BIOINFORMATICS ORIGINAL PAPER
Vol. 21 no. 7 2005, pages 1010–1019
doi:10.1093/bioinformatics/bti128

Structural bioinformatics

Non-sequential structure-based alignments reveal topology-independent core packing arrangements in proteins

Xin Yuan and Christopher Bystroff*

Department of Biology, Rensselaer Polytechnic Institute, Troy, NY 12180, USA

Received on June 18, 2004; revised on September 27, 2004; accepted on October 22, 2004
Advance Access publication November 5, 2004

ABSTRACT

Motivation: Proteins of the same class often share a secondary structure packing arrangement but differ in how the secondary structure units are ordered in the sequence. We find that proteins that share a common core also share local sequence–structure similarities, and these can be exploited to align structures with different topologies. In this study, segments from a library of local sequence–structure alignments were assembled hierarchically, enforcing the compactness and conserved inter-residue contacts but not sequential ordering. Previous structure-based alignment methods often ignore sequence similarity, local structural equivalence and compactness.

Results: The new program, SCALI (Structural Core ALIgnment), can efficiently find conserved packing arrangements, even if they are non-sequentially ordered in space. SCALI alignments conserve remote sequence similarity and contain fewer alignment errors. Clustering of our pairwise non-sequential alignments shows that recurrent packing arrangements exist in topologically different structures. For example, the three-layer sandwich domain architecture may be divided into four structural subclasses based on internal packing arrangements. These subclasses represent an intermediate level of structure classification, more general than topology, but more specific than architecture as defined in CATH. A strategy is presented for developing a set of predictive hidden Markov models based on multiple SCALI alignments.

Availability: An online topology-independent SCALI structure comparison server is available at http://www.bioinfo.rpi.edu/∼bystrc/scali.html

Contact: bystrc@rpi.edu

INTRODUCTION

Recurrent structural motifs in proteins can be found by structure-based alignment methods. Generally it is assumed that similar protein structures will align to each other in a sequential manner, conserving the direction of the chain and the order of the structural units. However, there are many examples of structural similarity that are non-sequential, produced possibly by sequence rearrangements (Janowski *et al*., 2001; Bennett *et al*., 1994; Schiering *et al*., 2000; Jeltsch, 1999; Gong *et al*., 1997; Iwakura *et al*., 2000; Viguera *et al*., 1995; Smith and Matthews, 2001; Jung and Lee, 2001) or by convergent evolution (Rost, 1997; Milik *et al*., 2003). Circular permutants and other rearrangements represent the topologically possible and energetically favorable ways of arranging secondary structure units along the chain.

Structural similarities that have permuted orders are interesting because they reveal recurrent structural packing themes in proteins (Efimov, 1995; Abagyan and Maiorov, 1989; Alexandrov, 1996). Examples are presented in this paper. These recurrent themes may be used to build predictive models. However, there are no sequence models for the well-known structural paradigms at the level of protein architecture. Instead, the focus has been on predicting structure at the family, superfamily or fold level (Eddy, 1998; Karplus *et al*., 1998; Gough and Chothia, 2002). The most successful of these are hidden Markov models (HMMs), which generally do not allow for non-sequential alignments. The lack of non-sequential HMMs may be due in part to the difficulty in obtaining good structural alignments without sequential constraints. Owing to this, a recurrent structural motif in a new protein may not be recognized, as such, if it is sequentially permuted.

Alignments of topologically different structures may be found by inspection using an interactive graphics program such as Rasmol (Bernstein, 2000; Sayle and Milner-White, 1995). The spatial alignment of permuted segments is often remarkably good, yet most structure-based alignment programs, such as DALI (Holm and Sander, 1993), CE (Shindyalov and Bourne, 1998), VAST (Gibrat *et al*., 1996), PrISM (Yang and Honig, 2000) and MAMMOTH (Ortiz *et al*., 2002), cannot find these superpositions because they assume the aligned segments to be sequentially ordered. Two exceptions to this rule are SARF (Alexandrov, 1996; Alexandrov and Fischer, 1996) and K2 (Szustakowski *et al*., 2002; Szustakowski and Weng, 2000), which consider non-topological alignments using secondary structure element information. However, neither of these programs considers sequence similarity. Our goal was to build sequence models from structure-based alignments; therefore, we developed a new program that optimizes both the structure and sequence similarity.

The new program, SCALI (Structural Core ALIgnment), was conceived based on the following five criteria defining a biologically relevant structure-based alignment (Koehl, 2001; Flores *et al*., 1993; Taylor and Orengo, 1989): (1) aligned residues should conserve structure locally (i.e. backbone angles); (2) contacts between pairs of aligned residues should be conserved; (3) the alignment as a whole should be spatially compact, rather than disperse; (4) aligned segments should have some degree of sequence similarity; and (5) the sequence order of aligned segments should be minimally permuted.

*To whom correspondence should be addressed.

In preliminary studies, we constructed non-sequential alignments manually for two cases of topologically different proteins with similar three-dimensional (3D) core packing arrangements, one of which is illustrated in Figure 1. We then attempted to reproduce the manually constructed alignments, automatically, using a fragment assembly strategy. The new program was compared with the two of the most commonly used structure alignment programs, DALI and CE, and with two non-sequential alignment programs, SARF and K2.

Pairwise SCALI alignments of representative protein structures were clustered to produce multiple structure alignments. Within these clusters we found recurrent core packing arrangements that could be used as models for structure prediction. HMMs based on these 'cores' are presented here in a diagrammatic form. These models represent a level of structural classification that is more general than 'fold' or 'topology' but more specific than 'architecture' or 'class'. Applications of the new non-topological HMMs for structure prediction and design are discussed. Recurrent core packing geometries may also tell us something about the folding process.

METHODS

SCALI: non-sequential sequence–structure alignment

SCALI aligns structures in a three-step process. First we generate a library of gapless local sequence–structure alignments ('fragments') using HMMSTR (HMM for protein STRucture) (Bystroff *et al.*, 2000). The second step is a tree search in alignment space, where each branch point is the addition of a new fragment to the alignment. Finally, the best alignments are pruned and extended.

HMMSTR is an almost comprehensive model for local sequence/structure correlations in proteins (Bystroff *et al.*, 2000). In HMMSTR, each Markov state represents a single position in an I-sites motif (Bystroff and Baker, 1998). Each state contains information about the amino acid preference and the preferred backbone angles. The transitions between the states model the adjacencies of motifs in protein sequences. HMMSTR has been used for secondary structure prediction, remote homolog detection (Hou *et al.*, 2003) and for developing knowledge-based contact potentials (Shao and Bystroff, 2003). The algorithms for using this and other HMMs are described in Rabiner's classic tutorial (Rabiner, 1989).

To align two structures using SCALI, we first computed the position-specific HMMSTR state probabilities, denoted γ, using the Forward/Backward algorithm (Rabiner, 1989). The input to this program was a sequence profile derived from PSI-BLAST (Altschul *et al.*, 1997) as described previously (Bystroff and Baker, 1998).

Next, we made an exhaustive list of all aligned fragments. To obtain this list, we first calculated the alignment matrix A as the dot-product of the state probabilities:

$$A_{ij} = \sum_q \gamma_{iq}^{\text{target}} \gamma_{jq}^{\text{template}}, \tag{1}$$

where q represents a Markov state in the HMMSTR model, and γ_{iq} is the probability of state q at position i. The score $S(i, j, L)$ for a fragment of length L, starting at position i in the target and j in the template is simply the sum over a diagonal segment of the alignment matrix A:

$$S(i, j, L) = \sum_{k=0,L-1} A_{(i+k)(j+k)}, \tag{2}$$

All possible fragments, defined by the positions i, j, and the length L, were compiled as a list, subject to the following constraints. A fragment

(1) must have no gaps or insertions,

(2) must have no backbone angle difference >90°,

(3) must be at least five residues in length and

(4) must not be contained within a longer fragment that has a higher score.

Fragments were sorted by their alignment score, S [Equation (2)]. In every example of two aligned segments that have no backbone angle differences >90°, the two segments are superimposable with a low root-mean-square deviation (RMSD). There is no upper limit on the length of a fragment.

A breadth-first tree search in alignment space was conducted using a contact map scoring function. A contact map, C, is an $N \times N$ matrix where $C_{ij} = 1$ if the β-carbons (Cα for glycine) of residues i and j are separated by <8 Å, and 0 otherwise. The n (where $n = 200$) fragments with the highest scores, S [Equation (2)], were used as seed alignments for the tree search. At each branch point, the parent alignment y was extended using fragment x if and only if:

(1) no residue in x is already aligned;

(2) there is at least one conserved contact between fragment x and a residue in y;

(3) distance geometric constraints are not violated, meaning Distance$(i, j) <$ $3.8 \times |l - k|$, and Distance$(k,l) < 3.8 \times |j - i|$, for all positions i aligned to l, and j aligned to k;

(4) the resulting alignment has one of the top n scores [NS, as defined in Equation (6)].

The top n scoring alignments (parents and children) become the parent alignments of a new search, until no new fragments could be added.

The similarity between two contact maps is more sensitive than the global RMSD when comparing distantly related proteins (Yang and Honig, 1999), since conformational plasticity can result in a high overall RMSD even when most of the pairwise contacts are conserved. The contact score, CS, is the sum over the dot-products of the contact maps for all aligned segments:

$$T = \sum_{ij} (C_{ij} \times C_{kl}), \tag{3}$$

$$F = \sum_{ij} [(1 - C_{ij}) \times C_{kl}] + [C_{ij} \times (1 - C_{kl})], \tag{4}$$

$$\text{CS} = T - \text{NCpenalty} \times F, \tag{5}$$

where, C_{ij} is the contact property at position i, j in the target, C_{kl} is the contact property at position k, l in the template, where i is aligned to l, and j is aligned to k. The F (false positive and false negative contacts) was penalized by non-contact penalty (NCpenalty). The score NS is calculated as follows:

$$\text{NS} = \text{CS} - \text{NSpenalty} \times \text{Nb}, \tag{6}$$

where NSpenalty is a constant for non-sequential penalty, and Nb is defined as the number of breaks needed to convert the alignment into sequential order. For example, if we have the alignment with the blocks labeled as ACBD, where each letter represents a sequentially aligned segment, to reorder the blocks sequentially to ABCD, three breaks are required, therefore Nb = 3. For the arrangement CDAB, a circular permutation, Nb = 1. The optimal settings for NSpenalty and NCpenalty were determined empirically by reproducing manually constructed alignments.

The final step in generating the alignment is pruning and extension, based on global RMSD. Occasionally SCALI aligned fragments that conserved a pattern of contacts, but the two structures were mirror images of each other. The 3D superposition followed by pruning eliminated these types of errors. After the superposition of structures, an aligned block was removed if it:

(1) had any distance difference >9 Å, or

(2) had any backbone angle difference >100°.

Similarly, an aligned block was extended on either end if the extension:

(1) had no backbone angle difference >100°, and

(2) had no distance difference >9 Å.

The larger cutoff in distance allowed for distorted packing arrangements. Pruning and extension were applied iteratively as long as the RMSD and aligned length continued to improve.

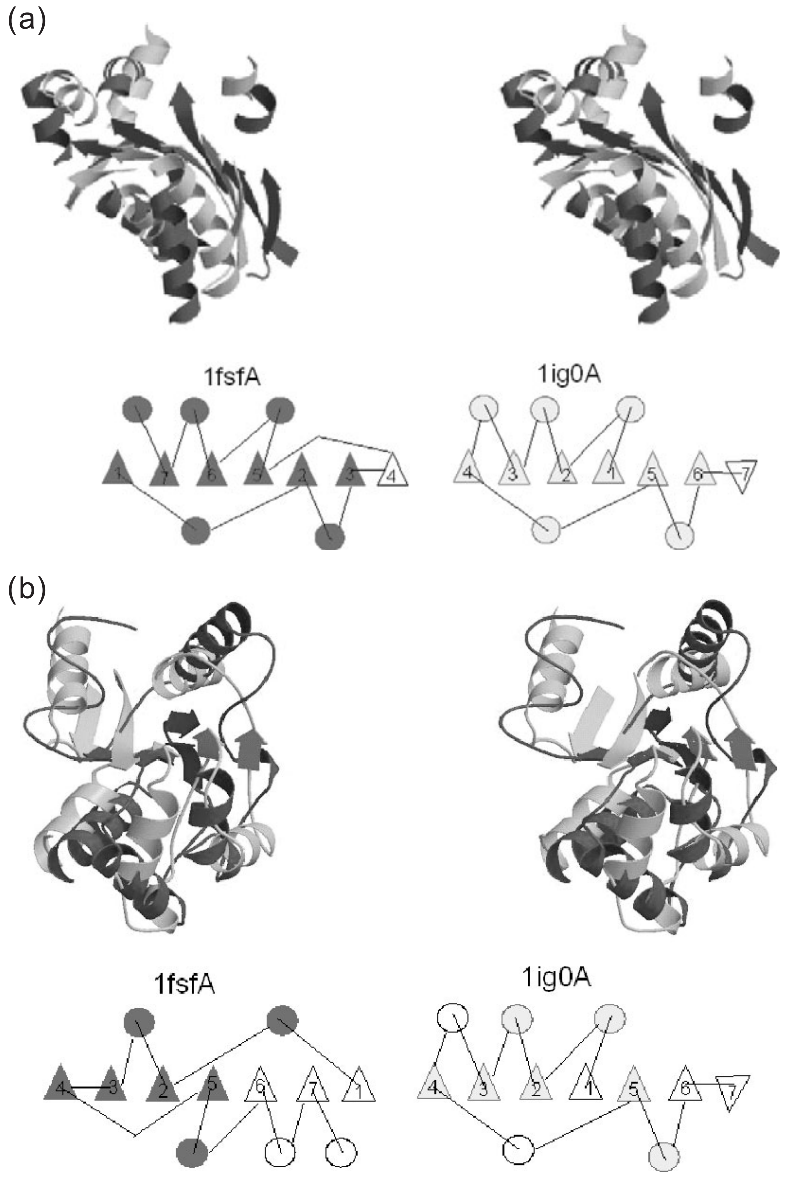

Fig. 1. Structural comparison between *Escherichia coli* Glucosamine-6-phosphate deaminase (PDB code 1fsfA, 266 residues) and yeast thiamin pyrophosphokinase (PDB code 1ig0A, 319 residues) , using four methods. Each figure shows only the aligned residues, dark color for 1fsfA and light color for 1ig0A. Below each figure is a topology (TOPS) cartoon, with strands as triangles, helices as circles. The diagrams are oriented roughly as the proteins are aligned, with the aligned segments shaded. For simplicity, only the secondary structure units that are in common between the two proteins are shown. The alignments may contain additional small aligned fragments that are not included in the TOPS diagrams. (**a**) SCALI alignment, 104 aligned residues, RMSD = 5.4 Å, one permutation. Aligned segments (1fsfA/1ig0A): 1–9/114–122, 10–24/125–139, 33–40/182–189, 43–58/197–212, 63–70/213–220, 133–140/38–45, 190–195/53–58, 200–218/64–82, 237–241/93–97, 247–256/101–110. (**b**) CE alignment, 111 aligned residues, RMSD = 5.1 Å. Aligned segments (1fsfA/1ig0A): 14–28/49–63, 35–43/64–72, 46–64/73–91, 67–74/92–99, 85–86/100–101, 89–102/102–115, 104–110/116–122, 115–116/123–124, 119/125, 120–130/129–139, 131–146/185–200, 150–156/201–207. (**c**) DALI alignment, 106 aligned residues, RMSD = 4.9 Å. Aligned segments (1fsfA/1ig0A): 23–27/32–36, 33–42/37–46, 45–53/48–56, 66–72/63–69, 85–93/71–79, 95–100/85–90, 104–110/92–98, 111–116/119–124, 119–129/128–138, 132–139/186–193, 145–148/194–197, 189–192/202–205, 196–203/217–224, 220–223/225–228, 230–233/303–306, 237–240/308–311. (**d**) SARF alignment, 105 aligned residues. RMSD = 2.9 Å. Aligned segments (1fsfA/1ig0A): 1–7/114–120, 10/125, 12–23/126–137, 34–42/183–191, 62–74/212–224, 91–98/203–210, 132–137/38–43, 188–204/54–70, 210–217/104–111, 221–225/83–79*, 235–246/92–103, 257–263/154–160. An asterisk (*) denotes reversed segments.

Empirical values for the permutation penalty (NSpenalty) and non-contact penalty (NCpenalty) were determined by attempting to reproduce the manual non-sequential alignments for two study cases: 1fsfA versus 1ig0A, and 1jx6A versus 1qo7A. The two manual alignments were made by inspection using the molecular modeling program InsightII (Accelrys, Inc.). Different parameter settings were assessed by inspecting the automated alignments and comparing them with the manual ones. The SCALI result for a difficult sequential alignment between remote homologs (sequence identity of 4.5%), 1rec_ versus 1eg4A, was also inspected. The final settings were NCpenalty = 0.15 and NSpenalty = 6.0.

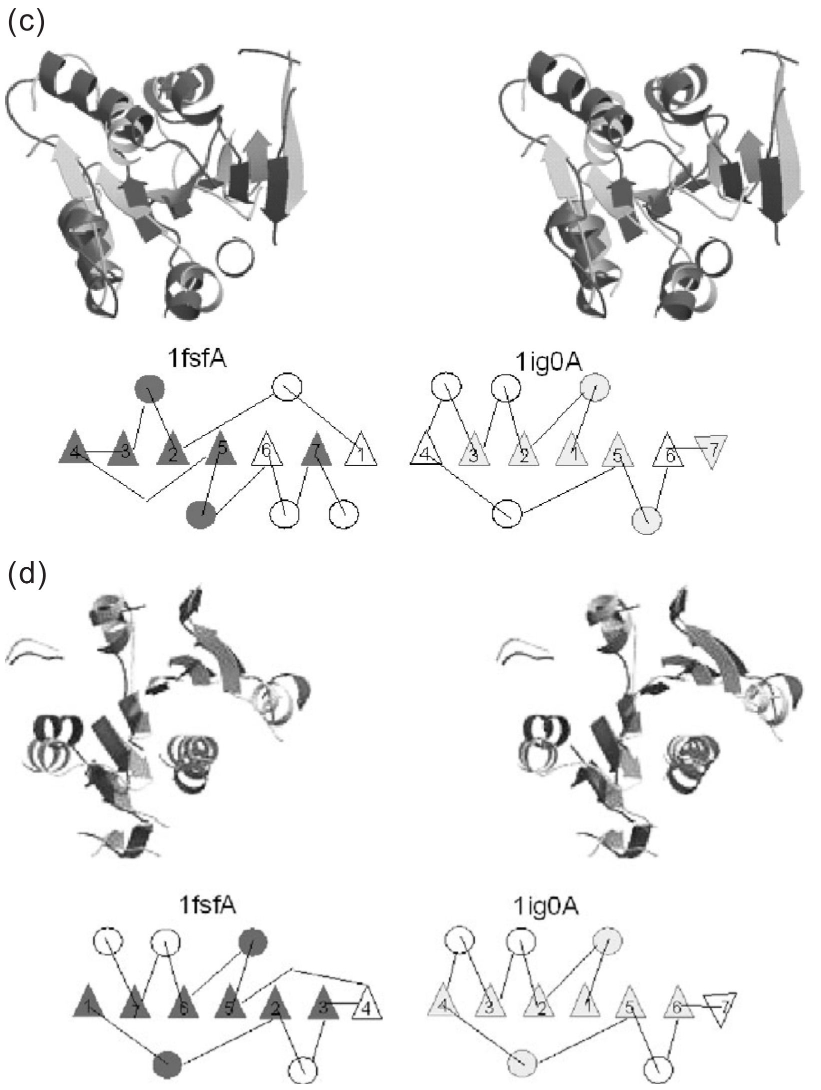

Fig. 1. *Continued*

Many of the automatically generated alignments (available from the website) were inspected in order to validate the method. In particular, we looked for the types of errors as described in the Results section and in Table 1. No further changes were made to the algorithm once the validation was undertaken.

DALI, CE, SARF alignments

CE and SARF programs were downloaded and the alignments were generated locally. DALI alignments were obtained from the server (www.ebi.ac.uk/dali/Interactive.html). For each program, the default settings were used. We were unable to run the K2 program in-house for technical reasons. Programs were written to test each alignment for specific types of errors, as described in the Results section.

Comparison and evaluation of the alignments from different methods

A comparison of various structural alignment methods with SCALI was performed using a reference set of 111 alignments derived from CATH. The alignments of SCALI were compared to those from DALI, CE and SARF, which were generated as described above. The alignment results are summarized in Table 1. While the alignments were judged by visual inspections, an evaluation was also undertaken using a figure-of-merit (*FOM*) scoring function denoted as:

$$\begin{aligned} FOM &= w_1(Len) + w_2(RMSD - 3.5) + w_3(Nonlocal) \\ &\quad + w_4(Disj) + w_5(Berr), \end{aligned} \tag{7}$$

where, *FOM* is computed as the scaled sum of the five criteria which are represented as *Len*, *RMSD*, *Nonlocal*, *Disj* and *Berr* in Equation (7). Each criterion is assigned a weight ($w_1 - w_5$). *Len* is the number of locally correct aligned residues in the alignment, where the positions having backbone angle deviations $>120°$ were ignored. *RMSD* is the root mean square deviation of the aligned Cα coordinates. *Nonlocal* is the percentage of the aligned residues that were locally non-equivalent, having backbone angle deviations $>120°$. *Disj* is the number of disjoint segments in the alignment, and *Berr* is the number of misaligned beta strands (as described in the Results section). The weights were chosen so as to roughly equalize the contribution of each factor: $w_1 = -0.2$, $w_2 = 2.0$, $w_3 = 15$, $w_4 = 4$ and $w_5 = 8$. A small *FOM* is better.

Table 1. Systematic comparison of 111 SCALI alignments with CE, SALI and SARF

Method	Average alignment length	RMSD (Å)	Local non-equivalent (cases)	Strand mis-alignment error (cases)	Disjoint error (cases)	Not aligned (cases)	Error-free (cases)	Average FOM
CE	83.3	5.9	111	28	1	0	0	1.6
DALI	81.5	5.7	76	10	18	35	0	0.1
SARF	79.7	2.7	111	22	41	0	0	−6.8
SCALI	64.7	4.3	10	7	7	0	91	−10.5

In this table, the information of the averaged alignment length, RMSD and FOM is derived for all 111 cases using the method of CE, SARF and SCALI. Only 76 alignments are used for the evaluation of DALI method since 35 out of 111 are not alignable.

The one-tailed paired Wilcoxson's sum of signed ranks test (Sokal and Rohlf, 1973) was calculated to measure the significance of the difference in *FOM* between the four alignment methods (CE, DALI, SARF and SCALI). The differences in *FOM* scores for the two methods being compared were ranked based on their absolute values, and the positive and negative ranks were summed separately. The smaller sum (T_s) was used to calculate a Z-score [Equation (8), $n = 111$].

$$Z = \frac{|Ts - [n(n+1)]/4|}{\sqrt{[n(2n+1)(n+1)]/24}}. \tag{8}$$

The calculated Z was compared to tabulated critical values given by Sokal and Rohlf (1973) to obtain a P-value, or significance level. Lower P-value is more significant. The DALI method aligned only 76 cases, so for comparisons to DALI, $n = 76$ and only those alignments were used.

Clustering multiple SCALI structural alignments

Pairwise SCALI non-sequential alignments were clustered using a simple greedy algorithm. The set of all pairwise alignments defines a graph where each vertex represents one protein, and an edge exists if the alignment between the two structures had RMSD ≤ 4.0 Å and at least 50 residues aligned. The first cluster was the vertex with the most edges and all of its connected vertices. The second cluster was the vertex with the most edges and all of the connected vertices after removing the first cluster, and so on.

Theoretically, this simple clustering method could group together different structures by transitive association (structure A is similar to B, and B is similar to C, but A is not similar to C). Surprisingly this did not happen. Instead, alignments within a cluster conserved the same spatial location. We should note again here that the structures in the CATH database are single domains, and therefore we did not expect to find multiple cores within one structure. Also, domain folding is usually an all-or-none, two-state phenomenon, and this may explain why we did not observe transitive association.

Hidden Markov models based on SCALI multiple structure alignments

An HMM state was defined for each column in the multiple structure alignment after clustering of our alignments. The state sequence profiles were initialized by summing the profiles of the aligned positions. The aligned proteins were representatives of the CATH topologies. Each protein was given equal weight in the summation. Any non-aligned residues were condensed to a single 'Loop' state that connects the aligned states. The loop states emit sequences whose length was drawn from a probability distribution. The probability distribution may be flat, allowing any size loop with equal probability.

State–state transitions were defined according to the sequential ordering of the states in each member protein. In many cases, since the alignments were non-sequential, cyclic state paths were possible. These paths are not physically meaningful, since they would imply that two residues can occupy the same position in space. Therefore, 'self-avoiding' states were defined as Markov states that could be visited at most once in any state pathway. The development of a modified Forward-Backward algorithm (Rabiner, 1989) that handles self-avoiding states is ongoing and will produce the correct results on our newly defined HMMs.

In the figures, the Markov states for aligned positions are grouped into single icons representing secondary structures, according to the TOPS convention (Westhead *et al.*, 1999). Loop states are not drawn, but would occur on each of the arrows.

Information content of HMM states

The information content is defined as the likelihood of obtaining a similar distribution of polar and non-polar residues, by chance, given the number of observations. To estimate this likelihood, we ran 5000 simulations for each Markov state. We randomly chose amino acids from the background distribution N times, where N was the number of observations. The P-value for non-polar residues was calculated as the fraction of the 5000 randomly generated profiles where the percentage of non-polar residues matched or exceeded that of the observation. If the Markov state represented a polar position, then a P-value for polar residues was calculated using a similar method.

Run-time complexity, implementation and availability

The alignment algorithm was implemented in Fortran90. The run time complexity for the main alignment program is $O[\min(L1, L2)^2 \times L1 \times L2]$, where $L1$ and $L2$ are the lengths for the target and template protein, respectively. The typical run time for proteins of length 250 on one 700 MHz Pentium III CPU was ~15 min. A searchable database of pre-calculated alignments may be found at http://www.bioinfo.rpi.edu/~bystrc/scali.html. Development of an installation package is in progress and will appear at the same site.

RESULTS

Validation of structure-based alignments

To assess its ability to reproduce state-of-the-art sequential structure-based alignments, SCALI was tested on a set of 120 pairs of distant structural homologs, defined as members of the same topology class in the CATH database but having <25% sequence identity. The alignments were compared with those from CE and DALI programs. All three methods produced similar aligned substructures. However, in CE and DALI alignments there were segments that should not have been aligned by the intuitive criteria defined above. Specifically, aligned residues sometimes lacked local structural similarity and/or the aligned region was not compact.

To assess its ability to find non-sequential alignments, SCALI was compared with DALI, CE and SARF on a set of topologically different structures. One example is the alignment of structures 1fsfA and 1ig0A, shown in Figure 1. These two proteins were first aligned manually, and the manual alignment was used to develop the method. Both proteins contain parallel six-stranded β-sheets with five α-helices arranged anti-parallel to the strands, two on one side and three on the other. The sheet continues in both cases, but in 1ig0A the seventh strand is anti-parallel. The seven strands appear in order 1765234 in 1fsfA and 4321567 in 1ig0A (i.e. strand 1 in 1fsfA is the structural equivalent of strand 4 in 1ig0A, and so on). In our manual alignment, the six parallel strands and the five helices could be aligned with one circular permutation. SCALI reproduced this alignment, superimposing the 11 secondary structure units with RMSD of 5.4 Å (Fig. 1a).

Both CE and DALI produced sequential alignments with the β-sheets in the flipped orientation. CE aligned strands 4325 in 1fsfA with strands 4325 in 1ig0A, aligning unpaired strands to paired strands (Fig. 1b). Strand 1, which is between strands 2 and 5 in 1ig0A, was left unaligned. DALI aligned strands 43257 in 1fsfA to strands 32157 in 1ig0A, but one strand was skipped and hence the two strand 7's point in opposite directions (Fig. 1c). Both CE and DALI alignments contained additional alignments of non-equivalent secondary structures. SARF was able to find the correct six-stranded sheet alignment but did not align three of the helices (Fig. 1d). The alignment is disjoint and several aligned segments are locally different. Two segments are aligned in reverse. Some of the β-strand alignments are offset by one residue (Fig. 1 legend).

To further examine the ability and the quality of aligning different topologies with automatic SCALI method, a set of 111 pairs of different topologies that had similar core structures were used. This list includes randomly selected proteins that shared the same architecture but differed in topology, according to CATH classification (Orengo *et al.*, 1997; Orengo, 1994; Pearl *et al.*, 2000, 2003). All of the pairwise alignments returned from CE, DALI, SARF and SCALI were compared with each other (Table 1) without manual curation. CE returned the alignments for all of the structure pairs, whereas DALI returned only 76 alignments, with 35 pairs rejected as non-superimposable. As expected, neither CE nor DALI returned non-sequential alignments, since they were not designed to do so. SARF returned alignments, including non-sequential and reverse alignments, for all of the test cases. In the alignments, three specific types of errors were observed and they are as follows:

(1) *Local non-equivalent error*: Non-equivalent secondary structures are aligned in 3D space.

(2) *Beta-strands misalignment error*: The alignment contains either the cross-aligned strands or unpaired strands. Cross-aligned strand error is where the paired β-strands are aligned in the opposite order (e.g. a four-stranded 1234 β-sheet is aligned to a 1324 β-sheet, with strand 2 is aligned to 2 and 3 is aligned to 3). Unpaired strand error is where paired strands (strands that are making hydrogen bonds) are aligned to unpaired strands.

(3) *Disjoint alignment error*: The alignment contains two or more segments that are spatially separate.

To evaluate the alignments from difference methods, an *FOM* was defined. *FOM* rewards aligned residues sharing the same secondary structure, having a Cα–Cα distance of $<\sim$3.5 Å in the final 3D alignment, and penalizes the three types of errors defined above. A lower *FOM* is better. Wilcoxson's sum of ranks test was performed to evaluate the statistical significance of the differences between the different alignment methods (for details see the Methods section).

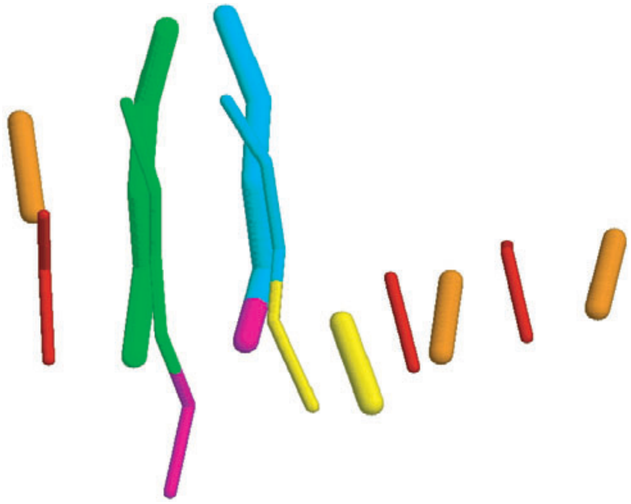

Fig. 2. Structural comparison between PDB:1cbf and PDB:2ts1 using SARF method. The alignment has RMSD of 2.76 with 76 residues aligned in space. The figure shows the C-alpha trace for the aligned β-strands only, with thicker line for 1cbf, and thinner line for 2ts1. The match positions in the alignment are shown in the same color for the two structures. The alignment contains the 'β-strands pairing errors' as defined in the Results section, which are illustrated as one strand having two different colors, each of which is aligned to the segment from another strand. Aligned segments (1cbf/2ts1): 22–23/187–188, 10/125, 24–30/216–222, 37–41/7–3*, 48–50/63–65, 57–66/50–59, 68–72/116–120, 78/172, 79–90/174–185, 95–97/31–33, 98–102/189–193, 109–121/198–210, 125–129/18–14*, 208/228. An asterisk (*) denotes reversed segments.

None of the CE, DALI and SARF methods returned error-free alignments. Both types of strand pairing misalignment ('unpaired strand' and 'cross-aligned strand' in Table 1) occurred in CE and DALI, which are similar to the strand pairing errors as shown in Figure 1b. SARF also made strand pairing errors, and one example is shown in Figure 2. In this alignment, one β-strand in 1cbf is aligned to two β-strands (in green and pink) and one β-strand in 2ts1 is aligned to two strands (in blue and pink), which result in non-equivalent paired hydrogen bonding among the aligned strands. The alignments returned by SCALI did not contain strand pairing errors due to its algorithmic design, which requires conserved contacts among the aligned segments. Compared with CE, the better performance of DALI on these difficult cases seems to be due to its ability to decide when to align and when not to. It will fail to return the alignment if a sequential structural comparison is too difficult, while CE will return the alignment anyway, even though it may contain many errors. Both types of strand misalignment result in a parallel displacement of one strand and cause only a minor increase in the RMSD.

SARF often produced a subset or superset of the SCALI alignment, but SARF aligned segments in reverse and allowed non-equivalent local structures to align. SCALI does not allow segments to be aligned in reverse. SARF alignments usually had disjoint pieces (41 out of 111). SARF alignments were often displaced by one residue from SCALI alignments. This would change the direction of the side chain in β-strands. Of 111, 91 of our alignments were compact and contained no obvious errors.

To further evaluate the quality of the 111 alignments generated from various structural alignment methods, the *FOM* was defined and computed [Equation (7) in the Methods section]. In principle, the *FOM* should capture the quality of the structural alignment by rewarding correct spatial alignments and penalizing the errors (as defined above). The *FOM* scoring function gives approximately equal penalties to the local non-equivalence errors and β-strand misalignment errors, and less penalty to the disjoint alignment errors. Using the *FOM scores* to evaluate the quality of each alignment method, SCALI performed the best on 81 out of 111 non-sequential alignments. If we ignore the 35 cases where DALI failed to align the structures, the best among the four methods is SCALI, followed by SARF, DALI and CE. The Wilcoxon's sum of signed ranks test was performed for all possible six paired comparisons and the results show that the differences between methods are statistically significant at the level of 0.1%. We should note that the choice of topologically different protein pairs favored SARF and SCALI over the other two methods.

Cluster analysis

By clustering pairwise SCALI alignments, we obtained non-sequential multiple structure alignments for several CATH architectures, including the 'up–down α bundle', 'β-sandwich', 'β-roll', '3-layer $\alpha\beta\alpha$ sandwich' proteins, and others. As an example, we choose the three-layer $\alpha\beta\alpha$ 'sandwich' (CATH code 3.40) for the all-against-all comparisons. This protein architecture is the most common and diverse, comprising 61 different topologies (Orengo *et al.*, 1997; Orengo, 1994; Pearl *et al.*, 2000, 2003). After clustering all 1830 alignments, 56 out of the 61 structures were divided into four subclasses (Fig. 3), each with a conserved core packing arrangement. The other five proteins (1b0pA, 1adn, 1div, 1inp and 1qhkA) had unique core packing arrangements. Proteins within a cluster, conserved at least 50 residues in a compact region that aligned with RMSD <4.0 Å. This cutoff produced some false negatives but very few false positives. Proteins within a cluster that fell below this significant value were still found to conserve the recurrent core, albeit more distorted.

The clustered alignments may be modeled as HMMs, where each aligned segment is a state and the variable sequential connections between the segments define the state–state transitions. Figure 4 shows diagrammatic HMMs for each of the $\alpha\beta\alpha$ clusters. In each model, some topological connections between the substructures are observed and others are not, probably reflecting the physical constraints on secondary structure packing (Honig, 1999). There are often compact subsets of connections that dominate, consistent with the previous argument that certain motifs, described as 'attractors', occur as the core of a protein's structure more frequently than others (Holm and Sander, 1996). An example is the right-handed parallel $\beta\alpha\beta$ motif.

In each cluster, all observed topologies are represented as pathways through the HMM. Based on these models, certain pathways exist that might represent proteins that have not been observed in crystal structures. For example, we may predict that the topology shown in Figure 5 is possible, based on the HMM for subclass A in Figure 4a. However, this topology has not yet been observed and would be considered as a novel fold if found.

When we analyzed the sequence information per position from the multiple structure alignments, we could clearly see a concentration of sequence information in the core positions. If we define the P-value as the likelihood of obtaining a similar distribution of polar and non-polar residues by chance given the number of observations (as described in the Methods section), all the high-information content positions (low P-value) tend to be deeply buried in the core. One example of such a result is shown in Figure 6, which shows the sequence information content for CATH architecture 3.40 subclass A (as in Fig. 4a).

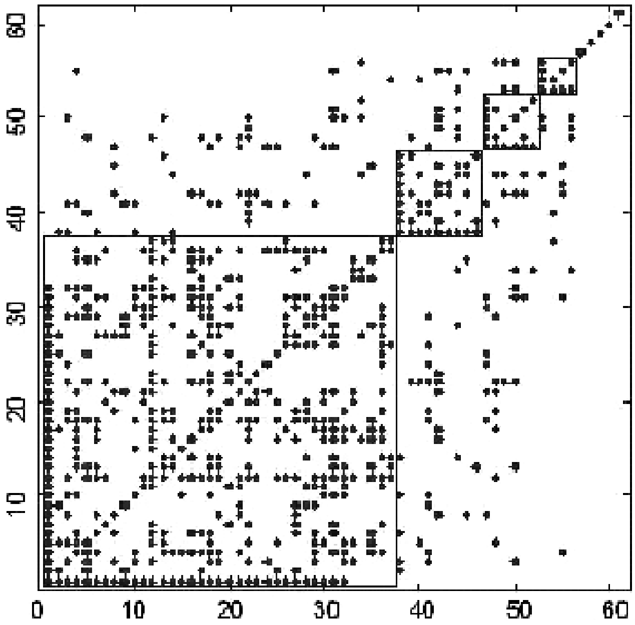

Fig. 3. All-against-all structure comparison and clustering for three-layer ($\alpha\beta\alpha$) proteins in CATH 3.40. A dot indicates that paired structures have a significant SCALI alignment. Bordered regions are four subclasses, A, B, C and D, listed here using the PDB code, chain and domain identifier (26). Subclass A: 1cbf01, 1aba00, 1ag8A1, 1ag8A2, 1alkA0, 1ami02, 1aua01, 1bg200, 1c8kA2, 1chmA1, 1dhs00, 1di6A0, 1dioB0, 1ekjA0, 1fuiA2, 1glaG2, 1iso00, 1lam01, 1lba00, 1poiB0, 1ra900, 1svq00, 1tplA2, 1udg00, 1vpt00, 1vsrA0, 2cevA0, 2ctc00, 2minB2, 2ts101, 3pmgA1, 3pmgA3, 1avpA0, 1rhs01, 1hfc00, 1ble00, 1cfr00. Subclass B: 1ctt02, 1a3aA0, 1cby00, 1cl8A0, 1eq6A0, 1fua00, 1g8tB0, 1uch00, 2bltA0. Subclass C: 1b94A0, 1b4uB0, 1e8gA3, 1eovA2, 1nox00, 1pvuA0. Subclass D: 1tdj03, 1br6A1, 1cfe00, 1pinA0.

Multiple SCALI alignments have been carried out for up–down bundle α-proteins, sandwich β-proteins, three-layer ($\alpha\beta\alpha$) and roll proteins in CATH database. The resulting models are shown in Figures 7 and 8. A full analysis of these conserved core packing arrangements is ongoing.

DISCUSSION

SCALI alignments are comparable to CE and DALI methods for comparing proteins that share the same topology, better if we agree that structure-based alignments should be compact and that aligned pieces should be locally similar in their backbone angles. In the cases where proteins share only a core packing arrangement but with different topologies, SCALI is able to find the proper structural equivalences, while previous methods fail, either because they assume a sequential ordering (DALI, CE) or because they do not

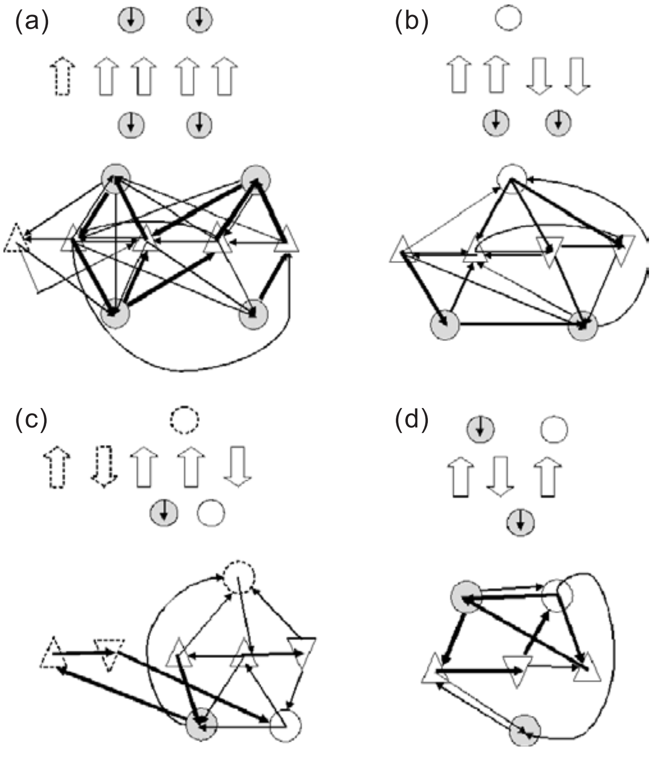

Fig. 4. Diagrammatic HMMs for the four subclasses of three-layer ($\alpha\beta\alpha$) proteins, A, B, C and D as defined in Figure 2. In each subclass, the upper panel shows the topology diagram without connectivities for that core structure. Strands are shown as arrows, and helices as circles. Shaded helices are pointing down (or into the page). Dotted lines indicate secondary structures that are sometimes present. The lower panel is the HMM drawn for that core. Strands are shown as triangles, and helices are shown as circles. The connectivities between the sub-structures are shown as arrows. Thicker lines indicate more frequent connections. (**a**) Subclass A: 37 proteins; (**b**) Subclass B: 9 proteins; (**c**) Subclass C: 6 proteins; and (**d**) Subclass D: 4 proteins.

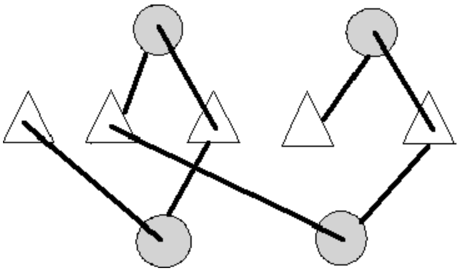

Fig. 5. A possible new fold topology. This fold has never been observed (according to the CATH released in January 2004) and yet is consistent with the model for subclass A of CATH architecture 3.40 (Fig. 4a).

enforce compactness, local equivalence and sequence similarity (SARF).

Multiple non-sequential alignments from SCALI have been used to construct non-linear profile HMMs, similar to the way profile HMMs have been constructed to model protein families and superfamilies (Eddy, 1998; Karplus *et al*., 1998; Gough and Chothia, 2002), and these may be useful in predicting structures that have recurrent core packing arrangements. Cluster analysis of our non-sequential alignments shows that some core packing arrangements have occurred dozens of times, each with a different topology. It is therefore reasonable to assume that there are many more permutants among the unsolved proteins.

Classification of cores

There are two widely cited classification schemes for protein domain structures, CATH (Orengo, 1994; Orengo *et al*., 1997; Pearl *et al*., 2000, 2003) and SCOP (Murzin *et al*., 1995). Both have a top-down hierarchy, starting from classes based on secondary structure content, then gross arrangement of secondary structure units, and then classes based on the topological connections between those units. At this level ('topology' in the CATH or 'fold' in SCOP), we expect structures to superimpose sequentially. The recurrent packing motifs discussed above represent a structure classification scheme that is more specific than 'architecture' but not as specific as 'topology'.

A new, intermediate classification level based on non-sequential multiple alignments may help us to understand the universe of protein folds. We may call these 'core' types, and apply codified names to each. For example, the model described in Figure 4a may be termed unambiguously as a '3-layer 2α(all down)-5β(all up)-2α(all down)' core. Figure 4b may be termed unambiguously as a '3-layer 1α(up)-4β(2 up, 2 down)-2α(all down)'. A numerical representation may be substituted for easy searching, such as 'a.00/b.11111/a.00' for Figure 4a, with /'s separating the structural layers and binary digits indicating the number and orientation of the secondary structures. However, some domains may not lend themselves easily to the 'layered' notation.

Can we make predictive models from non-sequential alignments?

The highest degree of sequence identity that we found in the SCALI non-sequential alignments at architecture level in the CATH database was only 12%. We cannot completely exclude the possibility of a common ancestor, but it is more likely that these core similarities are the result of convergent evolution, where energetic stability was the selection pressure. A conserved packing arrangement of secondary structures should energetically favor some sequence patterns, and this idea is supported by the results shown in Figure 6. Conserved secondary structure and 3D packing environment does appear to define conserved sequence patterns, at least binary (polar/non-polar) patterns.

We have observed certain core packing arrangements multiple times with different topologies. If these cores are recurrent themes in nature, then we might expect to see some future 'new folds' fall into these same classes. That is, 'new folds' may be permuted as 'old folds'. For example, the hypothetical protein yjiA from *Escherichia coli* (PDB code 1NIJ) solved in 2002 (Khil *et al*., 2004) was found to be a 'new fold' according to CASP5 (Moult *et al*., 2003). It was a new type of alpha–beta protein consisting of a single mixed β-sheet with strand order 15234 where strands 3 and 5 are anti-parallel to the others (Aloy *et al*., 2003). We have found a cluster of proteins that have the same core packing arrangement, all solved before 2002, and among the possible topologies given the HMM (Fig. 8a), was the topology of the new protein 1NIJ (Fig. 9). Since core alignments conserve sequence information,

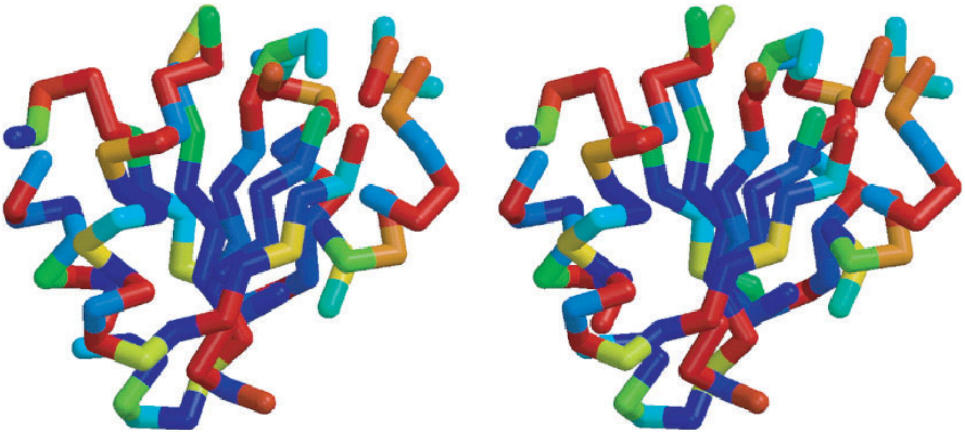

Fig. 6. The sequence information per position for subclass A in CATH3.40. The stereo image shows the core region of 1cbf (with C-alpha backbone trace only), a representative from CATH3.40 subclass A which consists of 34 topologies non-sequentially superimposable by SCALI (Fig. 3). Colors represent the information content of the combined sequence profiles at each aligned position, which is calculated as the P-value for obtaining the observed distribution of polar and non-polar amino acids by chance (as described in the Methods section). Blue represents a $P = 0.00$, red is $P = 0.30$ and higher. The P-value goes up in the hue scale from blue, through green, to red. The high-information content positions tend to be deeply buried in the core of the structure.

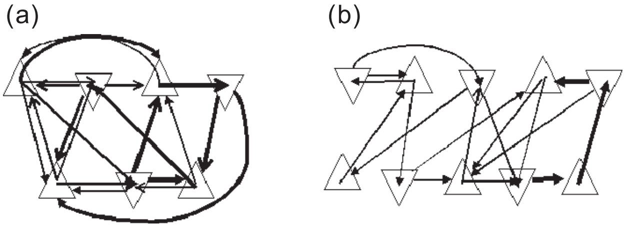

Fig. 7. Diagrammatic HMMs for the subclasses of 19 representative proteins in CATH 2.60 (β-sandwich) based on SCALI multiple alignments. Drawn as in Figure 4: (**a**) 12 proteins and (**b**) 3 proteins.

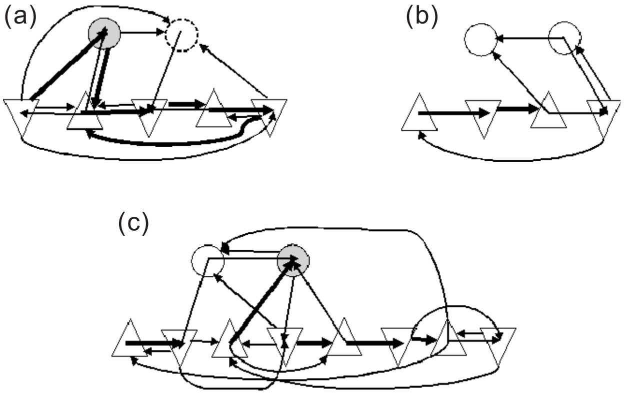

Fig. 8. Diagrammatic HMMs for the subclasses of 29 $\alpha\beta$ proteins in CATH architecture 3.10 (β-roll) based on SCALI alignments. Drawn as in Figure 4: (**a**) 6 proteins, (**b**) 5 proteins and (**c**) 2 proteins.

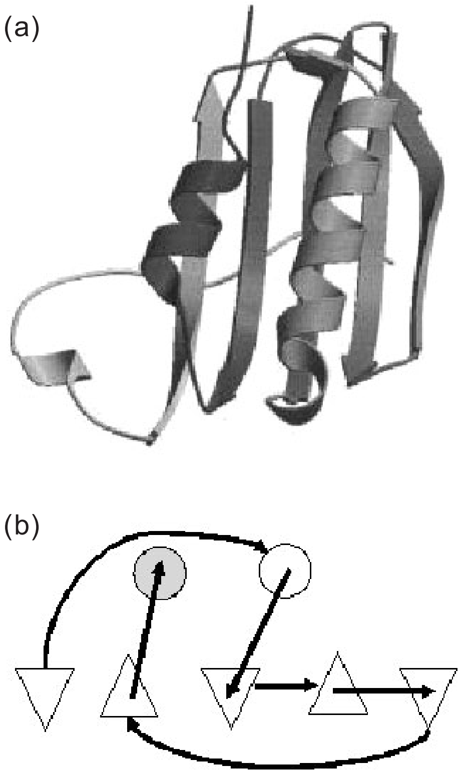

Fig. 9. Topology of 1NIJ, which was a new fold in 2002. (**a**) Structure of 1NIJ. (**b**) Topology of 1NIJ, which belongs to the first subclass of CATH 3.10 (Fig. 8a).

and cores are often recurrent; therefore, self-avoiding HMMs based on SCALI alignments have the potential for predicting the core structure of topologically novel proteins based on the sequence alone.

ACKNOWLEDGEMENT

This research was supported by NSF grant EIA-0229454 and DBI-343206.

REFERENCES

Abagyan,R.A. and Maiorov,V.N. (1989) An automatic search for similar spatial arrangements of alpha-helices and beta-strands in globular proteins. *J. Biomol. Struct. Dyn.*, **6**, 1045–1060.

Alexandrov,N.N. (1996) SARFing the PDB. *Protein Eng.*, **9**, 727–732.

Alexandrov,N.N. and Fischer,D. (1996) Analysis of topological and nontopological structural similarities in the PDB: new examples with old structures. *Proteins*, **25**, 354–365.

Aloy,P., Stark,A., Hadley,C. and Russell,R.B. (2003) Predictions without templates: new folds, secondary structure, and contacts in CASP5. *Proteins*, **53**(Suppl. 6), 436–456.

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Bennett,M.J., Choe,S. and Eisenberg,D. (1994) Domain swapping: entangling alliances between proteins. *Proc. Natl Acad. Sci. USA*, **91**, 3127–3131.

Bernstein,H.J. (2000) Recent changes to RasMol, recombining the variants. *Trends Biochem. Sci.*, **25**, 453–455.

Bystroff,C. and Baker,D. (1998) Prediction of local structure in proteins using a library of sequence–structure motifs. *J. Mol. Biol.*, **281**, 565–577.

Bystroff,C., Thorsson,V. and Baker,D. (2000) HMMSTR: a hidden Markov model for local sequence–structure correlations in proteins. *J. Mol. Biol.*, **301**, 173–190.

Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Efimov,A.V. (1995) Structural similarity between two-layer alpha/beta and beta-proteins. *J. Mol. Biol.*, **245**, 402–415.

Flores,T.P., Orengo,C.A., Moss,D.S. and Thornton,J.M. (1993) Comparison of conformational characteristics in structurally similar protein pairs. *Protein Sci.*, **2**, 1811–1826.

Gibrat,J.F., Madej,T. and Bryant,S.H. (1996) Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.*, **6**, 377–385.

Gong,W., O'Gara,M., Blumenthal,R.M. and Cheng,X. (1997) Structure of pvu II DNA-(cytosine N4) methyltransferase, an example of domain permutation and protein fold assignment. *Nucleic Acids Res.*, **25**, 2702–2715.

Gough,J. and Chothia,C. (2002) SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. *Nucleic Acids Res.*, **30**, 268–272.

Holm,L. and Sander,C. (1993) Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, **233**, 123–138.

Holm,L. and Sander,C. (1996) Mapping the protein universe. *Science*, **273**, 595–603.

Honig,B. (1999) Protein folding: from the levinthal paradox to structure prediction. *J. Mol. Biol.*, **293**, 283–293.

Hou,Y., Hsu,W., Lee,M.L. and Bystroff,C. (2003) Efficient remote homology detection using local structure. *Bioinformatics*, **19**, 2294–2301.

Iwakura,M., Nakamura,T., Yamane,C. and Maki,K. (2000) Systematic circular permutation of an entire protein reveals essential folding elements. *Nat. Struct. Biol.*, **7**, 580–585.

Janowski,R., Kozak,M., Jankowska,E., Grzonka,Z., Grubb,A., Abrahamson,M. and Jaskolski,M. (2001) Human cystatin C, an amyloidogenic protein, dimerizes through three-dimensional domain swapping. *Nat. Struct. Biol.*, **8**, 316–320.

Jeltsch,A. (1999) Circular permutations in the molecular evolution of DNA methyltransferases. *J. Mol. Evol.*, **49**, 161–164.

Jung,J. and Lee,B. (2001) Circularly permuted proteins in the protein structure database. *Protein Sci.*, **10**, 1881–1886.

Karplus,K., Barrett,C. and Hughey,R. (1998) Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, **14**, 846–856.

Khil,P.P., Obmolova,E., Teplyakov,A., Howard,A.J., Gilliland,G.L. and Camerini-Otero,R.D. (2004) Crystal structure of the *Escherichia coli* YjiA protein suggests a GTP-dependent regulatory function. *Proteins*, **54**, 371–374.

Koehl,P. (2001) Protein structure similarities. *Curr. Opin. Struct. Biol.*, **11**, 348–353.

Milik,M., Szalma,S. and Olszewski,K.A. (2003) Common structural cliques: a tool for protein structure and function analysis. *Protein Eng.*, **16**, 543–552.

Moult,J., Fidelis,K., Zemla,A. and Hubbard,T. (2003) Critical assessment of methods of protein structure prediction (CASP)-round V. *Proteins*, **53**(Suppl. 6), 334–339.

Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.

Orengo,C.A. (1994) Classification of protein folds. *Curr. Opin. Struct. Biol.*, **4**, 429–440.

Orengo,C.A., Michie,A.D., Jones,S., Jones,D.T., Swindells,M.B. and Thornton,J.M. (1997) CATH—a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.

Ortiz,A.R., Strauss,C.E. and Olmea,O. (2002) MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci.*, **11**, 2606–2621.

Pearl,F.M., Lee,D., Bray,J.E., Sillitoe,I., Todd,A.E., Harrison,A.P., Thornton,J.M. and Orengo,C.A. (2000) Assigning genomic sequences to CATH. *Nucleic Acids Res.*, **28**, 277–282.

Pearl,F.M., Bennett,C.F., Bray,J.E., Harrison,A.P., Martin,N., Shepherd,A., Sillitoe,I., Thornton,J. and Orengo,C.A. (2003) The CATH database: an extended protein family resource for structural and functional genomics. *Nucleic Acids Res.*, **31**, 452–455.

Rabiner,L.R. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, **77**, 257–286.

Rost,B. (1997) Protein structures sustain evolutionary drift. *Fold Des.*, **2**, S19–S24.

Sayle,R.A. and Milner-White,E.J. (1995) RASMOL: biomolecular graphics for all. *Trends Biochem. Sci.*, **20**, 374.

Schiering,N., Casale,E., Caccia,P., Giordano,P. and Battistini,C. (2000) Dimer formation through domain swapping in the crystal structure of the Grb2-SH2-Ac-pYVNV complex. *Biochemistry*, **39**, 13376–13382.

Shao,Y. and Bystroff,C. (2003) Predicting interresidue contacts using templates and pathways. *Proteins*, **53**(Suppl. 6), 497–502.

Shindyalov,I.N. and Bourne,P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.*, **11**, 739–747.

Smith,V.F. and Matthews,C.R. (2001) Testing the role of chain connectivity on the stability and structure of dihydrofolate reductase from *E. coli*: fragment complementation and circular permutation reveal stable, alternatively folded forms. *Protein Sci.*, **10**, 116–128.

Sokal,R.R. and Rohlf,F.J. (eds) (1973) *Introduction to Biostatistics*. W.H. Freeman and company, San Francisco, CA, pp. 220–222.

Szustakowski,J.D. and Weng,Z. (2000) Protein structure alignment using a genetic algorithm. *Proteins*, **38**, 428–440.

Szustakowski,J.D. and Weng,Z.(2002) Protein structure alignment using evolutionary computing. In Fogel,G. and Corne,D. (eds), *Evolutionary Computation in Bioinformatics*, Morgan Kaufman.

Taylor,W.R. and Orengo,C.A. (1989) Protein structure alignment. *J. Mol. Biol.*, **208**, 1–22.

Viguera,A.R., Blanco,F.J. and Serrano,L. (1995) The order of secondary structure elements does not determine the structure of a protein but does affect its folding kinetics. *J. Mol. Biol.*, **247**, 670–681.

Westhead,D.R., Slidel,T.W., Flores,T.P. and Thornton,J.M. (1999) Protein structural topology: automated analysis and diagrammatic representation. *Protein Sci.*, **8**, 897–904.

Yang,A.S. and Honig,B. (1999) Sequence to structure alignment in comparative modeling using PrISM. *Proteins*, Suppl. 3, 66–72.

Yang,A.S. and 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.