Published online 19 May 2011

Nucleic Acids Research, 2011, Vol. 39, No. 14 ***e94***
doi:10.1093/nar/gkr348

Biological insights from topology independent comparison of protein 3D structures

Minh N. Nguyen[1] and M. S. Madhusudhan[1,2,3,*]

[1]Bioinformatics Institute, 30 Biopolis Street, #07-01, Matrix, Singapore 138671, [2]Department of Biological Sciences, National University of Singapore and [3]School of Biological Sciences, Nanyang Technological University, Singapore

Received February 26, 2011; Accepted April 25, 2011

ABSTRACT

Comparing and classifying the three-dimensional (3D) structures of proteins is of crucial importance to molecular biology, from helping to determine the function of a protein to determining its evolutionary relationships. Traditionally, 3D structures are classified into groups of families that closely resemble the grouping according to their primary sequence. However, significant structural similarities exist at multiple levels between proteins that belong to these different structural families. In this study, we propose a new algorithm, CLICK, to capture such similarities. The method optimally superimposes a pair of protein structures independent of topology. Amino acid residues are represented by the Cartesian coordinates of a representative point (usually the C^{α} atom), side chain solvent accessibility, and secondary structure. Structural comparison is effected by matching cliques of points. CLICK was extensively benchmarked for alignment accuracy on four different sets: (i) 9537 pair-wise alignments between two structures with the same topology; (ii) 64 alignments from set (i) that were considered to constitute difficult alignment cases; (iii) 199 pair-wise alignments between proteins with similar structure but different topology; and (iv) 1275 pair-wise alignments of RNA structures. The accuracy of CLICK alignments was measured by the average structure overlap score and compared with other alignment methods, including HOMSTRAD, MUSTANG, Geometric Hashing, SALIGN, DALI, GANGSTA^{+}, FATCAT, ARTS and SARA. On average, CLICK produces pair-wise alignments that are either comparable or statistically significantly more accurate than all of these other methods. We have used CLICK to uncover relationships between (previously) unrelated proteins. These new biological insights include: (i) detecting hinge regions in proteins where domain or subdomains show flexibility; (ii) discovering similar small molecule binding sites from proteins of different folds and (iii) discovering topological variants of known structural/sequence motifs. Our method can generally be applied to compare any pair of molecular structures represented in Cartesian coordinates as exemplified by the RNA structure superimposition benchmark.

INTRODUCTION

The intimate relationship between protein structure and function has now been well established (1,2). Comparing protein structures with one another has helped make evolutionary and functional links amongst proteins. Traditionally, proteins related to one another by structural similarity were classified into families that mostly followed a pattern of sequence similarity (3–6). Several methods of protein structure comparison can be used for this purpose (2,7–20). Most of these methods match protein structures only when their overall topologies also match, i.e. amino acid sequence order is preserved in alignments. When one uses non-sequential and non-topological protein structure matching programs (13,15,20–25), new structural relationships between proteins emerge (10,26,27). The protein databank (PDB) (28) now has over 70 000 structures of proteins. These are variously categorized into protein families, usually numbering about a couple of thousand (3–5,29,30). In this study, we lay the basis to investigate relationships between proteins belonging to different families that are not obvious from these categorizations.

*To whom correspondence should be addressed. Tel: (65) 6478 8500; Fax: (65) 6478 9047; Email: madhusudhan@bii.a-star.edu.sg

When comparing a pair of structures, the measures of similarity usually include geometric accuracy, such as a root mean square deviation (RMSD) value and structural coverage of the match. The number of residues that align when the three-dimensional (3D) structures are superimposed often quantifies coverage. The aim of most structure superimposition programs is to accurately detect the extent of similarity (maximize coverage) and superimpose the structures to display geometric similarity (minimize RMSD). Many different methods have been proposed to achieve these measures of accuracy [for comprehensive reviews see refs. (20,31)]. Introduced here is our method, CLICK, in which small cliques of points from both structures are first matched by a least squares fit. The cliques comprise 3–7 amino acid (or nucleotide) residues, each represented by one or more points. The clique-matching step gives us a one-to-one mapping of the equivalent residues in the two structures. A structural superimposition of the equivalent residues gives the final structural alignment. The algorithm does not consider chain connectivity when considering the cliques of points. CLICK is hence a topology independent structure superimposition program. The method is not restricted to the comparison of the 3D structures of proteins alone; it can compare any two constellations of points that are represented in Cartesian coordinates. In this study we have devoted the most attention to aspects of protein 3D structural comparisons. However, to showcase the versatility of CLICK, we have included one small benchmark on RNA structure comparison.

Because CLICK primarily detects similarity in local packing of amino acid residues, it is ideally suited to detect similarities of biological significance that are a consequence of local structural similarity. Specifically, CLICK is best suited to investigate similarities in: (i) hydrophobic cores (31); (ii) enzymatic active sites; (iii) small molecule binding sites (32,33); and (iv) protein–protein interfaces (33–35). Our method also reports more than one alignment between a pair of proteins, if there are detectable conformational changes. CLICK produces as many alignments as necessary to maximize coverage in amino-acid residue equivalences. These features of CLICK make it ideally suited to discover instances of divergent evolution that are hard to detect, owing to changes in protein topology, and instances of convergent evolution where sub-structures of proteins are similar to one another at different length scales.

The results presented in the next section can be broadly separated into two halves. In the first part we describe the optimization of parameters and benchmark the algorithm against other popular methods. This is to validate the efficacy of the method under test conditions, including data sets of pairs of proteins that are topologically similar and pairs that are topologically different. Included in the benchmarking is a comparison of RNA structures to illustrate the multifaceted nature of our algorithm. In the second half, we illustrate the utility of CLICK with examples of its application including: (i) identifying conformational changes; (ii) recognizing the similarity in ATP binding sites in two evolutionarily unrelated proteins; and (iii) identifying sequence motif for nucleotide binding in proteins with different topologies.

MATERIALS AND METHODS

CLICK algorithm

The CLICK algorithm detects local structural similarity, independent of topology. The algorithm is generally applicable on any pair of structures, the position of whose constituents are specified in a given feature space. To explain the working principle of the algorithm, we mainly focus on the 3D structures of proteins where the feature space includes Cartesian coordinates, secondary structure and solvent accessible surface area. The feature space could be customized for comparison of other biomolecules. See the results section for a benchmark on RNA structure alignments.

The algorithm consists of four sequential steps, as follows.

Extracting features. Residues in a protein are represented by the Cartesian coordinates of one representative atom (typically the C^{α}), side-chain solvent accessibility and secondary structure. The secondary structures and solvent accessibilities of residues were computed using MODELLER9v7 (36). MODELLER uses the DSSP algorithm to assign secondary structures (37) and the algorithm of Richmond and Richards for solvent accessibilities (38). The solvent accessibility of an amino acid residue is considered as a buried if the side chain accessibility <8%, intermediate, if side chain accessibility is between 8 and 30%, or exposed otherwise.

Forming cliques. For each of the two structures to be compared, A and B, all possible internal pair-wise distances between the representative atoms are computed. We define a clique as a subset of n points, where the Euclidean distance between any pair within the clique is within a predefined threshold, d_{thr} (Equation 1).

Let S^n be the set of all possible cliques of n points. If $A^n \subset S^n$, then all pair-wise distances of A^n satisfy

$$D[A_i,A_j] < d_{\mathrm{thr}}, \quad (1)$$

where D is the pair-wise distance between two representative atoms A_i and A_j, and A_i, $A_j \subset A^n$. Optimal values of d_{thr}, for different values of n, were computed using a grid search (see 'Results' section).

Imposing the distance threshold criterion reduces the number of possible n-body cliques in a protein to a tractable range. For instance, in the case of phosphotyrosine protein phosphatase (PDB code 1phr, chain A, length 154 residues), all 596 904 possible subsets of three points are reduced to 4480 three-body cliques when a threshold d_{thr} of 10 Å is imposed.

Clique matching. The objective is to compute a one to one mapping between amino acid residues of the two structures A and B. To begin with, all possible 3-body cliques A^3 and B^3, where A^3 and $B^3 \subset S^3$, are compared to one

Table 1. The matrix of empirically determined equivalences between secondary-structure elements in proteins

SS	Coil	α-helix	β-strand
Coil	0	1	1
α-helix	1	0	2
β-strand	1	2	0

Table 2. The matrix of empirically determined equivalences between different solvent accessible area classes in proteins

SA	Buried	Intermediate	Exposed
Buried	0	1	2
Intermediate	1	0	1
Exposed	2	1	0

another (inclusive of all permutations). Equivalent pairs of cliques are deduced according to the relations in Equations (2–4). A pair of (A^3, B^3) is matched if their RMSD on superimposition is smaller than a preset threshold ($RMSD_3 = 0.15$ Å). RMSD between cliques is calculated by 3D least squares fit (39).

Additionally, amino acid residue secondary-structure state and side chain solvent accessible area also determines what pair of cliques are matched. Secondary structure provides the general three-dimensional form of local segments of proteins while side-chain solvent accessibility is the degree to which a residue in a protein is accessible to a solvent molecule. For matching of a pair of cliques in our algorithm, the secondary-structure score between two equivalent residues A_i and B_j are compared [Equation (2)]

$$SSM[A_i,B_j] < s, \tag{2}$$

where

$$SSM[A_i,B_j] = \left\{ \begin{array}{ll} 0, & \text{if } SS(A_i) = SS(B_j) \\ 1, & \text{if } (SS(A_i) \neq SS(B_j)) \\ & \text{and } (SS(A_i) = \text{Coil or } SS(B_j) = \text{Coil}) \\ 2, & \text{otherwise(Table1)} \end{array} \right\}$$

SSM is an empirically determined secondary-structure match matrix (Table 1), $SS(A_i)$ is the secondary-structure state of amino acid residue A_i, and s is a preset threshold for matching secondary-structure elements. The cut-off threshold for comparing secondary structure used in this study was 2, hence $SSM[A_i,B_j] < 2$ [Equation (2)]. This implies that, either residues of regular secondary structures can only match with other residues of the same secondary structure, or with residues in loops.

The solvent accessibility score between two residues A_i and B_j from solvent accessibility matrix (Table 2) are matched by using the inequality [Equation (3)]:

$$SAM[A_i,B_j] < a, \tag{3}$$

Table 3. Optimal values of the threshold RMSD for different values of clique size (n)

n	3	4	5	6	7	8	9
$RMSD_n$ (Å)	0.15	0.30	0.60	0.90	1.50	1.80	2.10

where

$$SAM[A_i,B_j] = \left\{ \begin{array}{ll} 0, & \text{if } SA(A_i) = SA(B_j) \\ 0, & \text{if } (SA(A_i) \neq SA(B_j)) \\ & \text{and } (|SS(A_i) = SS(B_j)| \leq 10) \\ 1, & \text{otherwise (Table2)} \end{array} \right\}.$$

SAM is an empirical solvent accessibility match matrix (Table 2), $SA(A_i)$ is the side-chain solvent accessibility of amino acid residue A_i, and a is a preset threshold for matching solvent accessible area states. The cut-off threshold for solvent accessibility matching is $a = 1$, implying that residues categorized in different accessible area classes cannot be matched. However, this criterion is relaxed to allow the matching of two residues in adjacent accessible area classes if their side chain accessible areas are within 10% of each other.

Next, A^3 and B^3 are extended to 4-body cliques A^4 and B^4, by including one residue, A_i and B_j respectively, subject to the distance threshold criterion [Equation (1)]. This new pair (A_i, B_j) and matched residues of (A^3, B^3) are used to superimpose the pair of cliques $A^4 = A^3 \cup A_i$ and $B^4 = B^3 \cup B_j$. Pairs of four-body cliques, A^4 and B^4, are matched if their RMSD is smaller than another preset threshold, $RMSD_4 = 0.30$Å [Equation (4), $n = 4$].

$$RMSD(\{A^{n-1} \cup A_i\},\{B^{n-1} \cup B_j\}) < RMSD_n. \tag{4}$$

Pairs of n body cliques, $A^n = A^{n-1} \cup A_i$ and $B^n = B^{n-1} \cup B_j$ are selected if their RMSD is smaller than a preset threshold $RMSD_n$ (Table 3). Every value of n has a different RMSD threshold, $RMSD_n$. See the section on RMSD threshold optimization for details. At every step the secondary structure and accessible area comparisons [Equations (2) and (3)] are also performed.

All matched pairs of 4-body cliques A^4 and B^4 are extended to all possible higher order cliques, A^n and B^n, where A^n, $B^n \subset S^n$ and $n > 4$. In this study, cliques are extended to a maximum of seven constituent residues.

Alignment. Matching cliques helps in identifying structurally equivalent residues in the two structures. Using these equivalences, a final 3D least squares fit is performed to superimpose the two structures (Figure 1). Given that the matching of cliques is not unique, i.e. many cliques comparisons could fit the criteria for a match, of all the possible least squares fits, the comparison that yields the best-structure overlap (SO_{best}) is considered [Equation (5)]:

$$SO_{best} = \max\{\text{SO of global alignment of all } (A^n,B^n)\}, \tag{5}$$

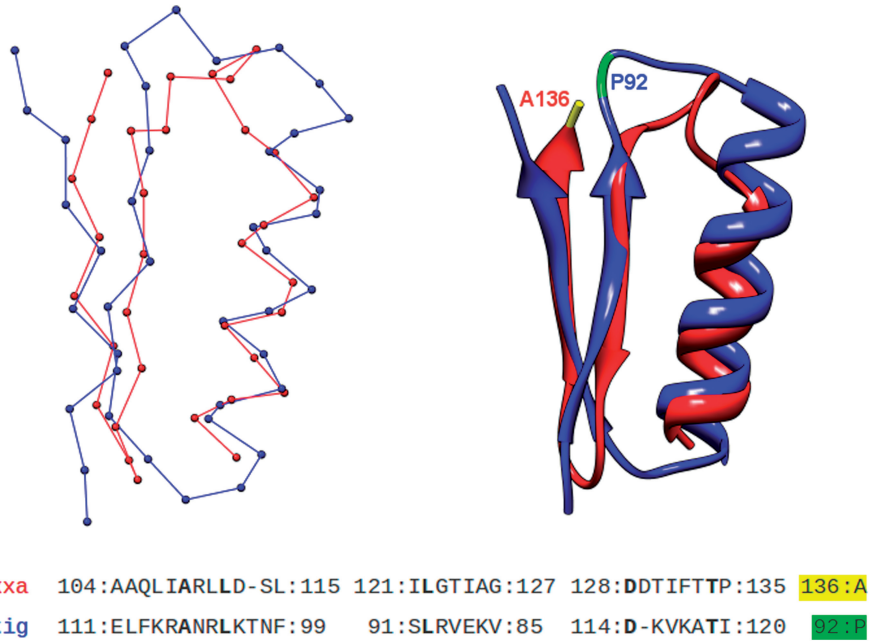


Figure 1. The structural alignment of two topologically different yet structurally similar proteins, PDB codes 1xxa (red) and 1tig (blue), according to CLICK. The sequence alignment implied by the structural superimposition is shown below. Residues in bold lettering represent conservation. The last residue, 136A, on 1xxa is matched with residue 92P of 1tig (colored yellow and green, respectively, in sequence and structure). This anomalous match is 'corrected' using heuristics (see Heuristics to maintain chain continuity section of methods).

where SO is the structure overlap of a matched pair of n-body cliques A^n and B^n.

Heuristics to maintain chain continuity

The matching of cliques is independent of the chain connectivity of the constituent residues. This sometimes results in anomalous matching of residues (Figure 1). To correct these anomalous matches, heuristic rules are introduced to maintain local chain connectivity. These rules are (i) matched segments must have a minimum length of five residues (gap matching not included); (ii) adjacent segments separated by five or less residues are joined; (iii) gaps are appropriately introduced when the region between the two adjacent segment were of unequal length in the two proteins. To illustrate, while aligning a pair of proteins 1xxa and 1tig (Figure 1), the residues in segments [104:115], [121:127], [128:135] of 1xxa were aligned with the residues in segments [111:99], [91:85], [114:120] of 1tig, respectively. The anomaly in the alignment is the matching of A136 in 1xxa with P92 of 1tig, because the sequence stretches that are in the immediate neighbourhood of these residues do not structurally superimpose on one another. Applying the heuristic rules corrects the alignment. We foresee such anomalies to occur frequently with residues that are a part of long floppy loop regions.

Detecting conformational changes

When one of the proteins undergoes a conformational change, such as a rigid body shift of a part of the structure, most existing methods usually only align the largest similar sub-structures. CLICK however provides alignments between the unmatched regions, should they be structurally similar.

Consider a pair of proteins, A and B, that have the same fold but are not structurally identical because of inter-domain (or inter-subdomain) reorientation. If the domain reorientation was larger than $RMSD_n$, CLICK first computes the superimposition of residues that result in the largest structure overlap. A match is then found between the residues of both proteins that were not aligned/superimposed in the first parse. This procedure is iterated till the number of unaligned residues is 10 or lower.

Alignment measures

RMSD. Given two proteins A and B, the RMSD is the norm of the distance vector between the two sets of coordinates of representative atoms, after superimposition. It is given by

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\left\|x_i^A - x_i^B\right\|^2\right)}, \quad (6)$$

where N is the number of atoms in the list of equivalence, and x_i^A and x_i^B are the Cartesian coordinates of representative atoms of structurally equivalent amino acid residues of proteins A and B (40).

Structure overlap. Structure overlap (SO; also called equivalent positions) is defined as the percentage of the representative atoms in the protein A that are within 3.5

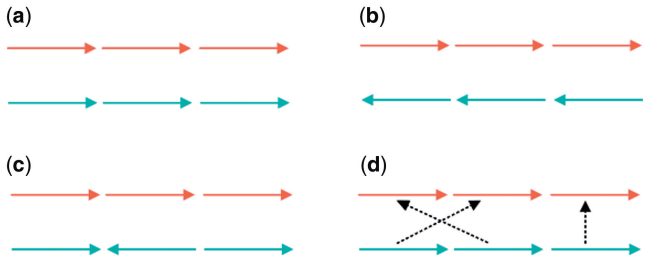


Figure 2. The topology scores of different structure alignments. In each of the four examples, the two proteins that are matched consist of three different sequence segments. The directions of the arrows that symbolize each segment show the direction from N- to C-termini. Unless explicitly indicated with black arrows, the segments in the top structure are aligned to segments directly below. Four different cases of sequence alignments implied by CLICK are illustrated here. (**a**) Alignment that maintains topology; topology score = 1. (**b**) The directionality from N to C of the sequence on top is the exact opposite of that to the one below; topology score = 0. (**c**) Two of the three sequence segments have the same directionality; topology score = 0.66. (**d**) Two of the three segments are matched but not in sequential order; topology score = 0.66.

Å (for the RNA alignments this value is set to 4.0 Å) of the corresponding atoms in the superimposed protein B (41).

Fragment score. On applying heuristic measures to maintain chain or fragment continuity, some residue matches are excluded from consideration, as they do not belong to (or are in the close proximity of) contiguously matched fragments. The fragment score is the ratio of the number of matched positions in the alignment before and after the application of heuristics. This is a handy measure to estimate the extent of similarity between two protein structures especially when they are of dissimilar fold. For structures of similar fold (and size) the fragment score is close to 1 (the maximum value). In the example of aligning residues of 1xxa with 1tig (Figure 1), the fragment score of the alignment was 0.89.

Topology score. The topology score is a measure of how similar the topologies of the matched structures are to one another. It is computed based on the directionality of the matched sequence fragments (Figure 2). Topology score varies between a maximum of 1 for topologically identical structures and 0 for those are the topologically completely dissimilar.

Methods compared

The performance of CLICK was compared with other popular structural alignment methods including MUSTANG (13), Geometric Hashing (C-alpha Match) (24,25,33–35), SALIGN (19), DALI (22,42,43), GANGSTA$^+$ (44) and FATCAT (18) on the protein benchmark data sets. All these programs were run using default parameters, and no effort was made to adjust the parameters for specific cases. For the comparison of 3D structures of RNA, CLICK was compared with ARTS (45,46) and SARA (47,48).

Tests for statistical significance

To estimate the statistical significance of the comparisons described above, the non-parametric Wilcoxon signed rank test was used (49). The software Octave (http://www.gnu.org/software/octave/index.html) was used to perform the Wilcoxon tests.

Alignment data sets

Homologous proteins. CLICK and the other methods were tested for alignment accuracy on a data set of structures of pairs of homologous proteins. 9537 pair-wise alignments implied by the HOMSTRAD (http://tardis.nibio.go.jp/homstrad/) database of multiple alignments were used for this data set. In all, this data set comprised of 3454 structures.

Difficult cases of aligning homologous protein pairs. This second data set is used to quantitatively assess the performance of CLICK when structure similarity is low, as in the case of distant homologues. They include 64 pair-wise alignments from HOMSTRAD database with $30\% < SO < 70\%$ and RMSD >2.5 Å. The alignment accuracy of CLICK is compared to results obtained from MUSTANG, Geometric Hashing, DALI, SALIGN, GANGSTA$^+$ and FATCAT for this data set.

Similar structure but different topology. This data set includes 199 pair-wise alignments, including circular permutations (5 pairs) (16,27), non-topological alignments (60 pairs) (23) (http://bioinfo3d.cs.tau.ac.il/MASS/examples.html#non-topological), swapped domains (24 pairs) (16,50) and 110 pair-wise alignments amongst 10 members of retinol binding protein family, five members of verotoxin family and four members of the pleckstrin homology domain family (Supplementary Figure S5 and Supplementary Tables S1a and S1b in Supplementary Data). Here, CLICK is compared to the other sequence-order-independent methods including MUSTANG, Geometric Hashing, DALI and GANGSTA$^+$. Structural similarities between proteins in this data set could be indicative of evolutionary relationships between the proteins despite pronounced structural differences in fold (topology) (16).

Proteins in different conformations. The structural alignments of alternative conformations produce a useful comparison of structures that exhibit domain motion or rigid body shifts. A characterization of such motions may lead to an improved understanding of the relationship between structure and function. The structure comparisons of alterative conformations were carried out for 22 proteins. This included 20 proteins from the Hinge Atlas (51), DNA Polymerase Beta proteins (2bpf, 2fmq) and maltodextrin binding proteins (1omp, 1anf). On this data set, CLICK results were compared to those of FATCAT.

RNA-structure comparisons. A subset of 1275 pair-wise alignments was benchmarked by the SARA server (47). This subset contained alignments between 51 RNA

structures each having between 19 and 157 nucleotide residues (http://sgu.bioinfo.cipf.es/datasets/RNA/NR95-HR.txt).

Implementation of CLICK

The CLICK algorithm has been implemented in C++. The program run time increases with increase in (i) sizes of input structures and (ii) number of best matched cliques. The exact dependence on these two factors is difficult to compute as they may be independent of one another. If the Euclidian distances between the representative atoms, secondary structures and solvent accessibilities of residues were pre-computed, on average CLICK took 1 s to perform a single alignment of a pair of proteins each of size ~150 residues on a Ubuntu 8.04 Linux platform with 3.00 GHz CPU (Core 2 Duo E8400) and 3.5GB primary memory. A web server of the program can be found at: http://mspc.bii.a-star.edu.sg/click.

All atomic, cartoon, ribbon and surface representation of proteins and RNA shown in this study were rendered using Chimera (52).

RESULTS

In this section, we begin by describing our efforts at optimizing parameters and extensive benchmarking. We establish that our method performs better than, or at the least, at par with popular existing methods. Subsequently, we demonstrate a few ways in which CLICK can be utilized to obtain new biological insights from structural comparisons including determining new evolutionary relationships, detecting hinge regions in proteins where domain or sub-domains show flexibility, and identifying binding site/motif similarities despite topological differences.

Optimal parameters

Optimization of clique size, distance threshold and RMSD. To optimize CLICK parameters such as clique size, distance threshold and RMSD cutoff for protein structure comparisons, a grid search was performed over the data set of difficult HOMSTRAD cases, comprising 64 pair-wise alignments. The grid search was performed by varying the number of clique members, n, in the range [3, 9] and cut-off distance, d_{thr}, in the interval 6–12 Å in steps of 0.5 Å (Figure 3). At each step, the structure overlap (SO) value was computed. While matching cliques, residue equivalences are decided upon after a least squares fit [Equation (5)]. These data show that the larger the clique size, the better is the SO value. The variation in SO score was small, almost negligible, when $n > 6$ and the cut-off distance was >9 Å. Running time also increases with clique size. For practical considerations, we chose to limit clique size at $n = 7$. Using a grid search, the optimal cut-off distance, d_{thr}, for $n = 7$ was determined to be 10 Å (Figure 3). The SO values appear to saturate after $n = 7$. The cut-off distance value at which the SO value shows no further change is 10 Å, which is chosen as the cutoff value in the algorithm.

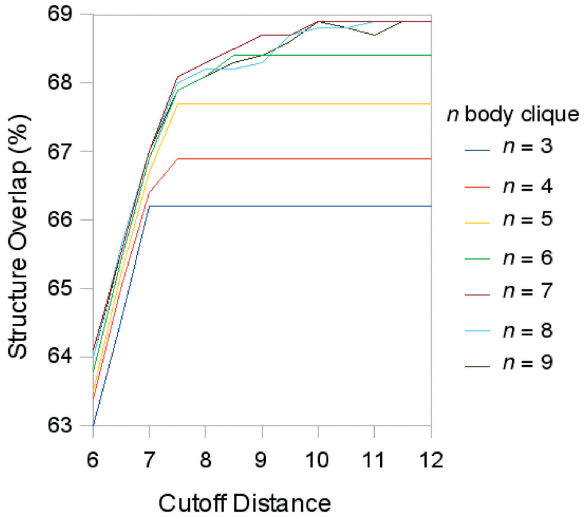


Figure 3. The choice of optimal threshold cut-off distances (in this case C^{α}–C^{α} distances within a protein) that describe different n-body cliques (i.e. different value of n). The optimal distance cut-off is that after which the structure overlap value changes by <0.05.

Another grid search was performed with n in the range [3, 9] and RMSD cut-off threshold in the interval 0.0–3.0 Å in steps of 0.05 Å. Again, SO was computed at each step. The optimal value of RMSD cut-off for a particular clique size was chosen as that value above which there was no change in the SO value (Table 3). In general, the RMSD cut-off value increases monotonically with clique size. Note that in all these optimizations, the C^{α} atom was chosen to represent the amino acid residues.

Contribution of secondary structure and solvent ccessibility. To evaluate the contribution of secondary structure and solvent accessibility to the accuracy of the structure alignments, CLICK runs with and without these features were compared (Supplementary Figure S1 in Supplementary Data). CLICK_C^{α} used only the coordinates of C^{α} atoms without using secondary structure and solvent accessibility information of amino acid residues. In over 40% of the alignments (3899 out of 9537), using the secondary structure and accessibility features improved the alignment in terms of SO. In only ~7% of the cases (689 out of 9537) did the exclusion of these structural features result in better alignments. Of these 689 cases, the improvements in SO were <5% in 663 alignments. Clearly, the inclusion of secondary structure and solvent accessibility features improved alignment accuracy.

Comparing CLICK with other methods

Protein structure comparisons. CLICK was compared to other structure alignment methods over three data sets: (i) 9537 pair-wise alignments of the HOMSTRAD data set (Table 4 and Supplementary Figure S2a–c in Supplementary Data); (ii) 64 pair-wise alignments of the difficult HOMSTRAD data set (Table 5 and Figure 4); and (iii) 199 pair-wise alignments of structurally similar but topologically different proteins (Table 6 and

Table 4. The comparison of performances of CLICK and MUSTANG, SALIGN on 9537 pair-wise alignments of HOMSTRAD database using average structure overlap (SO) and root mean square deviation (RMSD) scores

	CLICK	MUSTANG	SALIGN
Average SO (%)	86.3	80.5	85.7
RMSD (Å)	1.50	1.52	1.52

	Number of alignment with better CLICK SO values	Number of alignments where CLICK better by 5% SO	Statistical significance of difference (*P*-value)
MUSTANG	6854	3454	Yes ($<10^{-4}$)
SALIGN	4073	1298	No (0.94)

Table 5. The comparison of performances of CLICK and MUSTANG, Geometric Hashing, DALI, SALIGN, GANGSTA$^+$ and FATCAT on the 64 difficult pair-wise alignments

	CLICK	MUSTANG	Geometric Hashing	DALI	SALIGN	GANGSTA$^+$	FATCAT
Average SO (%)	68.9	49.4	59.5	63.0	67.2	61.9	59.1
RMSD (Å)	1.96	2.30	1.91	2.00	2.02	1.99	2.36

	Number of alignment with better CLICK SO values	Number of alignment where CLICK better by 5% SO	Statistical significance of difference (*P*-value)
MUSTANG	56	47	Yes ($<10^{-4}$)
Geometric Hashing	60	50	Yes ($<10^{-4}$)
DALI	45	23	Yes ($<10^{-4}$)
SALIGN	27	15	No (0.72)
GANGSTA$^+$	42	32	Yes ($<10^{-4}$)
FATCAT	48	38	Yes ($<10^{-4}$)

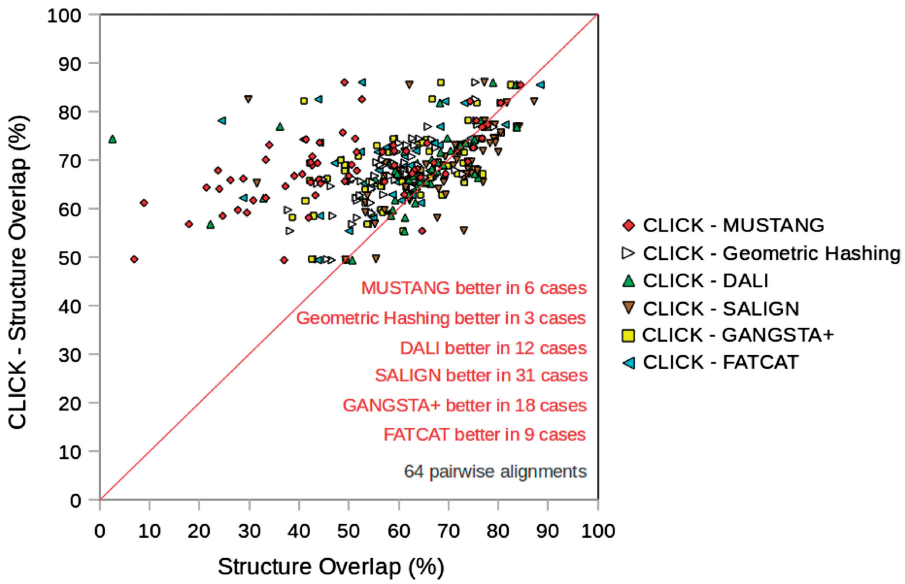


Figure 4. The comparison of the structure overlap values of CLICK against MUSTANG (red diamond), Geometric Hashing (white triangle), DALI (green triangle), SALIGN (brown triangle), GANGSTA+ (yellow square) and FATCAT (cyan triangle) over the 64 difficult pair-wise alignments from the HOMSTRAD database.

Figure 5). For a detailed description of the constituents of each data set refer to the 'Materials and Methods' section.

With the exception of SALIGN, the SO values of CLICK alignments are statistically significantly better than those of MUSTANG, Geometric Hashing, DALI, GANGSTA+ and FATCAT over the difficult HOMSTRAD data sets, and significantly better than those of MUSTANG, Geometric Hashing, DALI,

Table 6. The comparison of performances of CLICK and MUSTANG, Geometric Hashing, DALI, GANGSTA$^+$ on the 199 topologically different pair-wise alignments

	CLICK	MUSTANG	Geometric Hashing	DALI	GANGSTA$^+$
Average SO (%)	68.9	19.8	61.3	60.9	61.7
RMSD (Å)	1.90	3.2	1.86	3.50	2.74

	Number of alignment with better CLICK SO values	Number of alignment where CLICK better by 5% SO	Statistical significance of difference (*P*-value)
MUSTANG	197	190	Yes ($<10^{-4}$)
Geometric Hashing	157	91	Yes ($<10^{-4}$)
DALI	150	107	Yes ($<10^{-4}$)
GANGSTA$^+$	155	106	Yes ($<10^{-4}$)

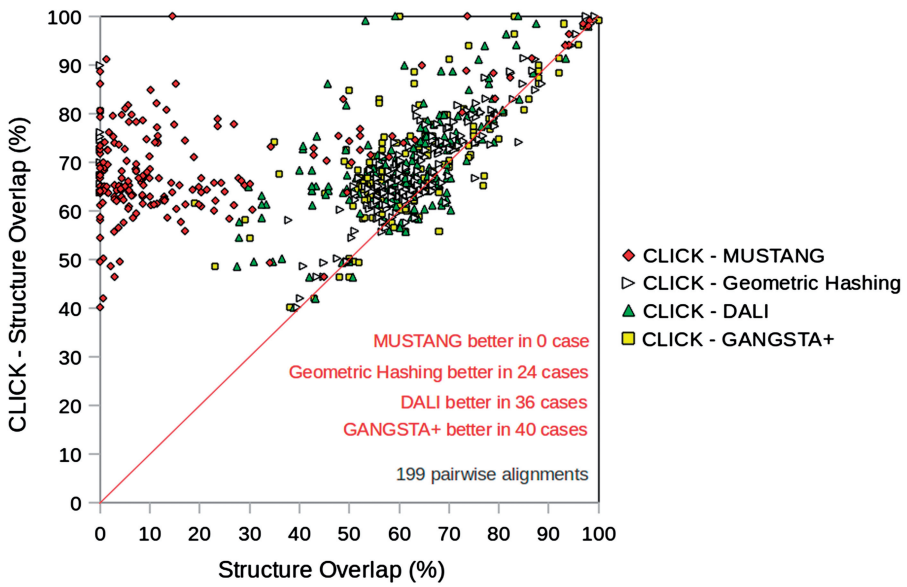


Figure 5. The comparison of the structure overlap values of CLICK against MUSTANG (red diamond), Geometric Hashing (white triangle), DALI (green triangle) and GANGSTA$^+$ (yellow square) over 199 pair-wise alignments that are structurally similar but topologically different.

GANGSTA+ over the set of 199 topologically different alignments. In all of these data sets, the SO from CLICK alignments is never below 40% (Supplementary Figure S2a–c, Figures 4 and 5). In the case of the HOMSTRAD alignments, the CLICK alignments follow the topology of the aligned structures in 9442 of the 9537 cases. In the 95 cases where the topology is not maintained in the CLICK alignment (and also fragment score <1), the infringements are minor—<26 residues on average. Since CLICK maximizes SO values, there are several pairs whose structural similarities but different topology have been detected in HOMSTRAD. For instance, the thermoacidophilic archaeal ferredoxins (PDB code 1xer) (53) is aligned with the cytochrome c553–ferredoxin complex (PDB code 1dwl:A) (54) with an SO of 79.7%, RMSD of 1.84 Å, and fragment and topology scores of 0.94 and 0.51, respectively. Were CLICK to align the two proteins by following the topology of the two structures, which are similar, it would result in a lower SO score and poorer RMSD value. The SO and RMSD values for MUSTANG that follows the chain topology in this alignment are 57.6% and 2.28 Å. The statistical differences in the SO values holds even when these 95 cases that have topology score <1 are not taken into consideration. Also, this statistical difference was preserved when the SO values were computed at different values (1–4 Å) of RMSD (Supplementary Figures S3 and S4 in Supplementary Data).

RNA structure comparisons. To showcase the ability of CLICK to align various different kinds of molecules, we tested its applicability in aligning the 3D structures of RNA pairs. 1275 pair-wise alignments of 51 RNA were chosen from SARA (47,48). We estimated a distance threshold of 15 Å. The C3′ atom was chosen to represent each of the nucleotide residues. The accuracy of CLICK alignments in terms of SO was compared with that of SARA and ARTS (45,46). Over here the RMSD cut-off for SO was taken as 4.0 Å, as used in the SARA server (47). Of the 1275 alignments, CLICK performed better

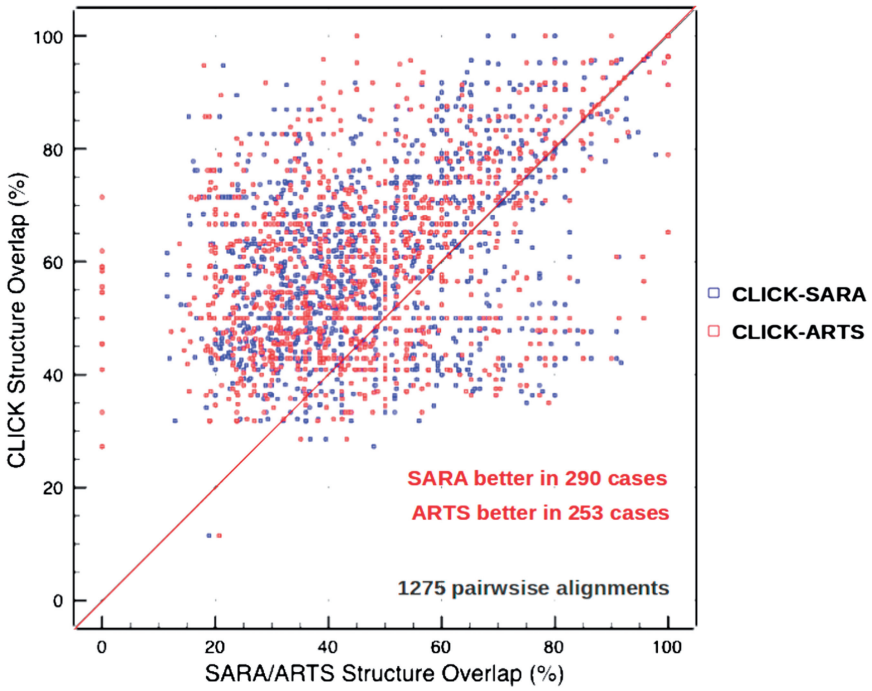


Figure 6. The comparison of the structure overlap values of CLICK against ARTS (red squares) and SARA (blue squares) over 1275 pair-wise alignments of 51 RNA structures.

Table 7. The comparison of performances of CLICK and ARTS, SARA on the 1275 RNA pair-wise alignments

	CLICK	ARTS	SARA
Average SO (%)	60.6	48.1	48.7

	Number of alignment with better CLICK SO values	Number of alignment where CLICK better by 5% SO	Statistical significance of difference (*P*-value)
ARTS	997	876	Yes ($<10^{-4}$)
SARA	970	860	Yes ($<10^{-4}$)

than SARA and ARTS in 970 and 997 cases, and worse than SARA and ARTS in 290 and 178 cases, respectively (Figure 6 and Table 7). While CLICK alignments are statistically significantly better (in terms of SO), most of the cases where SARA performs better than CLICK happen when the SO values are large. On average the SO values obtained from CLICK, SARA, and ARTS are 60.6, 48.7 and 48.1%, with standard deviations (SD) of 16.1, 18.6 and 19.2, respectively.

Detecting alternate conformations—flexible alignment

A useful feature of CLICK is that it can produce more than one alignment between a pair of proteins. The best matching fragments (according to SO) of the aligned pair are first matched. The subsequent matches are between fragments that were not previously aligned. This feature can be utilized to characterize conformational changes in protein structure (see 'Materials and Methods' section). CLICK is used to detect such conformational changes at the level of domains or even sub-domains. Twenty-two pairs of the same protein in different conformations were chosen from the Hinge Atlas database (51), DNA Polymerase Beta proteins, and maltodextrin binding proteins. In each case, CLICK produced multiple alignments for each pair of conformers, corresponding to the number of domains that exhibited conformational change with respect to one another (Supplementary Table S2 in Supplementary Data). CLICK and FATCAT perform comparably over this set (Table 8). One example of such flexible alignments is between the conformers of rat DNA polymerase beta (55) and human DNA polymerase beta (56) (pdb codes 2bpf and 2fmq) (Figure 7a and b). While such large domain motions are easy to detect, smaller conformational changes are harder to spot. Consider the alignment between two metallothioneins-2 of human (57) and rat liver (58) (PDB codes 2mhu and 2mrt,

Table 8. The comparison of performances of CLICK and FATCAT on 22 pairs of the same protein in different conformations from the Hinge Atlas database, DNA Polymerase Beta proteins and maltodextrin binding proteins

	CLICK	FATCAT
Average SO (%)	92.8	86.3
RMSD (Å)	1.31	2.05

	Number of alignment with better CLICK SO values	Number of alignment where CLICK better by 5% SO	Statistical significance of difference (*P*-value)
FATCAT	11	4	No (0.69)

respectively). The MUSTANG alignment matches the whole length of both proteins (Figure 8). CLICK however produces two alignments, implying a conformational change (Figure 9a and b). The hinge regions of the conformational changes captured by these alignments is at the C- and N-termini of the complementary alignments.

Ligand binding site similarity across protein folds

In our effort to unearth cases of convergent structural evolution, we have detected several instances of ATP binding sites that look geometrically similar between proteins belonging to different fold families (data not shown). An example would be the CLICK superpositions for purt-encoded glycinamide ribonucleotide transformylase complexed with Mg-ATP (59) (PDB code 1kj9:B and SCOP entry: b.84.2.1) and *A.fulgidus* rio2 kinase complexed with ATP and manganese ions (60) (PDB code: 1zao:A and SCOP entry: a.4.5.56) (Figure 10a). Though these two proteins belong to different SCOP classes and have different global topologies, their bound ATPs and ATP binding atoms are spatially superimposed to within 1.5 Å (Figure 10b). The base and sugar of the bound ATPs superimpose to within 0.5 Å. No knowledge of the bound ATPs or their binding sites was used in the computation of equivalent residues. The sequence alignment resulting from the 3D superimposition of the two structures shows that three of the 15 residues constituting the binding site are identical, while seven others are similar in the two structures (Figure 10b). The binding of ATP to the active site results in 12 hydrogen bonds, of which in eight cases the same or similar atoms are involved as hydrogen bond donors and acceptors.

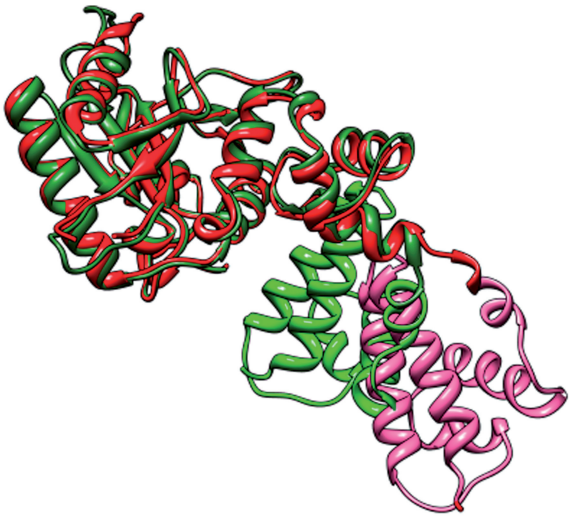
Detecting binding sequence motifs from different topologies

Another application of CLICK was to again detect similar ligand binding sites in topologically distinct proteins, only in this case the binding site comprised of a well-studied sequence motif, the Walker motif (61,62). Using the structure of the multidrug ABC transporter Sav1866 from *Staphylococcus aureus* (63) (pdb 2onj:A) as query, which had the conserved Walker A and Walker B motifs, we searched the PDB for topologically different regions of proteins that matched these motifs structurally. Searching over a data set of 17 712 non-redundant representatives of the PDB (resolution <3 Å, *R*-factor = 0.3 or better, and excluding non crystallographic and C^{α}-only entries), 19 hits were recovered with a topologically different Walker B motif (Supplementary Table S3 in Supplementary Data) and RMSD <3 Å. The Walker B motif is a β-strand with the sequence φφφφD, where φ represents any one of the hydrophobic residues. In the topologically different yet structurally similar hits, the direction of the β-strand is reversed. In all of the 19 hits the Aspartic acid was always conserved and within 8.3 Å from the Walker A motif with a SD of 0.29 Å in C^{α}–C^{α} distance (Figure 11). In the Walker motif that follows the sequential topology, the corresponding C^{α}–C^{α} distance is on average 7.1 Å with a SD of 0.32 Å, as computed from 13 structures. In 18 of the 19 'reverse' Walker B hits, nucleotides (ATP, GDP, GTP, GNP) are bound to the structures and directly interacting with the Walker A motif. In all of these 18 cases, the Aspartic acid from the reverse walker motif interacts indirectly with the nucleotide though a metal ion or via water-mediated hydrogen bonds. It appears that the Aspartic acid plays the same role both the Walker B motif and its reverse. Further, we used the sequence of the reverse Walker B motifs in PSI-BLAST (64) searches to check for sequence conservation. The Aspartic acid is absolutely conserved among all homologues in both the regular and reverse Walker B motifs. The residues that precede the Aspartic acid in the Walker B motif are hydrophobic, on average, 980 times out of 1000; while the residues that follow the conserved Aspartic acid (in the reverse motif) were hydrophobic, on average, only six times out of 100. In the Walker B motif, the hydrophobic residues that precede the Aspartic acid have no direct role in interacting with the nucleotide ligand. Their high degree of conservation is probably not directly related to nucleotide binding. The presence of non-hydrophobic equivalents in the reverse Walker B motif also probably has no direct effect on nucleotide binding. These findings suggest that it is possible for non-sequential versions of motifs to perform the same role as their sequential counterparts. The reverse Walker B motif qualifies as yet another example of convergent evolution in a ligand-binding site.

DISCUSSION

In this study we formalize the protein superimposition problem as one of comparing two sets of points in 3D space. Each of the points is given attributes, such as its

(**a**)



```
2bpf   91:DDTSSSINFLTRVTGIGPSAARKLVDEGIKTLEDLRKNEDKLNHHQRIGLKYFEDFEKRIPREEMLQMQDIVL
NEVKKLDPEYIATVCGSFRRGAESSGDMDVLLTHPNFTSESSKQPKLLHRVVEQLQKVRFITDTLSKGETKFMGVCQLPSEN-
---YPHRRIDIRLIPKDQYYCGVLYFTGSDIFNKNMRAHALEKGFTINEYTIRPLGVTGVAGEPLPVDSEQDIFDYIQWRYRE
PKDRSE:335
2fmq   91:DDTSSSINFLTRVSGIGPSAARKFVDEGIKTLEDLRKNEDKLNHHQRIGLKYFGDFEKRIPREEMLQMQDIVL
NEVKKVDSEYIATVCGSFRRGAESSGDMDVLLTHPSFTSESTKQPKLLHQVVEQLQKVHFITDTLSKGETKFMGVCQLPSKND
EKEYPHRRIDIRLIPKDQYYCGVLYFTGSDIFNKNMRAHALEKGFTINEYTIRPLGVTGVAGEPLPVDSEKDIFDYIQWKYRE
PKDRSE:335
```

(**b**)

```
2bpf   10:TLNGGITDMLVELANFEKNVSQAIHKYNAYRKAASVIAKYPHKIKSGAEAKKLPGVGTKIAEKIDEFLATGKL
RKLEKI:88
2fmq   10:TLNGGITDMLTELANFEKNVSQAIHKYNAYRKAASVIAKYPHKIKSGAEAKKLPGVGTKIAEKIDEFLATGKL
RKLEKI:88
```

Figure 7. Complementary CLICK alignments that maximize chain coverage. Two structures of the DNA polymerase beta protein from Rat and Human, PDB codes 2bpf (red/pink) and 2fmq (bright/dark green), respectively, are superimposed against one another. (**a**) The red and dark green domains, spanning residues 91–335 are aligned with one another. Because of a conformational change (domain motion about a hinge), the pink and bright green domains do not align structurally. (**b**) An alignment of the pink and bright green domains. The corresponding sequence alignments are shown below the ribbon diagrams.

Cartesian coordinates, secondary structure, the accessible area of the amino acid it represents, etc. In principle, every point can be bestowed with many different attributes, as long as we lay down the rules for matching such points. For instance, Cartesian coordinates are matched by a many-body least squares fit and are only considered if the over all RMSD is within a certain threshold and transitions are only possible between certain secondary structure and solvent accessible classes. The resulting protein structure alignments detect structural similarity by

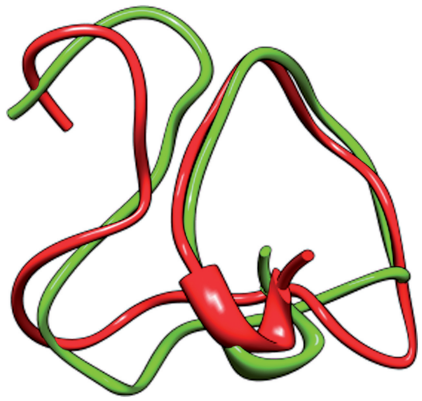


Figure 8. The MUSTANG alignment of two metallothioneins, PDB codes 2mhu (red) and 2mrt (green). The MUSTANG alignment matches the whole length of both proteins.

matching local residue packing and are independent of protein topology.

We have shown that our method, CLICK, performs at par or slightly better than other methods over three sets of benchmarks. The benchmarks include 9537 pair-wise alignments implied by multiple structural alignments of HOMSTRAD families, 64 of which form another set of so called difficult cases where the structural relationships are distant. Also included in the benchmarks is a set of 199 pair-wise alignments between proteins that are structurally similar but topologically distinct. Variously, over these test sets, CLICK is compared against MUSTANG, DALI, Geometric Hashing, GANGSTA^{+}, FATCAT and SALIGN. CLICK produces statistically better results, in terms of structure overlap, than all other methods, except SALIGN.

The superior performance of CLICK over the other data sets was not an exercise to show the superiority of one method over another. We merely wanted to test CLICK over conventional data sets of alignments, to ensure fidelity of alignments. Since we optimized the parameters of CLICK to perform better at comparisons of the 64 difficult HOMSTRAD pairs, we are not surprised that CLICK outperforms other methods on this data set. The performance of CLICK over the rest of HOMSTRAD and the topologically different pairs, gives us confidence to use it to extract structural similarities that are not obvious from sequence or sequential structural comparisons.

We have showcased the utility of CLICK at obtaining biological insights with three different examples. First, we showed that CLICK aligns pairs of protein structures, even when conformational changes alter the positioning of domains or sub-domain with respect to one another. The alignments can be used to detect the locations of hinges in the protein around which these domain/sub-domains rearrange. CLICK alignments are sensitive enough to recognize conformational changes even in sub-domains which as small as 30 residues in length. We propose that CLICK could be an integral part of molecular dynamics trajectory analysis tools to detect

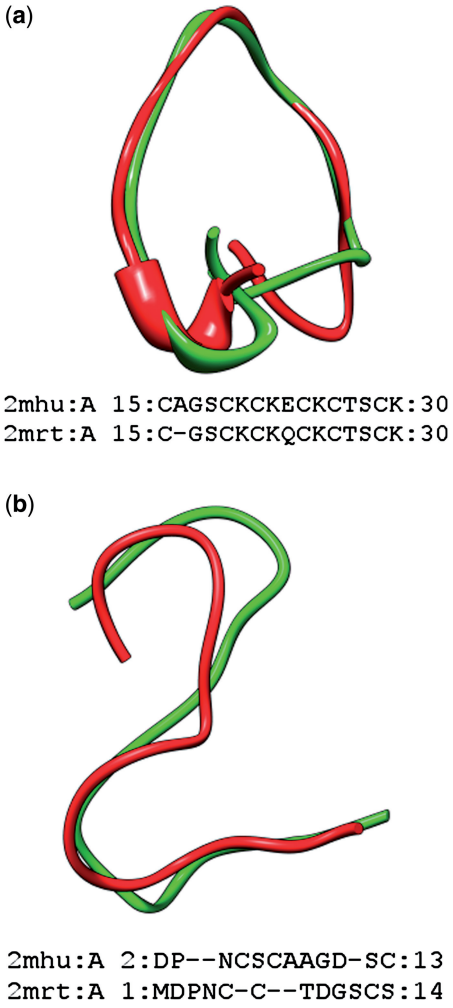
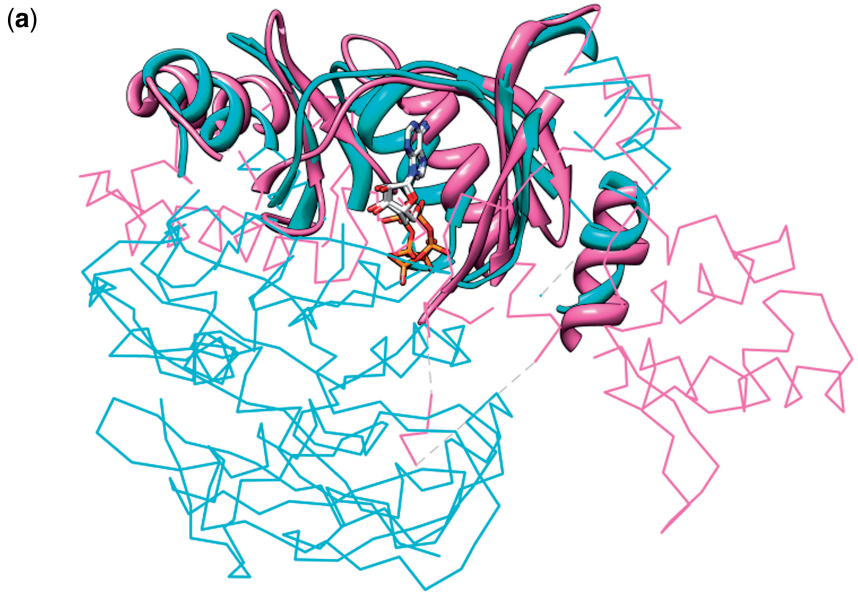


Figure 9. CLICK fragments the alignment between the two metallothioneins, PDB codes 2mhu (red), and 2mrt (green) into two, to account for a small conformational change. The complementary CLICK alignments are shown in (**a** and **b**). The sequence alignment implied by the structural alignment is displayed under the ribbon diagram.

conformational changes (corresponding to low-frequency modes) during the course of protein dynamics.

Second, we have shown that our method is ideally suited to look for similarities in binding site sub-structures. We demonstrated this by aligning two ATP bound proteins with one another. The alignment perfectly matched the binding site residues, and the bound ATP. We have since used this approach to construct a library of ATP binding site geometries, as defined by the atoms of the binding site residues. Such ligand-binding site geometry libraries could prove very useful in constructing models of proteins that are known to, or suspected to bind ligands. The conformation of residues identified as part of the binding site can be refined according to this library when transitioning from apo to holo structural forms. We believe that such examples when sytematized would significantly change our perception of evolutionary relationships between proteins.




```
1kj9:B 112:M--NREGIRRLAAEELQLPTSTYRF:134 146:ADIGYPCIVKPVMS:159 160:SSGKGQTFI:168
1zao:A 150:SARNEF---A---K---LAVPKVYA:171 113:KF-GE-CVVK---H:122 103:ESA---FNC:109

1kj9:B 178:W--KYAQQ--G:184 192:V----IVEGVVKFDFEI:204 238:QM-SP-LALERAQE--I:250
1zao:A  82:LSYLS--LGIF:74  172:WEGNAVLMELIDA-KEL:187 191:RVENPDEVLDMI-LEEV:206

1kj9:B 266:V---ELFVCGDEVIFSE:279
1zao:A 219:LSQYNVLVSE-GIWIID:235
```

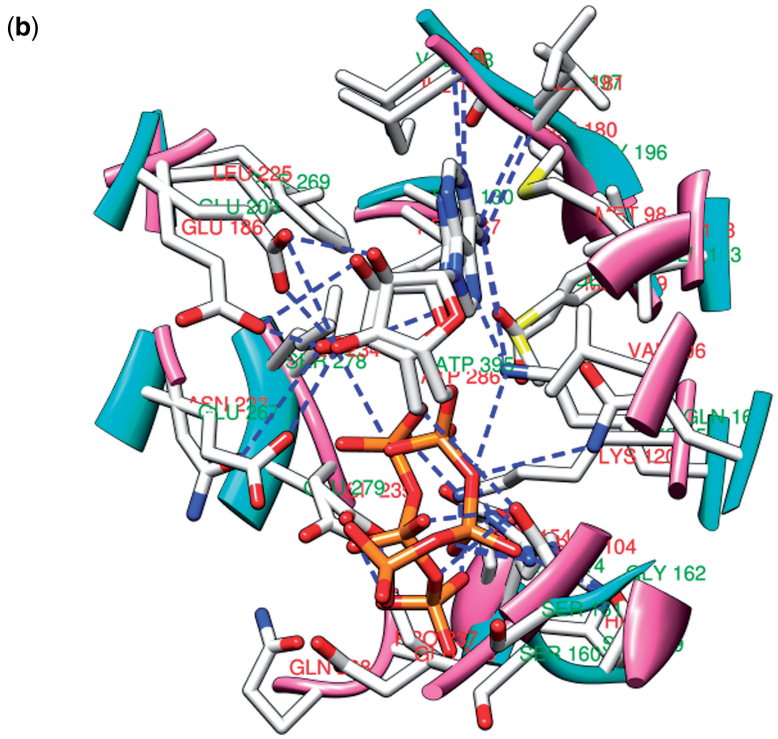


Figure 10. (**a**) CLICK superimposition of two ATP binding proteins 1kj9:B (cyan ribbon & SCOP entry: b.84.2.1) and 1zao:A (magenta ribbon and SCOP entry: a.4.5.56). The regions that superimpose are shown in ribbon representation while the unmatched regions are shown in trace representation. The location of the ATP molecules bound to both structures is shown in stick representation. The sequence alignment between the two proteins as inferred from the structural alignment is show below. The superimposed residues that are in contact with the ATP (within 4 Å) are shown in bold lettering, and the conserved residues in red. (**b**) Representation of the superimposition of the residues in contact (within 4 Å) with the ATP. Residue side chains and the bound ATP molecules are represented as stick and color coded by atom type. Hydrogen bonds are shown as blue dotted lines.

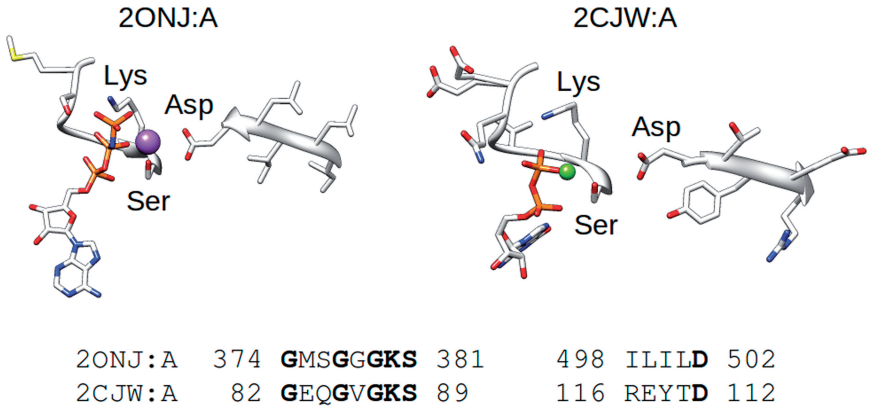


Figure 11. Topologically different Walker B motifs. The walker motif of multi-drug ABC transporter from *S. aureus*, PDB code 2onj:A is superimposed against the equivalent (topologically different) structure of the gem GTPase, PDB code 2cjw:A. The bound nucleotides and the side chains of the residues in the Walker A and B motifs are represented as sticks. The metal ions coordinating the binding of the nucleotide in the two structures are shown as spheres—Mg^{++} (green) and Na^{+} (purple).

Lastly, we have identified variants of sequence motifs that mediate ligand binding. In 19 PDB structures, we show the existence of a 'reverse' Walker B motif, where the beta strand that defines the motif has opposite N–C directionality. Here, we make the case that several such known sequence motifs that define binding sites could have topological variants which will evade sequence search detection. CLICK is ideally designed to identify such variants.

The CLICK program is generally designed to compare any two sets of points. In this study we chose to mainly showcase the utility of CLICK in aligning protein 3D structures. CLICK can just as easily align other biomolecules with one another, as exemplified by the RNA structure comparison benchmark. We hope to use CLICK to extract biological insights from various comparisons of the 3D structures of DNAs, RNAs, DNA–protein complexes, etc. The web server of the program can accommodate searches that compare biomolecules other than proteins. We hope to develop this algorithm to make macromolecular comparisons across different length scales.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors wish to thank Dr Marc A. Marti-Renom for sharing SARA data; Dr Chandra Shekar Verma, Dr Raghavan Varadarajan and Kuan Pern Tan for valuable comments and insights. Thanks are also due to various members of the Biomolecular Molecular Simulation and Design division of the Bioinformatics Institute for their feedback.

FUNDING

Funding for open access charge: Biomedical Research Council (A*STAR), Singapore.

Conflict of interest statement. None declared.

REFERENCES

1. Perutz,M.F., Rossmann,M.G., Cullis,A.F., Muirhead,H., Will,G. and North,A.C. (1960) Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-A. resolution, obtained by X-ray analysis. *Nature*, **185**, 416–422.
2. Kolodny,R., Koehl,P. and Levitt,M. (2005) Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. *J. Mol. Biol.*, **346**, 1173–1188.
3. 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.
4. Cuff,A.L., Sillitoe,I., Lewis,T., Redfern,O.C., Garratt,R., Thornton,J. and Orengo,C.A. (2009) The CATH classification revisited–architectures reviewed and new ways to characterize structural divergence in superfamilies. *Nucleic Acids Res.*, **37**, D310–D314.
5. Stebbings,L.A. and Mizuguchi,K. (2004) HOMSTRAD: recent developments of the Homologous Protein Structure Alignment Database. *Nucleic Acids Res.*, **32**, D203–D207.
6. Marti-Renom,M.A., Pieper,U., Madhusudhan,M.S., Rossi,A., Eswar,N., Davis,F.P., Al-Shahrour,F., Dopazo,J. and Sali,A. (2007) DBAli tools: mining the protein structure space. *Nucleic Acids Res.*, **35**, W393–W397.
7. 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.
8. Friedberg,I., Harder,T., Kolodny,R., Sitbon,E., Li,Z. and Godzik,A. (2007) Using an alignment of fragment strings for comparing protein structures. *Bioinformatics*, **23**, e219–e224.
9. Csaba,G., Birzele,F. and Zimmer,R. (2008) Protein structure alignment considering phenotypic plasticity. *Bioinformatics*, **24**, i98–i104.
10. Leslin,C.M., Abyzov,A. and Ilyin,V.A. (2007) TOPOFIT-DB, a database of protein structural alignments based on the TOPOFIT method. *Nucleic Acids Res.*, **35**, D317–D321.
11. Konagurthu,A.S., Stuckey,P.J. and Lesk,A.M. (2008) Structural search and retrieval using a tableau representation of protein folding patterns. *Bioinformatics*, **24**, 645–651.
12. Abyzov,A. and Ilyin,V.A. (2007) A comprehensive analysis of non-sequential alignments between all protein structures. *BMC Struct. Biol.*, **7**, 78.
13. Konagurthu,A.S., Whisstock,J.C., Stuckey,P.J. and Lesk,A.M. (2006) MUSTANG: a multiple structural alignment algorithm. *Proteins*, **64**, 559–574.

14. Veeramalai,M., Ye,Y. and Godzik,A. (2008) TOPS++FATCAT: fast flexible structural alignment using constraints derived from TOPS+ Strings Model. *BMC Bioinformatics*, **9**, 358.
15. Holm,L., Kaariainen,S., Rosenstrom,P. and Schenkel,A. (2008) Searching protein structure databases with DaliLite v.3. *Bioinformatics*, **24**, 2780–2781.
16. Andreeva,A., Prlic,A., Hubbard,T.J. and Murzin,A.G. (2007) SISYPHUS–structural alignments for proteins with non-trivial relationships. *Nucleic Acids Res.*, **35**, D253–D259.
17. 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.
18. Ye,Y. and Godzik,A. (2004) FATCAT: a web server for flexible structure comparison and structure similarity searching. *Nucleic Acids Res.*, **32**, W582–W585.
19. Madhusudhan,M.S., Webb,B.M., Marti-Renom,M.A., Eswar,N. and Sali,A. (2009) Alignment of multiple protein structures based on sequence and structure features. *Protein Eng. Des. Sel.*, **22**, 569–574.
20. Hasegawa,H. and Holm,L. (2009) Advances and pitfalls of protein structural alignment. *Curr. Opin. Struct. Biol.*, **19**, 341–348.
21. Ilyin,V.A., Abyzov,A. and Leslin,C.M. (2004) Structural alignment of proteins by a novel TOPOFIT method, as a superimposition of common volumes at a topomax point. *Protein Sci.*, **13**, 1865–1874.
22. Holm,L. and Sander,C. (1993) Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, **233**, 123–138.
23. Dror,O., Benyamini,H., Nussinov,R. and Wolfson,H. (2003) MASS: multiple structural alignment by secondary structures. *Bioinformatics*, **19(Suppl. 1)**, i95–i104.
24. Nussinov,R. and Wolfson,H.J. (1991) Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques. *Proc. Natl Acad. Sci. USA*, **88**, 10495–10499.
25. Bachar,O., Fischer,D., Nussinov,R. and Wolfson,H. (1993) A computer vision based technique for 3-D sequence-independent structural comparison of proteins. *Protein Eng.*, **6**, 279–288.
26. Pascual-Garcia,A., Abia,D., Ortiz,A.R. and Bastolla,U. (2009) Cross-over between discrete and continuous protein structure space: insights into automatic classification and networks of protein structures. *PLoS Comput. Biol.*, **5**, e1000331.
27. Grishin,N.V. (2001) Fold change in evolution of protein structures. *J. Struct. Biol.*, **134**, 167–185.
28. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
29. Andreeva,A., Howorth,D., Chandonia,J.M., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2008) Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.*, **36**, D419–D425.
30. 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.
31. Lesk,A.M. and Chothia,C. (1980) How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *J. Mol. Biol.*, **136**, 225–270.
32. Stuart,A.C., Ilyin,V.A. and Sali,A. (2002) LigBase: a database of families of aligned ligand binding sites in known protein sequences and structures. *Bioinformatics*, **18**, 200–201.
33. Tsai,C.J., Lin,S.L., Wolfson,H.J. and Nussinov,R. (1996) A dataset of protein-protein interfaces generated with a sequence-order-independent comparison technique. *J. Mol. Biol.*, **260**, 604–620.
34. Keskin,O., Tsai,C.J., Wolfson,H. and Nussinov,R. (2004) A new, structurally nonredundant, diverse data set of protein-protein interfaces and its implications. *Protein Sci.*, **13**, 1043–1055.
35. Tuncbag,N., Gursoy,A., Guney,E., Nussinov,R. and Keskin,O. (2008) Architectures and functional coverage of protein-protein interfaces. *J. Mol. Biol.*, **381**, 785–802.
36. Sali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 779–815.
37. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
38. Richmond,T.J. and Richards,F.M. (1978) Packing of alpha-helices: geometrical constraints and contact areas. *J. Mol. Biol.*, **119**, 537–555.
39. Kearsley,S.K. (1989) On the orthogonal transformation used for structural comparisons. *Acta Cryst.*, **A45**, 208–210.
40. Koehl,P. (2001) Protein structure similarities. *Curr. Opin. Struct. Biol.*, **11**, 348–353.
41. John,B. and Sali,A. (2003) Comparative protein structure modeling by iterative alignment, model building and model assessment. *Nucleic Acids Res.*, **31**, 3982–3992.
42. Holm,L. and Sander,C. (1996) Mapping the protein universe. *Science*, **273**, 595–603.
43. Holm,L. and Park,J. (2000) DaliLite workbench for protein structure comparison. *Bioinformatics*, **16**, 566–567.
44. Guerler,A. and Knapp,E.W. (2008) Novel protein folds and their nonsequential structural analogs. *Protein Sci.*, **17**, 1374–1382.
45. Dror,O., Nussinov,R. and Wolfson,H. (2005) ARTS: alignment of RNA tertiary structures. *Bioinformatics*, **21(Suppl. 2)**, ii47–ii53.
46. Dror,O., Nussinov,R. and Wolfson,H.J. (2006) The ARTS web server for aligning RNA tertiary structures. *Nucleic Acids Res.*, **34**, W412–W415.
47. Capriotti,E. and Marti-Renom,M.A. (2008) RNA structure alignment by a unit-vector approach. *Bioinformatics*, **24**, i112–i118.
48. Capriotti,E. and Marti-Renom,M.A. (2009) SARA: a server for function annotation of RNA structures. *Nucleic Acids Res.*, **37**, W260–W265.
49. Marti-Renom,M.A., Madhusudhan,M.S., Fiser,A., Rost,B. and Sali,A. (2002) Reliability of assessment of protein structure prediction methods. *Structure*, **10**, 435–440.
50. Liu,Y. and Eisenberg,D. (2002) 3D domain swapping: as domains continue to swap. *Protein Sci.*, **11**, 1285–1299.
51. Flores,S.C., Lu,L.J., Yang,J., Carriero,N. and Gerstein,M.B. (2007) Hinge Atlas: relating protein sequence to sites of structural flexibility. *BMC Bioinformatics*, **8**, 167.
52. Pettersen,E.F., Goddard,T.D., Huang,C.C., Couch,G.S., Greenblatt,D.M., Meng,E.C. and Ferrin,T.E. (2004) UCSF Chimera–a visualization system for exploratory research and analysis. *J. Comput. Chem.*, **25**, 1605–1612.
53. Fujii,T., Hata,Y., Wakagi,T., Tanaka,N. and Oshima,T. (1996) Novel zinc-binding centre in thermoacidophilic archaeal ferredoxins. *Nat. Struct. Biol.*, **3**, 834–837.
54. Morelli,X., Dolla,A., Czjzek,M., Palma,P.N., Blasco,F., Krippahl,L., Moura,J.J. and Guerlesquin,F. (2000) Heteronuclear NMR and soft docking: an experimental approach for a structural model of the cytochrome c553-ferredoxin complex. *Biochemistry*, **39**, 2530–2537.
55. Pelletier,H., Sawaya,M.R., Kumar,A., Wilson,S.H. and Kraut,J. (1994) Structures of ternary complexes of rat DNA polymerase beta, a DNA template-primer, and ddCTP. *Science*, **264**, 1891–1903.
56. Batra,V.K., Beard,W.A., Shock,D.D., Krahn,J.M., Pedersen,L.C. and Wilson,S.H. (2006) Magnesium-induced assembly of a complete DNA polymerase catalytic complex. *Structure*, **14**, 757–766.
57. Messerle,B.A., Schaffer,A., Vasak,M., Kagi,J.H. and Wuthrich,K. (1990) Three-dimensional structure of human [113Cd7]metallothionein-2 in solution determined by nuclear magnetic resonance spectroscopy. *J. Mol .Biol.*, **214**, 765–779.
58. Schultze,P., Worgotter,E., Braun,W., Wagner,G., Vasak,M., Kagi,J.H. and Wuthrich,K. (1988) Conformation of [Cd7]-metallothionein-2 from rat liver in aqueous solution determined by nuclear magnetic resonance spectroscopy. *J. Mol. Biol.*, **203**, 251–268.
59. Thoden,J.B., Firestine,S.M., Benkovic,S.J. and Holden,H.M. (2002) PurT-encoded glycinamide ribonucleotide transformylase. Accommodation of adenosine nucleotide analogs within the active site. *J. Biol. Chem.*, **277**, 23898–23908.
60. LaRonde-LeBlanc,N., Guszczynski,T., Copeland,T. and Wlodawer,A. (2005) Autophosphorylation of Archaeoglobus

fulgidus Rio2 and crystal structures of its nucleotide-metal ion complexes. *FEBS J.*, **272**, 2800–2810.
61. Ramakrishnan,C., Dani,V.S. and Ramasarma,T. (2002) A conformational analysis of Walker motif A [GXXXXGKT (S)] in nucleotide-binding and other proteins. *Protein Eng.*, **15**, 783–798.
62. Rees,D.C., Johnson,E. and Lewinson,O. (2009) ABC transporters: the power to change. *Nat. Rev. Mol. Cell Biol.*, **10**, 218–227.
63. Dawson,R.J. and Locher,K.P. (2007) Structure of the multidrug ABC transporter Sav1866 from *Staphylococcus aureus* in complex with AMP-PNP. *FEBS Lett.*, **581**, 935–938.
64. 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.