–815.
13. Rost, B. & Sander, C. (1993). Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proc. Natl Acad. Sci. USA*, **90**, 7558–7562.
14. Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292**, 195–202.
15. Karplus, K., Karchin, R., Barrett, C., Tu, S., Cline, M., Diekhans, M. *et al.* (2001). What is the value added by human intervention in protein structure prediction? *Proteins: Struct. Funct. Genet.*, 86–91.
16. Asai, K., Hayamizu, S. & Handa, K. (1993). Prediction of protein secondary structure by the hidden Markov model. *Comput. Appl. Biosci.* **9**, 141–146.
17. Lu, H. & Skolnick, J. (2001). A distance-dependent atomic knowledge-based potential for improved protein structure selection. *Proteins: Struct. Funct. Genet.* **44**, 223–232.
18. DeBolt, S. E. & Skolnick, J. (1996). Evaluation of atomic level mean force potentials *via* inverse folding and inverse refinement of protein structures: atomic burial position and pairwise non-bonded interactions. *Protein Eng.* **9**, 637–655.
19. Bonneau, R., Strauss, C., Rohl, C., Chivian, D., Bradley, P., Malmstrom, L. *et al.* (2002). *De novo* prediction of three-dimensional structures for major protein families. *J. Mol. Biol.* **322**, 65.
20. Kihara, D., Lu, H., Kolinski, A. & Skolnick, J. (2001). TOUCHSTONE: an *ab initio* protein structure prediction method that uses threading-based tertiary restraints. *Proc. Natl Acad. Sci. USA*, **98**, 10125–10130.
21. Kihara, D., Zhang, Y., Lu, H., Kolinski, A. & Skolnick, J. (2002). *Ab initio* protein structure prediction on a genomic scale: application to the *Mycoplasma genitalium* genome. *Proc. Natl Acad. Sci. USA*, **99**, 5993–5998.
22. Kolinski, A., Jaroszewski, L., Rotkiewicz, P. & Skolnick, J. (1998). An efficient Monte Carlo model of protein chains. Modeling the short-range correlations between side group center of mass. *J. Phys. Chem.* **102**, 4628–4637.
23. Skolnick, J., Kolinski, A. & Ortiz, A. (2000). Derivation of protein-specific pair potentials based on weak sequence fragment similarity. *Proteins: Struct. Funct. Genet.* **38**, 3–16.
24. Kabsch, W. (1978). A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallog.* A**34**, 827–828.
25. Gerstein, M. & Levitt, M. (1996). Using iterative dynamic programming to obtain accurate pairwise and multiple alignments of protein structures. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **4**, 59–67.
26. May, A. C. (1996). Pairwise iterative superposition of distantly related proteins and assessment of the significance of 3-D structural similarity. *Protein Eng.* **9**, 1093–1101.
27. Taylor, W. R. & Orengo, C. A. (1989). Protein structure alignment. *J. Mol. Biol.* **208**, 1–22.
28. Holm, L. & Sander, C. (1993). Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233**, 123–138.
29. Grindley, H. M., Artymiuk, P. J., Rice, D. W. & Willett, P. (1993). Identification of tertiary structure resemblance in proteins using a maximal common subgraph isomorphism algorithm. *J. Mol. Biol.* **229**, 707–721.
30. Mizuguchi, K. & Go, N. (1995). Comparison of spatial arrangements of secondary structural elements in proteins. *Protein Eng.* **8**, 353–362.
31. Bachar, O., Fischer, D., Nussinov, R. & Wolfson, H. (1993). A computer vision based technique for 3-D sequence-independent structural comparison of proteins. *Protein Eng.* **6**, 279–288.
32. Shindyalov, I. N. & Bourne, P. E. (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.* **11**, 739–747.
33. Godzik, A. (1996). The structural alignment between two proteins: is there a unique answer? *Protein Sci.* **5**, 1325–1338.
34. Pearson, W. R. & Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
36. Shindyalov, I. N. & Bourne, P. E. (2000). An alternative view of protein fold space. *Proteins: Struct. Funct. Genet.* **38**, 247–260.
37. de la Cruz, X., Sillitoe, I. & Orengo, C. (2002). Use of structure comparison methods for the refinement of protein structure predictions. I. Identifying the structural family of a protein from low-resolution models. *Proteins: Struct. Funct. Genet.* **46**, 72–84.
38. Holm, L. & Sander, C. (1996). Mapping the protein universe. *Science*, **273**, 595–603.
39. Shindyalov, I. N. & Bourne, P. E. (2000). An alternative view of protein fold space. *Proteins: Struct. Funct. Genet.* **38**, 247–260.
40. Boutonnet, N. S., Kajava, A. V. & Rooman, M. J. (1998). Structural classification of alphabetabeta and betabetaalpha supersecondary structure units in proteins. *Proteins: Struct. Funct. Genet.* **30**, 193–212.
41. Harrison, A., Pearl, F., Mott, R., Thornton, J. & Orengo, C. (2002). Quantifying the similarities within fold space. *J. Mol. Biol.* **323**, 909–926.
42. Yang, A. S. & Honig, B. (2000). An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance. *J. Mol. Biol.* **301**, 665–678.

Appendix

Z-score of the relative RMSD

The conventional RMSD can be calculated only for two, same length proteins. Here it is necessary to compare different length structure alignments; we introduce the relative RMSD *Z*-score as the similarity measure.[A1] (The smaller the number of residues, the lower the RMSD must be for it to be significantly better than random.)

The relative RMSD is defined as the ratio between the conventional RMSD and that of the average of two randomly related proteins, $\overline{RMSD}$

of the same length:

$$(\text{Relative RMSD}) \equiv \text{RMSD}/\overline{RMSD} \quad \text{(A1)}$$

The statistics for $\overline{RMSD}$ the are taken from almost 1300 non-homologous continuous chains.[A1,A2]

The *Z*-score is calculated from the average and the standard deviation of the relative RMSD, and for an alignment of length N, is given approximately by:

$$\langle \text{Relative RMSD} \rangle = 1.0 \quad \text{(for any protein length)} \quad \text{(A2)}$$

$$\text{Std(Relative RMSD)(N)} \approx 0.09 + 1.16\, e^{-(N-1)/1.6} + 0.25\, e^{-(N-1)/36} \quad \text{(A3)}$$

The *Z*-score is defined is defined using equations (A1)–(A3) as:

$$Z\text{-score} = (\text{relative RMSD} - \langle \text{Relative RMSD} \rangle)/\text{Std(Relative RMSD)} \quad \text{(A4)}$$

Recalculation of the relative RMSD for structural alignments with gaps

The ensemble of randomly selected pairs of proteins used to obtain the relative RMSD *Z*-score ((A1)–(A3)), were continuous chains. On the other hand, most structural alignments in the current work have gaps. Here, we re-examined the relative RMSD and its *Z*-score using gapped alignments. The relative RMSD and its *Z*-score for 922 selected structure alignments from 41 target proteins were recalculated using a reference ensemble 3339 alignments having the same number of aligned residues and the same arrangements of gaps (i.e. the same alignment pattern). The recalculated relative RMSD and the original relative RMSD have a correlation coefficient of 0.95. The *Z*-score of the recalculated relative RMSD also correlates well with the original *Z*-score, with a correlation coefficient of 0.91. Since the original relative RMSD and its *Z*-score correlate very well with the recalculated ones, we used the original relative RMSD and its *Z*-score in the analysis.

Cut-off for structural similarity

We use a RMSD threshold of 6.5, 5.0 and 3.5 for structure similarity. Although Skolnick *et al.* addressed the issue of the statistical significance of RMSD years ago,[A3] here, we examine it in terms of the *Z*-score of the relative RMSD calculated using equation (A4) (which is length dependent). The *Z*-score monotonically decreases (i.e. becomes more significant) as N increases. For example, the *Z*-score of RMSD of 6.5 Å is −5.3 when N is 100. For shorter proteins, the *Z*-score is −2.42 when N is 50, which is the minimum length of target proteins discussed (e.g. Figures 4 and 7).

References

A1. Betancourt, M. R. & Skolnick, J. (2001). Universal similarity measure for comparing protein structures. *Biopolymers*, **59**, 305–309.

A2. Betancourt, M. R. & Skolnick, J. (2001). Finding the needle in a haystack: educing native folds from ambiguous *ab initio* protein structure predictions. *J. Comp. Chem.* **22**, 339–353.

A3. Reva, B. A., Finkelstein, A. V. & Skolnick, J. (1998). What is the probability of a chance prediction of a protein structure with a RMSD of 6Å? *Fold. Des.* **3**, 141–147.

Edited by B. Honig

(Received 15 November 2002; received in revised form 30 September 2003; accepted 9 October 2003)