PROTEINS: Structure, Function, and Bioinformatics 56:143–156 (2004)

# A Method for Simultaneous Alignment of Multiple Protein Structures

**Maxim Shatsky,[1]* Ruth Nussinov,[2,3] Haim J. Wolfson[1]**
[1]*School of Computer Science, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv, Israel*
[2]*Sackler Institute of Molecular Medicine, Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel*
[3]*Basic Research Program, SAIC-Frederick, Inc. Laboratory of Experimental and Computational Biology, NCI-Frederick Frederick, Maryland*

***ABSTRACT*** **Here, we present MultiProt, a fully automated highly efficient technique to detect multiple structural alignments of protein structures. MultiProt finds the common geometrical cores between input molecules. To date, most methods for multiple alignment start from the pairwise alignment solutions. This may lead to a small overall alignment. In contrast, our method derives multiple alignments from simultaneous superpositions of input molecules. Further, our method does not require that all input molecules participate in the alignment. Actually, it efficiently detects high scoring partial multiple alignments for all possible number of molecules in the input. To demonstrate the power of MultiProt, we provide a number of case studies. First, we demonstrate known multiple alignments of protein structures to illustrate the performance of MultiProt. Next, we present various biological applications. These include: (1) a partial alignment of hinge-bent domains; (2) identification of functional groups of G-proteins; (3) analysis of binding sites; and (4) protein-protein interface alignment. Some applications preserve the sequence order of the residues in the alignment, whereas others are order-independent. It is their residue sequence order-independence that allows application of MultiProt to derive multiple alignments of binding sites and of protein-protein interfaces, making MultiProt an extremely useful structural tool. Proteins 2004;56:143–156.** © 2004 Wiley-Liss, Inc.

**Key words: simultaneous multiple structure alignment; protein structural comparison; structural core; protein interfaces; order independent structural comparison; multiple structure alignment of binding sites**

## INTRODUCTION

The protein structure analysis task has become especially acute with the increase in the number of determined protein structures. It is essential to develop appropriate methods so that structural analysis be as effective as possible. However, currently structural analysis is not as advanced as sequence analysis. For example, methods to characterize a family of protein structures are still largely lacking. In sequence analysis, multiple characteristics of sequence families are utilized in sequence profiles, Hidden Markov Models etc.[1,2] Structural analysis requires approaches in the same spirit. One of the major ingredients for this task is a method for accurate multiple structural alignment. Many methods have been proposed to solve the pairwise structural alignment of protein molecules.[3–7] For comprehensive reviews see Lemmen & Lengauer[8] and Eidhammer et al.[9] However, for a family of structures, an analysis using pairwise structural alignment methods is inadequate. Despite this need, there are only a few methods that address the multiple structure alignment problem.

Let us define the **M**ultiple **ST**ructural **A**lignment (MSTA) problem. There are several points that should be considered for a proper definition:

First, an important aspect of any multiple alignment is a detection of a subset of molecules that are more similar than a whole input set. Consider an example where among 100 input molecules there are 40 structurally similar molecules from family "A," 50 structurally similar molecules from family "B," and 10 additional molecules, which are structurally dissimilar to any other molecule in the input. A multiple alignment which aligns all 100 molecules would probably detect at most one secondary structure element which appears in all 100 structures. Therefore it is very important for a multiple alignment method to be able to recognize two sets "A", "B" from the 100 structures.

Second, there might be only a sub-structure (motif, domain) that is similar between some molecules. Families "A" and "B" may have a common structurally similar motif. Thus, partial similarities between the input molecules should also be reported by a MSTA method.

However, the number of all possible solutions could be exponential in the number of input molecules. For example, consider proteins that contain α-helices. Each pair

Availability: MultiProt is available for download at http://bioinfo3d.cs.tau.ac.il/MultiProt/.



*Correspondence to: Maxim Shatsky, School of Computer Science, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv 69978, Israel. E-mail: maxshats@post.tau.ac.il

Received 24 April 2003; Accepted 24 August 2003

Published online 28 April 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.10628

of α-helices could be structurally aligned (at least partially, if they are different in their lengths). Any combination of α-helices from different molecules gives us some solution to the MSTA problem. The number of such combinations is exponential. Thus, it is not practical to output (even if the algorithm is capable of detecting) all possible solutions. The MSTA problem indeed has been proved to be computationally a hard one even in the one-dimensional space for the case of an exact congruence (zero error matching between the points).[10]

As in the case of multiple sequence alignment, one may apply a *center-star* approach. One, center, molecule is selected, usually the one which is structurally the closest to the rest of the molecules. Then, all other structures are joined into a multiple alignment based on their pairwise alignments with the center structure.[11,12] However, such an approach has some limitations. First, it does not detect small common motifs. A detection of small motifs is analogous to a local alignment, while the center-star approach, in essence, builds a global alignment. Second, the approach does not recognize structurally similar subsets.

Another approach to multiple alignment is to apply a hierarchical combination of sub-multiple alignments. For example, the tree-progressive alignment.[13,14] The structures are iteratively aligned according to a distance tree. Therefore, first the most similar structures are aligned, then the process proceeds to less similar molecules. An advantage of such approaches is their capability to detect sub-set alignments of structurally different families. However, since at each stage of the hierarchical alignment only one, the best, solution is selected, the methods cannot detect small structurally similar motifs. Figure 1(a) shows a simple example where a straightforward application of a pairwise alignment method will fail to recognize a pattern common to more than two sequences/structures.

The *MUSTA* algorithm[15,16] tries to solve the multiple structural alignment problem without using pairwise alignments and dealing with multiple molecules simultaneously. The method is based on the *Geometric Hashing* technique. This technique was successfully applied in docking algorithms[17,18] and in a number of pairwise structure alignment methods.[6,19] One of the strengths of this technique is its efficiency and independence of the order of the structural features representing the molecules. However, there are several disadvantages inherent to the *MUSTA* algorithm. First, the algorithm requires that all input molecules participate in the multiple structural alignment, i.e., the algorithm cannot detect the structural cores between a non-predefined subset of the input molecules. Second, its runtime is practical only for sets of about 10–15 molecules.

A combination of sequence properties and structural information is used in the SPratt method[20] which discovers local packing motifs in a number of protein structures. For each residue the method describes its spatial neighborhood as a string. Then, SPratt applies an efficient sequence pattern discovery method to detect patterns common to subsets of these strings. Therefore, the method is particularly suitable for detection of small structural motifs. The common motifs should follow the backbone order.

Recently, a new method for multiple structural alignment has been proposed.[21,22] Both methods, MASS and our proposed method MultiProt, are capable of detecting structural motifs shared only by a subset of the molecules. The MASS method exploits the secondary structure representation, which aids in filtering out noisy results and in making the method highly efficient. It requires that at least two secondary structure elements (SSE's) be aligned. The method disregards the sequence order of the SSE's. Thus, it can find non-sequential and even non-topological structural motifs.

Here, we propose a method that aims to solve the multiple structural alignment problem with detection of partial solutions. It includes sub-sets detection as well as a detection of small structurally similar motifs. The multiple alignments are achieved by simultaneous structural superposition of input molecules in all possible ways. The only required condition is that at least short contiguous fragments (three amino acids or more) of the backbone chains should be structurally similar. The method computes the best scoring structural alignments, which can be either according to a sequence order, like in sequence alignment, or be sequence-order independent, if one seeks geometric patterns which do not follow the sequence order. Therefore, it is able to detect non-topological similarities and can deal with proteins consisting of several chains.

One of the advantages of our method is illustrated in Figure 1(a, b). Consider the set of proteins from Figure 1(a). The goal of our method is to detect local multiple alignments of all four patterns. This is achieved by performing all possible local multiple alignments of ungapped fragments. The final solutions are concluded from these locally aligned multiple fragments. This makes our approach different from existing ones, which generally derive a multiple alignment from the high scoring pairwise superpositions. If a pattern appears more than once in some protein, our method recognizes only one combination of this pattern from all possible appearances, although all sets of possible combinations are reported by the program.

---

Fig. 1. **a**: Consider three sequences/structures that share a common pattern X. Applying a pairwise alignment method that detects the most similar common pattern will result in pattern A for S1 and S2, pattern B for S1 and S3, and pattern D for S2 and S3. Therefore, no common pattern can be derived from the patterns A, B, and D. One possible solution is to store two (or more) high scoring solutions for each pairwise comparison. However, in this case the number of iterations to compare all pairwise results to detect the best combination of multiple alignments becomes, theoretically, exponential. **b**: The goal of the MultiProt method is to detect the local multiple alignments of all four patterns. Pattern X will appear in the multiple alignment of three molecules. Patterns A, B and D will appear in the set of alignments consisting of two molecules.

Fig. 3. Four-helix bundle. **a**: There are four structures in our study, 1f4n, 2cbl: A, 1b3q, and 1rhg: A. **b**: Structural alignment between 1f4n, 2cbl: A, 1b3q, 1rhg: A. Four-helix bundle is aligned. **c**: A multiple alignment of the sequences according to the multiple structural alignment of the 4-helix bundle. Note that the directions are different. Since proteins 1f4n, 1b3q have two chains, it leads to an additional difficulty for sequence alignment methods in identifying a four-helix bundle.

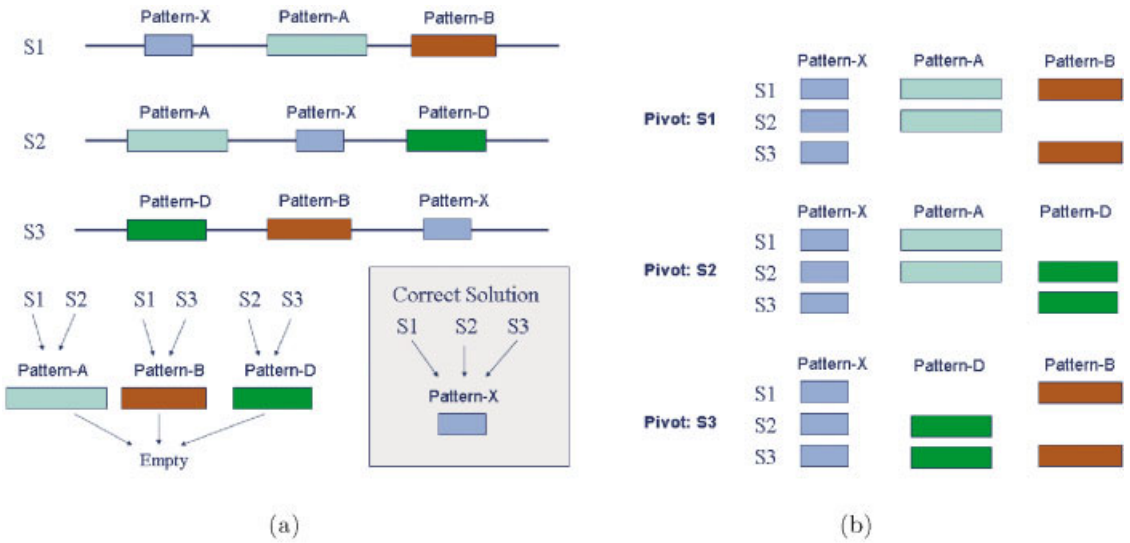



Figure 1

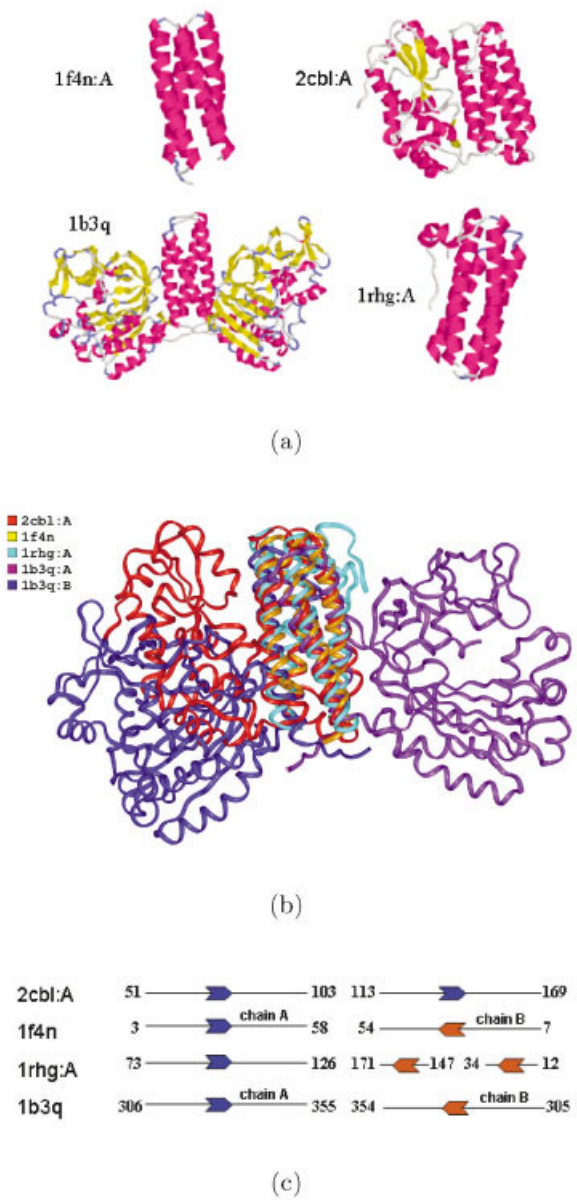



Figure 3

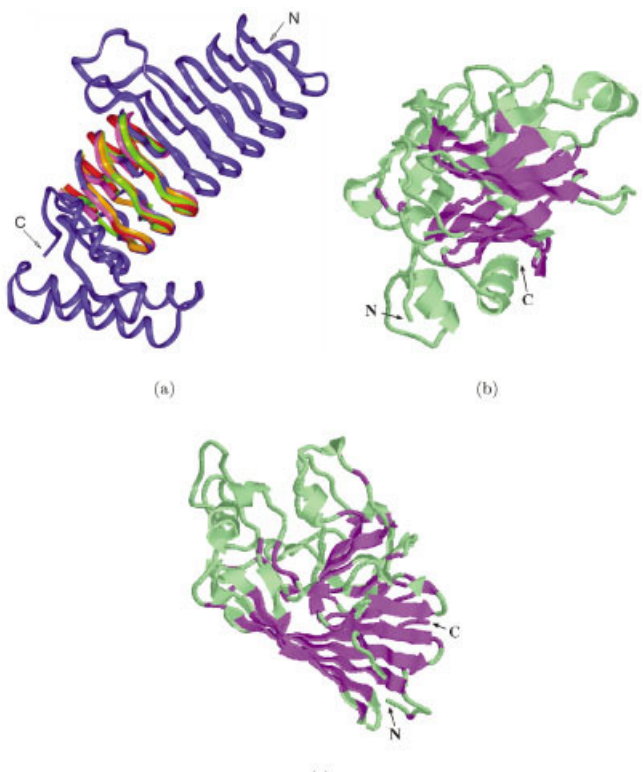



Fig. 4. **a**: Superhelix. This figure shows the structural core between five molecules, [1lxa, 1qq0, 1xat, 2tdt, 1fwy(A: 252–328)]. The complete backbone of 1lxa is shown in blue. For the other molecules only the common detected core is shown, by assigning a different color to each molecule. **b**: *Concanavalin A-like lectins/glucanases*. Structural core of six molecules—1a8d: 1–247, 1d2sA, 1gbg, 1sacA, 2galA, 2sli: 81–276. The *two-sheet sandwich* is conserved. The backbone of molecule 1a8d: 1–247 is shown completely in light green. The aligned core is in purple color. **c**: *Supersandwich*. Multiple structural alignment of proteins 1bgmI: 731–1023, 1cb8A: 336–599 and 1oacA: 301–724. The backbone of 1bgmI: 731–1023 is shown in its entirety in light green. The aligned core is in purple color. Three β-sheets are aligned.

## THE MULTIPROT ALGORITHM

In MultiProt we try to give an efficient heuristic solution to the MSTA problem, which we define as:

(*) *Given* $\boldsymbol{m}$ molecules, a parameter $\kappa$ and a threshold value $\epsilon$, *for each r* ($2 \le r \le m$), *find the* $\kappa$ *largest* $\epsilon$*-congruent multiple alignments containing exactly r molecules.*

Here an $\epsilon$-*congruent multiple alignment* is defined as: given a pivot molecule $M_1$ and $r - 1$ molecules ($M_2, \ldots, M_r$), we define an $\epsilon$-congruent multiple alignment as a set of $r - 1$ 3D transformations ($T_2, \ldots, T_r$) ($T_j$ is a transformation which superimposes molecule $M_j$ onto $M_1$) and a set of K $r$-tuples (aligned points) $\{(v_{i_k^1}, v_{i_k^2}, \ldots v_{i_k^r})\}_{k=1}^{K}$, $v_{i_k^j} \epsilon M_j$ such that, $\forall k \forall i_k^j \| v_{i_k^1} - T_j(v_{i_k^j}) \| \le \epsilon$, i.e., the matched points are within $\epsilon$ distance from the appropriate pivot molecule point.

This definition is based on the selection of the pivot molecule. In our algorithm, in order not to be dependent on the choice of the pivot, we iteratively choose every molecule to be the pivot one.

The above MSTA problem definition (*) is general and can be applied for comparison of any 3-D objects represented as unconnected point sets in 3-D space. However, we wish to utilize the fact that a protein structure can be represented as an ordered set of points, e.g., by the sequence of the centers of the $C_\alpha$ atoms. Thus, we exploit the natural assumption that any solution for the MSTA of proteins should align, at least short, contiguous fragments (minimum 3 points) of input atoms. For example, these fragments could be secondary structure elements which could be aligned between the input molecules. First we detect all possibly aligned fragments of maximal length between the input molecules. Then, we select solutions that give high scoring global structural similarity based on the (*) definition. Aligning protein fragments is not a new idea. It has been previously applied in several methods for pairwise protein structural alignment.[5,7] In our method we use an algorithm which detects structurally similar fragments of maximal length, i.e., fragment pairs which cannot be extended while preserving $\epsilon$-congruence.[19,23]

For any multiple alignment problem we can always assume that the selected pivot molecule has to participate in all the alignments. If all the molecules are iteratively selected as pivots, then all solutions can be detected. Therefore, our method is based on the pivoting technique, i.e., the rest of the molecules are aligned with respect to the pivot molecule.

```
Input: m molecules S = {M_1. . .M_m}
for i = 1 to m − 1
  M_pivot = M_i
  S′ = S \M_pivot
  Alignments = MultipleFragmentAlignment(M_pivot, S′)
  GlobalMultipleAlignment(Alignments)
End
```

The *MultipleFragmentAlignment* stage detects all aligned fragments with the pivot molecule and then detects all possible combinations of structurally similar fragments between two or more molecules (*cuts* detection).

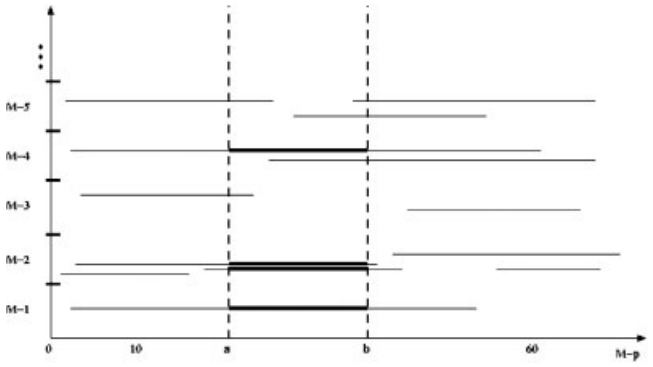



Fig. 2. *Cuts.* The *x*-axis is the sequence of $M_p$. Molecules $M_1, \ldots, M_m$ are assigned into bins on the *y*-axis. Fragment pairs that include completely fragment $[\alpha,\beta]$ (shown in bold) are incorporated into $Cut[\alpha,\beta]$.

The algorithm requires that the pivot molecule participates in the multiple alignments, but it does not require that all input molecules from set $S'$ are included in the multiple alignment. In essence, based on fragment alignment, this stage detects a set of multiple transformations $\{(T_{i^1}, \ldots, T_{i^r})\}$ (multiple transformation $(T_{i^1}, \ldots, T_{i^r})$ aligns molecules $(M_{i^1}, \ldots, M_{i^r})$ with $M_{pivot}$).

The fragment alignment from the *MultipleFragment Alignment* stage defines only 3D transformations, but it does not identify which points (amino acids) are matched in 3D space. It only finds that some, possibly very short—3-amino-acid-long, fragments can be matched. Therefore, once multiple transformations are detected, the goal of the *GlobalMultipleAlignment* stage is to detect the largest structural cores between the aligned molecules. At this stage the sequence order of matched points can be optionally preserved or be sequence-order independent.

Thus, our method detects large partial[1] multiple alignments.

### Stage 1. Multiple Fragment Alignment
### *Detection of all fragment pairs*

Given a pivot molecule $M_p$, for each molecule $M_k$ from the set $S'$ we detect all structurally similar fragment pairs between $M_p$ and $M_k$. Namely, a structurally similar (or $\epsilon$-congruent) fragment pair is defined as $F_i^p F_j^k(l)$ (a fragment starts at point $i$ ($j$) in molecule $M_p$ ($M_k$) and has length $l$). It also satisfies the following condition: $\mathrm{RMSD}_{\mathrm{opt}}(F_i^p F_j^k(l)) \le \epsilon$. $RMSD_{opt}$ is defined as:

$$RMSD_{opt}(F_i^p F_j^k(l)) = \min_T RMSD(F_i^p(l), T(F_j^k(l))),$$

where $T$ is a rigid 3D transformation.

To calculate all $\epsilon$-congruent fragment pairs two options are implemented in the program. One can use an exact algorithm,[2] but in order to achieve a favorable running

[1]For the MSTA problem the term *partial* has two different meanings. First, only a subset of the molecules is aligned. Second, the structural core contains only subsets of the aligned molecule residues.

[2]All $\epsilon$-congruent fragments, $\{F_i^p F_j^k(l)\}$, can be obtained by an exhaustive verification in polynomial time.

**TABLE I. Pairwise Structural Alignment Test**[†]

| Molecule 1 (size) | Molecule 2 (size) | VAST $S_{al}$/rms | Dali $S_{al}$/rms | CE $S_{al}$/rms | GH $S_{al}$/rms | MultiProt$_1$ $S_{al}$/rms | MultiProt$_2$ $S_{al}$/rms |
|---|---|---|---|---|---|---|---|
| 1fxi:A (96) | 1ubq (76) | 48/2.1 | — | — | 51/1.6 | 44/1.7 | 50/1.8 |
| 1ten (89) | 3hhr:B (195) | 78/1.6 | 86/1.9 | 87/1.9 | 81/1.7 | 81/1.3 | 82/1.3 |
| 3hla:B (99) | 2rhe (114) | — | 63/2.5 | 85/3.5 | 62/1.8 | 60/1.8 | 67/1.9 |
| 2aza:A (129) | 1paz (120) | 74/2.2 | — | 85/2.9 | 74/1.9 | 75/2.0 | 85/2.5 |
| 1cew:I (108) | 1mol:A (94) | 71/1.9 | 81/2.3 | 69/1.9 | 66/1.6 | 76/1.8 | 75/1.9 |
| 1cid (177) | 2rhe (114) | 85/2.2 | 95/3.3 | 94/2.7 | 70/1.5 | 84/1.8 | 88/1.9 |
| 1crl (534) | 1ede (310) | — | 211/3.4 | 187/3.2 | 180/1.9 | 161/2.3 | 232/2.4 |
| 2sim (381) | 1nsb:A (390) | 284/3.8 | 286/3.8 | 264/3.0 | 197/2.0 | 233/2.3 | 268/2.3 |
| 1bge:B (159) | 2gmf:A (121) | 74/2.5 | 98/3.5 | 94/4.1 | 72/1.8 | 78/2.5 | 88/2.2 |
| 1tie (166) | 4fgf (124) | 82/1.7 | 108/2.0 | 116/2.9 | 87/1.7 | 95/2.1 | 99/2.3 |

[†]The protein pairs are classified as "difficult" for structural analysis.[26] The alignments are performed by VAST,[27] Dali,[28] CE,[7] GH Geometric Hashing method[29] http://bioinfo3d.cs.ac.il/c_alpha_match/) and MultiProt. The information in this table, except for the Geometric Hashing method and MultiProt results, is taken from Shindyalov and Bourne.[7] MultiProt$_1$ results do preserve the sequence order, while MultiProt$_2$ are sequence order independent. $S_{al}$ is the number of aligned atoms.

time of the program we can apply the same efficient greedy method as in the *FlexProt* algorithm.[19] We start by aligning a single matching atom pair $(v_a, u_b)$, where $v_a \in M_p$ and $u_b \in M_k$. Now, we iteratively try to extend the initial match-list. We do this by adding one matching atom pair to the left and to the right (following the backbone direction) of the current fragment alignment. This is done iteratively, until the RMSD of the fragment alignment exceeds a predefined threshold. That is, we stop when the match list cannot be extended either to the left, nor to the right. Given $F_i^p F_j^k(l)$, the next alignment is initiated at $(v_{i+(l+1)}, u_{j+(l+1)})$. The process can be viewed as proceeding along the diagonals of the 2D matrix, which represents the indices of $M_p$ and $M_k$. Since RMSD can be continuously updated by a constant number (O(1)) of operations at each step, the time complexity of computing $F_i^p F_j^k(l)$ is only $O(l)$ and our greedy iterative approach takes only $O(|M_p| \cdot |M_k|)$. All presented experimental results are computed with the greedy method.

At the end of this step we have a set of congruent fragment pairs, $\{F_i^p F_j^k(l) : k \neq p, \mathrm{RMSD}_{\mathrm{opt}}(F_i^p F_j^k(l)) \leq \epsilon\}$.

***Cut detection*** Let us represent the above set by a 2-dimensional plot. The $x$-axis represents the sequence of the pivot molecule $M_p$. The $y$-axis is divided to bins, one bin for each molecule from set $S'$. Fragment $F_j^k(l)$, from the pair $F_i^p F_j^k(l)$, is plotted in bin $M(k)$ ($y$-axis) and aligned to $F_i^p(l)$, i.e., its projection onto the $x$-axis is exactly the set of the corresponding (according to the $F_i^p F_j^k(l)$ alignment) points of the $F_i^p(l)$. The order of the fragments inside the $M(k)$ bin (on the $y$-axis) is arbitrary (see Fig. 2).

Drawing two vertical lines at points α and β defines a *cut* on the interval [α,β]. A fragment belongs to the *cut* if the interval [α,β] lies inside the fragment, i.e. $F_i^p F_j^k(l)$ is in the *cut* if and only if $i \leq \alpha$ *and* $\beta \leq (i + l - 1)$. Since several fragments from the same molecule might participate in the *cut*, such a *cut* provides us with a set of multiple choices for the alignment. Choosing from the *cut* only one fragment for each molecule gives us some multiple alignment. The number of choices equals $\Pi_i k_{M_i}$ where $k_{M_i}$ is the number of fragments from molecule $M_i$ in the *cut*. Thus, the number of possible multiple alignments (for the given *cut*) might grow exponentially with the number of molecules. This is the nature of the *multiple alignment* problem (there is an exponential number of choices to align α-helices from, for example, all-alpha-proteins). We shall return to this problem later in *Stage 2* after we explain how to detect the *cuts*.

**TABLE II. Structural Classification of Protein Sets Used for the Comparison With the MUSTA Method[16]**

| Superfamily | PDB code |
|---|---|
| TIM-Barrels | |
| *Triosephosphate isomerase* | 7timA |
| *Cellulases* | 1tml |
| *Glycosyltransferases* | 1btc |
| *Enolase C-terminal domain-like* | 4enl |
| *Ribulose-phosphate-binding barrel* | 1pii |
| *Xylose isomerase* | 6xia |
| *RuBisCO, C-terminal domain* | 5rubA |
| Helix-Bundle | |
| *Protein designs* | 1flx |
| *Apolipophorin III* | 1aep |
| *4-helical cytokines* | 1bgeB,1rcb,3inkC |
| *Apoliporotein* | 1le2 |
| *Cytochromes* | 256bA,2ccyA,1bbhA |
| *Hemerythrin* | 2hmzA |

From Figure 2 we can observe that it might be possible to extend a given *cut* to the left and to the right so that the *cut* contains the same fragments. Thus, we define a locally maximal cut $Cut[\alpha,\beta]$ as an interval [α,β] such that for any $\delta > 0$, $Cut[\alpha - \delta,\beta]$ and $Cut[\alpha,\beta + \delta]$ contain different fragments than $Cut[\alpha,\beta]$. It is obvious that any $Cut[\alpha,\beta]$ starts at the beginning of a fragment and ends at the end of a (possibly, another) fragment. From now on, we use the terms *cut* and $Cut[\alpha,\beta]$ interchangeably.

All possible *cuts* can be detected efficiently by a simple algorithm which is similar to the sweeping technique from Computational Geometry.[24] The idea is to represent the projections on the $x$-axis of the fragment start(end)-points as events on the $x$-axis. The fragments left (right) end-point is the *start* (*end*) event. Starting

**TABLE III. Multiple Structural Alignment Test**

| Proteins | Number of Mols | Average Size | MUSTA $S_{al}$ | MultiProt $S_{al}$ | MultiProt run-time |
|---|---|---|---|---|---|
| (a) Comparison with MUSTA[16] algorithm | | | | | |
| *Serpins* | 13 | 372 | 163 | 237 | 9m04s |
| 7apiA, 8apiA, 1hleA, 1ovaA, 2achA, 9apiA, 1psi, 1atu, 1kct, 1athA, 1attA, 1antl, 2antl | | | | | |
| *Serine Proteinase* | 5 | 277 | 220 | 227 | 25s |
| 1cseE, 1sbnE, 1pekE, 3prkE, 3tecE | | | | | |
| *Calcium-binding* | 6 | 140 | 31 | 36 | 9s |
| 4cpv, 2scpA, 2sas, 1top, 1scmB, 3icb | | | | | |
| *TIM-Barrels* | 7 | 391 | 40 | 44 | 3m12s |
| 7timA, 1tml, 1btc, 4enl, 1pii, 6xia, 5rubA | | | | | |
| *Hclix-Bundle* | 10 | 140 | 27 | 27 | 2m10s |
| 1flx, 1aep, 1bbhA, 1bgeB, 1le2, 1rcb, 256bA, 2ccyA, 2hmzA, 3inkC | | | HOMSTRAD $S_{al}$ | | |
| (b) Comparison with the HOMSTRAD data base | | | | | |
| *Calcium-binding* | 7 | 107 | 101 | 101 | 2s |
| 1rtp, 1pvaA, 5cpv, 1pal, 5pal, 1omd, 1a75A | | | | | |
| *Subtilase* | 8 | 274 | 217 | 221 | 48s |
| 1dbiA, 1thm, 1bh6A, 1alyE, 1meeA, 1sup, 1gci, 2prk | | | | | |
| *Serine Proteinase Inhibitor* | 8 | 376 | 270 | 269 | 2m21s |
| 2ach, 1qlpA, 1athA, 1attA, 1hle, 1ovaA, 1a7cA, 1sek | | | | | |
| *Reductases* | 7 | 266 | 131 | 121 | 38s |
| 2cnd, 1ndh, 1que, 1qfzA, 1fdr, 1aSp, 1qfjA | | | | | |
| *TPR domain* | 6 | 153 | 68 | 86 | 6s |
| 1a17, 1elwA, 1elrA, 1c96B, 1fchA, 1ibgA | | | | | |
| *Plant virus coat protein* | 7 | 175 | | | |
| 2tbvA, 4sbvA, 1smvA, 1stmA, 1bmv1, 1cwpA, 2stv | | | | | |
| $\epsilon = 3$Å | | | 33 | 34 | 16s |
| $\epsilon = 4$Å | | | 74 | 76 | 21s |

$S_{al}$ is the number of aligned atoms.

from the first point ($C_\alpha$ atom) of the pivot molecule, $M_p$, we move along the $x$-axis ($M_p$ sequence) with a vertical line and generate *cuts*. At every moment (i.e., position $\alpha$ of the vertical line) we remember the fragments that the vertical line passes through. At every *start* event we add its fragment to the list of current fragments and generate new *cuts*. These new *cuts* start at the encountered *start* event and end at *end* points of the fragments from the current list.

At an *end*-event we just remove the fragment from the current list of fragments. It is easy to see that the described algorithm generates all possible *cuts* according to the definition of $Cut[\alpha,\beta]$.

### Stage 2. Global Multiple Alignment

Consider one of the previously detected *cuts*, $Cut[\alpha,\beta]$. One of the possible approaches is to leave the *cuts* as is, i.e., not to choose the multiple alignment(s) from the set of possible ones of the specific *cut* (several fragments of the same molecule might be included in the *cut*). These complete *cuts* are included in the output of the program. Thus, an end-user can apply additional criteria to filter out the non-relevant fragments (or 3D transformations).

However, our goal is to detect the best multiple alignments based on the global structural similarity, as defined in (*). As we have already pointed out, this is a hard problem,[3] so we provide only a heuristic solution.

In this step, we select from every *cut* only one fragment for each molecule. We aim to perform this selection, in a way that the resulting multiple alignment would give, possibly, the highest score. Namely, given a $Cut[\alpha,\beta]$ containing $\{F_i^p F_j^k(l) : i \le \alpha, (i + l - 1) \ge \beta\}$, for each different $k$ (for each molecule $M_k$) select the fragment (if there is more than one fragment from molecule $M_k$) so that its transformation gives the largest, global, structural (pairwise) alignment with the pivot molecule ($M_p$). Let us explain this step in more detail. Given a fragment pair $F_i^p F_j^k(l)$ and the transformation $T_{opt}$ that optimally superimposes $F_i^p F_j^k(l)$ onto $F_j^k(l)$, we apply the transformation $T_{opt}$ on molecule $M_k$. Now, when the two molecules are aligned, we calculate the size of their maximal structural alignment which preserves $\epsilon$-congruence. For more details see the Appendix in Shatsky et al.[25]

[3]In the first two stages we compute a set of transformations. However, even computing all possible transformations will not reduce the complexity of the MSTA problem. The MSTA problem is NP hard even in the case of exact congruence ($\epsilon = 0$). When $\epsilon = 0$ all possible transformations can be computed in polynomial time (in 3D it is enough to match all possible triplets of points). Therefore to select the correct set of transformations is still a NP hard problem.

At this stage of the algorithm every solution for the MSTA problem has a non-ambiguous representation, i.e., each solution contains at most one representative from each molecule. Now, the task is to score these solutions based on the size of the multiple alignment. Notice, that the transformation for each molecule is now fixed, thus we only need to calculate the *multiple correspondence* size. Computing multiple correspondence is also combinatorially not a simple task. We use very efficient procedure which for small $\epsilon$ produces optimal results. For the details of this procedure see the Appendix in Shatsky et al.[25]

To enlarge the obtained solutions we apply an iterative improvement procedure as follows. For each solution, after the *multiple correspondence* between the pivot molecule with the other molecules is established, we apply a rigid transformation that minimizes the RMSD between the matching points. Then, we compute the *multiple correspondence* once again and repeat this procedure (the default number of iterations is three).

***Solution scoring.*** When a common geometric core is detected, we compute the multiple RMSD (mRMSD) of the alignment. It is computed as an average of the RMSD values between the geometric core of the pivot molecule $M_p$ with the corresponding geometric core of each molecule from the multiple alignment. Thus, solutions are grouped according to the number of aligned molecules and each group is sorted according to the size of the alignment and according to the mRMSD, giving priority to the alignment size.

**TABLE IV. Structural Classification of Studied Proteins by SCOP Database**

| Family | PDB code |
|---|---|
| Superhelix | |
| *UDP N-acetylglucosamine acyltransferase* | 1lxa |
| *Carbonic anhydrase* | 1qq0 |
| *Xenobiotic acetyltransferase* | 1xat |
| *Tetrahydrodipicolinate-N-succinlytransferase, THDP-succinlytransferase, DapD* | 2tdt |
| *N-acetylglucosamine* | 1fwy |
| *1-phosphate uridyltransferase GlmU, C-terminal domain* | A:252–328 |
| Concanavalin A-like lectins/glucanases | |
| *Legume lectins* | 2bqpA |
| *beta-Glucanase-like* | 1gbg |
| *Galectin (animal S-lectin)* | 2galA |
| *Laminin G-like module* | 1d2sA |
| *Pentraxin (pentaxin)* | 1sacA |
| *Clostridium neurotoxins, the second last domain* | 1a8d:1–247 |
| *Vibrio cholerae sialidase, N-terminal and insertion domains* | 1kit:25–216 |
| *Leech intramolecular trans-sialidase, N-terminal domain* | 2sli:81–276 |
| *Endoglucanase/cellulase I catalitic core* | 6cel |
| *Xylanase/endoglucanase 12* | 1xnb |

### *Optimization schemes and bio-core detection*

As described above we treated the problem as a pure geometrical structural alignment. We can apply a somewhat different scoring scheme, which requires that aligned points are of the same biological type (still, the points should be close enough in 3D space). In our method the input points can be either positions of $C_\alpha$ or $C_\beta$ atomic centers, or geometric centers of amino acids, or residue specific points (see description in Results section for the case of G-proteins). Therefore, each point represents a specific amino acid and thus, we can require that only points with similar characteristics be aligned. For instance, we can require residue identity matching, but it is usually too restrictive. Thus, we adopted the following classification: hydrophobic (Ala, Val, Ile, Leu, Met, Cys), polar/charged (Ser, Thr, Pro, Asn, Gln, Lys, Arg, His, Asp, Glu), aromatic (Phe, Tyr, Trp), glycine (Gly). Let us name this the *bio-core* classification.

At Stage-1 the method detects a set of possible multiple transformations, while at Stage-2 the solutions are scored based on the size of the multiple structural alignment. Therefore, we can apply various scoring schemes, like the bio-core classification, to Stage-2 to obtain different solution ranking.

One of the ways used to measure sequence similarity of several proteins is to compute an average sequence identity. However this technique is based on pairwise properties and is dependent on gap penalty parameters. The bio-core classification possibly provides a more robust sequence characteristic than an average pairwise identity, since it is based on multiple structural alignment itself. In case that the bio-core is small relative to pure structural alignment size, then the common sequence properties are evolutionary distant. In addition, an optimization according to the bio-core classification may give a different structure superposition, which may be more appropriate in some cases. In the Results section we calculate the pure geometrical structural alignment as well as the bio-core alignment.

## RESULTS

Below we provide, along with known results, new multiple protein structural alignments. We present applications of the **MultiProt** method for (1) a non-sequential structural similarity, (2) a partial alignment of hinge-bent domains, (3) identification of functional groups of G-proteins, (4) application to the analysis of binding sites, and (5) protein-protein interface alignment. Protein structures are taken from the Protein Data Bank.[30] All experiments were performed on a Pentium©IV 1800 MHz processor with 1024 MB internal memory. MultiProt is available for download at http://bioinfo3d.cs.tau.ac.il/MultiProt/.

The program output consists of ten (changeable parameter) highest-scoring results for each number of molecules, i.e., if the number of input molecules is 15, then there are sets of results for 2, 3,. . .,15 aligned molecules. Each result lists (1) 3D rigid transformation for each aligned molecule, (2) matrix of aligned amino acids, (3) RMSD of the multiple alignment (mRMSD), calculated as described above. We

**TABLE V. Multiple Structural Alignment Results Performed by MultiProt**

| Proteins | Number of Mols | Average Size | $S_{al}$ | Run-time |
|---|---|---|---|---|
| *4-helix bundle* | | | | 3s |
| 1f4n, 2cblA, 1b3q, 1rhgA | 4 | 284 | 75 | |
| 1b3q, 1rhgA | 2 | 451 | 102 | |
| *Superhelix* | | | | 14s |
| 1lxa, 1qq0, 1xat, 2tdt, 1fwyA:252–328 | 5 | 205 | 64 | |
| 1lxa, 1qq0, 1xat, 2tdt | 4 | 238 | 84 | |
| 1lxa, 1qq0, 2tdt | 3 | 248 | 114 | |
| 1lxa.pdb, 1qq0.pdb | 2 | 235 | 143 | |
| *Supersandwich* | | | | 12s |
| 1bgmI:731–1023, 1cb8A:336–599, 1oacA:301–724 | 3 | 360 | 118 | |
| 1cb8A:336–599, 1oacA:301–724 | 2 | 393 | 187 | |
| *Concanavalin* | | | | 54s |
| 2bqpA, 1gbg, 2galA, 1d2sA, 1sacA, 1a8d:1–247, 1kit:25–216, 2sli:81–276, 6cel, 1xnb | 10 | 220 | 54 | |
| 1a8d:1–247, 1d2sA, 1gbg, 1sacA, 2galA, 2sli:81–276 | 6 | 194 | 75 | |
| 1a8d:1–247, 1gbg, 6cel | 3 | 298 | 128 | |
| *tRNA synthetase* | | | | 39s |
| 1adjA, 1hc7A, 1qf6A, 1atiA-AntiCodon | 4 | 409 | 75 | |
| 1adjA, 1hc7A, 1qf6A | 3 | 508 | 176 | |
| *G-proteins* | | | | 29s |
| 1agr, 1tad, 1gfi, 1tx4, 1grn, 1wql | 6 | 370 | 13 | |
| 1agr, 1tad, 1gfi | 3 | 350 | 199 | |
| *PTB domain* | | | | 8s |
| 1x11,1irs,1shc,1ddm,2nmb,1evh | 6 | 147 | 66 | |
| 1shc,1ddm,2nmb | 3 | 168 | 111 | |

$S_{al}$ is the number of aligned atoms.

nickname the largest structural (bio) core to be *struct-core* (*bio-core*).

## Comparisons With Other Methods

### *Pairwise alignment cases*

First, we test our method on "hard to detect" *pairwise* alignments. We repeated the experiment presented in Shindyalov and Bourne.[7] The experiment presents a set of ten protein pairs and pairwise alignments performed by different (pairwise) methods. The results are presented in Table I. Two kinds of MultiProt results are given: alignments which preserve protein backbone order and sequence order independent alignments. As can be observed from the table, our pairwise results are very competitive. The maximal running time (pair 1crl: 534, 1ede: 310) is less than 4 s.

### *Globins*

The globin family has been extensively studied in the literature.[31,32] We applied MultiProt on seven globin structures (5mbn, 1ecd, 2hbg, 2lh3, 2lhb, 4hhbA, 4hhbB). We compared our results with those obtained in Wu et al.[32] The largest geometrical core detected by their method[32] consists of 105 $C_\alpha$ atoms (or "corresponding landmarks" as called in the paper). Our program obtains similar results. The size of the detected common *struct-core* varied between 93 $C_\alpha$ atoms ($\epsilon$ = 3 Å) to 111 $C_\alpha$ atoms ($\epsilon$ = 4 Å). The structural similarity is detected primarily between α-helices, while loop regions were left un-aligned. The detected *bio-core* is comparatively small. It ranged from 18 $C_\alpha$ atoms ($\epsilon$ = 3 Å) to 31 $C_\alpha$ atoms ($\epsilon$ = 4 Å). The running time was about 15 seconds.

### *Comparison with sequence order independent multiple alignment method*

A comparison with the results achieved by the MUSTA algorithm,[16] illustrates that our method achieved similar alignment results. The test includes the following cases:

- *Serpins* family—7apiA, 8apiA, 1hleA, 1ovaA, 2achA, 9apiA, 1psi, 1atu, 1kct, 1athA, 1attA, 1antl, 2antl.
- *Serine Proteinase*: *Subtilases* family—1cseE, 1sbnE, 1pekE, 3prkE, 3tecE.
- *Calcium-binding*: *EF hand-like* superfamily. The protein 4cpv is from the *parvalbu-min* family; 2scpA, 2sas, 1top, and 1scmB from the *calmodulin-like* family; and 3icb from the *Cal-binding D9K* family.
- *TIM-Barrels*: The proteins are taken from the seven different superfamilies. See details in Table II.
- *Helix-Bundle* The proteins are taken from the 6 different superfamilies. Details are given in Table II.

In all cases the size of the geometric core is at least the same [see Table III(a)]. In addition, our method produced high-scoring partial alignments and runs significantly faster (on the same computer).

### *Comparison with multiple alignments taken from the HOMSTRAD database*

The HOMSTRAD[33] database contains multiple alignments of homologous protein families. An average se-

quence identity (sID) in the protein families varies between 8 and 94 percent. We performed a comparison of 6 families, calcium-binding protein (sID 56%), subtilase (sID 52%), serine proteinase inhibitor (sID 34%), reductases (sID 22%), TPR domain (sID 17%) and plant virus coat protein (sID 13%).

To make an adequate comparison we performed the following procedure. From the HOMSTRAD database, for each of the six families we extracted the multiple structural alignments. Then, we computed the alignment size according to the MultiProt scoring method with the default parameters. No optimization on multiple transformations was performed. The computed alignment size represents the scoring of the HOMSTRAD database. Finally, we run MultiProt on the unaligned structures. As can be observed from Table III(b) the performance of both methods is very close.

## Applications of MultiProt

Here we demonstrate a variety of biological applications of MultiProt. First some case studies of multiple structural alignments are presented. Then, we show a partial alignment of hinge-bent domains, identification of functional groups of G-proteins and an application to the analysis of binding sites and protein-protein interface alignment. The results are summarized in Table V and VI.

### *Non-sequential structural similarity*

We consider an alignment of a 4-helix bundle. A 4-helix arrangement appears in a large number of proteins. SCOP includes at least 40 folds with a 4-helix bundle. Holm & Sander[28] show an alignment of the Rop protein (1rop) with cytochrome b56 (256b). Both proteins have 4-helix bundle, but the topological arrangement is different, i.e., when the two structures are aligned, at least one helix-pair is aligned in an opposite sequential order. Here we show a multiple structural alignment of four proteins (1f4n, 2cbl: A, 1b3q, 1rhg: A) which share a 4-helix bundle (see Fig. 3). Figure 3(c) shows the direction of the protein sequences according to a structural alignment when all four helices are aligned. As one can see the direction is different for the last two helices. Thus, none of the commonly used sequence alignment methods can align simultaneously the four α helices. Figure 3(b) shows a multiple structural alignment with the four helices aligned. See details in Table V.

### *Superhelix*

In this experiment we compare five proteins [1lxa, 1qq0, 1xat, 2tdt, 1fwy(A: 252-328)] from the *Superfamily: Trimeric LpxA-like enzymes*. Each protein is taken from a different family (for details see Table IV). While the first four molecules are between 208 and 274 residues long, the last one (1fwy, A: 252-328), is a truncated form and has only 77 residues. Our algorithm detected multiple alignment between all five molecules with *struct-core* of size 64 with mRMSD 0.9 Å. The *bio-core* for these molecules consisted of 17 $C_\alpha$-atoms. See the alignment in Figure 4(a).

Four molecules (the first four) gave 88 (18) $C_\alpha$-atoms in the *struct-core* (*bio-core*). Molecules 1lxa, 1qq0 and 2tdt gave 114 (27) $C_\alpha$-atoms in the *struct-core* (*bio-core*). (There are additional combinations of the first four molecules which are presented in the solutions). Molecules 1lxa.pdb and 1qq0.pdb gave 143 (62) $C_\alpha$-atoms in the *struct-core* (*bio-core*). For the three molecules there are other molecule pairs that gave high similarities. The running time was about 14 seconds. For details see Table V.

### *Concanavalin A-like lectins/glucanases*

From the SCOP database we selected the *Concanavalin A-like lectins/glucanases* fold (sandwich; 12–14 strands in two sheets; complex topology). This fold contains only one *superfamily* which includes ten different families. We selected one protein from each family (for details see Table IV): 2bqpA, 1gbg, 2galA, 1d2sA, 1sacA, 1a8d: 1-247, 1kit: 25-216, 2sli: 81–276, 6cel, 1xnb.

Aligning all ten molecules results in a geometric core of size 54. Interestingly, Tetanus Neurotoxin (1a8d: 1–247) participated in all alignments containing different numbers of molecules. This protein has five (A, B, C, D, E) β-sheets. A, C, and D create almost one β-sheet (let us call it S1), and so do B and E (S2). In the alignment of all ten molecules only β-sheet C is aligned well (from seven β-strands three were aligned well and two received only small partial alignments) and β-sheet E obtains only a small partial alignment. Investigating multiple alignments containing fewer molecules, we notice that the common core of β-sheet S1 increases and so does that of S2.

See Figure 4(b) for a geometric core of 6 molecules—1a8d: 1–247, 1d2sA, 1gbg, 1sacA, 2galA, 2sli: 81–276. The size of the *struct-core* is 75 $C_\alpha$ atoms with mRMSD 2.0 Å. The *bio-core* size of these molecules is nine atoms. The running time was 54 seconds. For details see Table V.

### *Supersandwich*

In this experiment we selected from the SCOP database[34] the *Supersandwich* fold from the *All beta proteins* class. This fold contains three superfamilies. From each superfamily we selected one protein, β-Galactosidase (1bgmI: 731–1023), Chondroitinase Ac (1cb8A: 336–599) and Copper Amine Oxidase (1oacA: 301–724). Our multiple alignment result contains 118 $C_\alpha$ atoms with mRMSD 2.21 Å. β-Galactosidase (1bgmI: 731–1023) has 17 strands. Only two strands (914–921,894–901) are not in the alignment. This example demonstrates that our method does not totally depend on the order of the residues in the backbone chain. A number of strands were aligned in the opposite order. Below is part of the alignment (notice the alignment between 1bgmI and 1oacA):

1bgmI: 738 739 740 741 746 747 748 749 750
1cb8A: 444 442 340 341 348 342 350 351 339
1oacA: 327 325 324 323 332 331 330 329 328

See Figure 4(c) for the aligned core. The *bio-core* of this multiple alignment contains 23 atoms, which is a sub-set of the largest detected *struct-core*. The running time was about 12 seconds. For details see Table V.

### *Mixed Experiment*

In this experiment, in order to show the power of our method, we included in the input set 18 proteins. Five

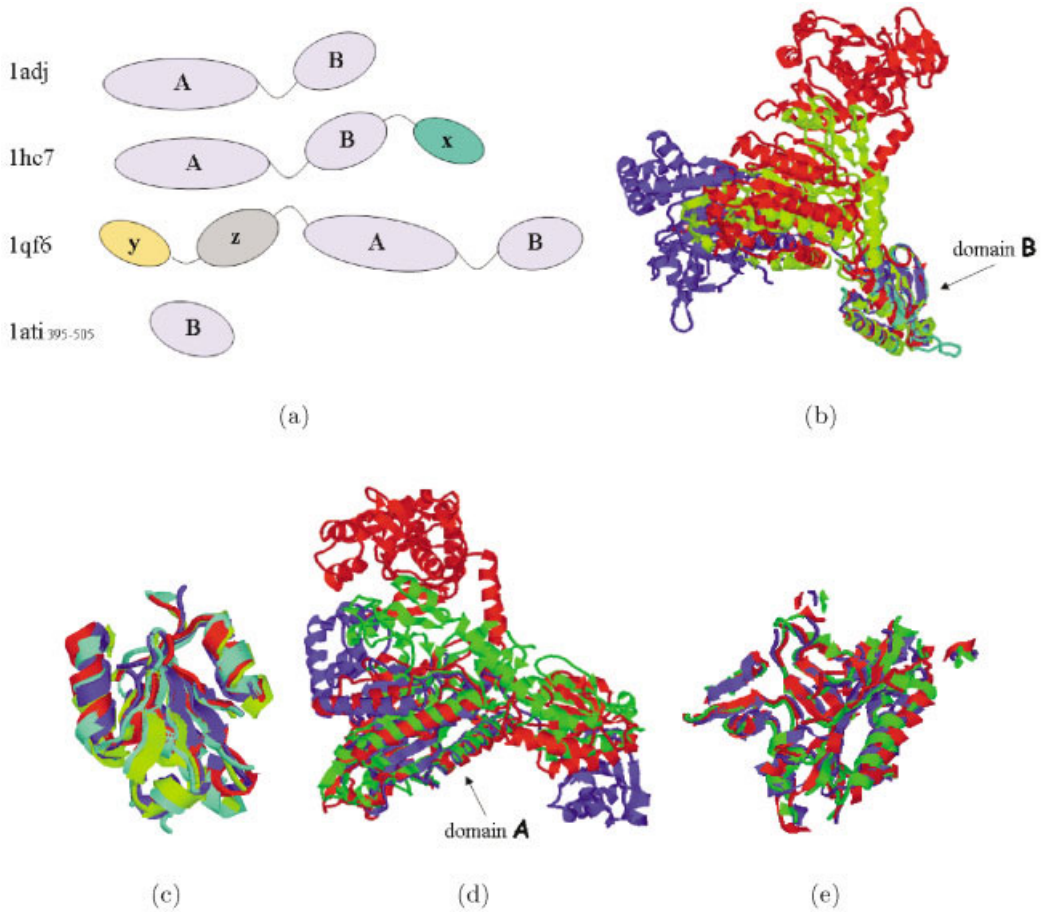



Fig. 5. Domain detection. **a**: Simplified schematic view of protein domains of 1adj, 1hc7, 1qf6, 1ati: 395–505. **b**: Four proteins are aligned with respect to the detected domain B. **c**: Only the multiply-aligned domain B is shown enlarged. **d**: Three proteins, 1adj, 1hc7, and 1qf6, are aligned with respect to domain A. Notice on the right that domain B is not aligned in this configuration, due to a hinge motion between A and B. **e**: Only multiply-aligned domain A is shown enlarged.

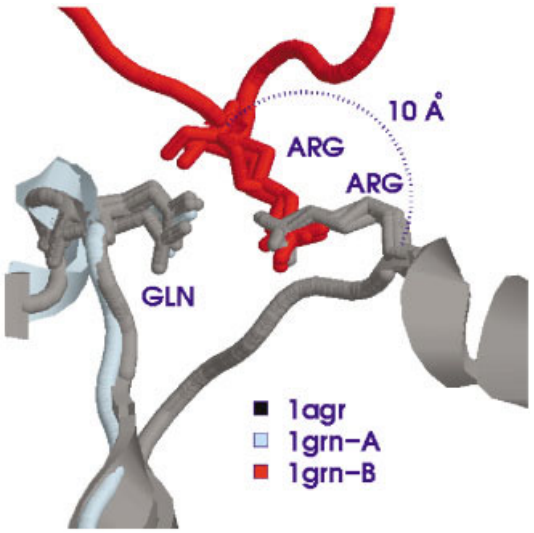



Fig. 6. The binding site of G-proteins has two functionally conserved amino acids Gln and Arg. In proteins 1agr, 1tad, and 1gfi the conserved amino acids are located on the same chain while in 1tx4, 1grn, and 1wql these amino acids are on different chains. For illustration, 1agr is colored dark-gray and two chains of 1grn are colored light-blue and red. The distance between the $C_\alpha$ atoms of the conserved Arg is about 10 $\alpha$, therefore a structural representation with $C_\alpha$ atoms is not appropriate. Using the amino acid representation described in text, MultiProt detected correctly the multiple alignment. The functional Gln and Arg were among the 13 structurally aligned residues.

proteins from the *Superhelix* experiment [1lxa, 1qq0, 1xat, 2tdt, 1fwy(A: 252–328)], three proteins from the *Supersandwich* experiment (1bgmI: 731–1023, 1cb8A: 336–599, 1oacA: 301–724) and ten proteins from the *Concanavalin A-like lectins/glucanases* experiment (2bqpA, 1gbg, 2galA, 1d2sA, 1sacA, 1a8d: 1–247, 1kit: 25–216, 2sli: 81–276, 6cel, 1xnb).

It took only eight minutes for the program to compare these 18 molecules. The multiple structural alignments for three molecules contained an alignment of the *Supersandwich* family. The results for five molecules contained the alignment of the *Superhelix* family. The results for six molecules contained the alignment of six proteins from *Concanavalin A-like lectins/glucanases* family (1a8d: 1–247, 1d2sA, 1gbg, 1sacA, 2galA, 2sli: 81–276), shown in Figure 4(b). Therefore, MultiProt detected the correct partial alignments. Since both families *Supersandwich* and *Concanavalin A-like lectins/glucanases* contain a number of β-sheets, 10–13 molecules contained the mixed alignments of proteins from these two families.

### *Detection of Partial Solutions*

In this example our goal is detection of partial solutions. We have four molecules each with several domains. Some domains are structurally similar. Our task is to identify these structurally similar domains.

The four proteins included in this study are Histidyl-tRNA Synthetase (1adj), Prolyl-tRNA Synthetase (1hc7), Threonyl-tRNA Synthetase (1qf6), and Glycyl-tRNA Synthetase (1ati). All four proteins have two common domains: *Class II aminoacyl-tRNA synthetase (aaRS)-like, catalyic domain,* and *Anticodon-binding domain of Class II aaRS*. For simplicity we call these domains A and B. Proteins 1hc7 and 1qf6 have domains in addition to A and B (see Fig. 5).

For the sake of an experiment, we cut domain A from protein 1ati, so only the second domain, B, is left (1ati: 395–505). Our goal in studying such a set of proteins is to identify in all four proteins a common domain B and in the three proteins 1adj, 1hc7, and 1qf6 a common domain A. However, aligning proteins 1adj, 1hc7, and 1qf6 with respect to domain A, does not align domain B at the same time. This is due to a hinge motion between domain A and B. Thus, there is no 3D transformation that simultaneously aligns both domains. Therefore, the problem requires detection of two different solutions (different 3D transformations and an appropriate set of molecules) for recognition of domains A and B. MultiProt successfully carries out this task (see Fig. 5).

### *G Proteins*

In this experiment we performed a multiple alignment of six G-proteins.[35] A binding site of these proteins has two conserved amino acids Gln and Arg. In three proteins (1agr, 1tad, and 1gfi) the conserved amino acids are located on the same chain while in the 1tx4, 1grn, and 1wql these amino acids are on different chains. Therefore, no sequence analysis method can detect these structurally conserved functional groups. Thus, to detect this structurally conserved pattern, a proper protein representation should be selected. The distance between $C_\alpha$ atoms of the conserved Arg of the first group (1agr, 1tad, and 1gfi) and the second group (1tx4, 1grn, and 1wql) is about 10 Å. Therefore, a method that aligns $C_\alpha$ atoms will not be able to recognize the functionally conserved Arg, since a distance threshold of 10 Å is not realistic.

For this experiment we represented the proteins by side-chain specific points. For Phe, Tyr, His, and Pro a geometric center of the ring was selected. Trp, Val, Ile, Thr were represented by a side chain geometric center; Ala, Ser by $C_\beta$ atoms; Gly by a pseudo $C_\beta$ atom; and Cys, Met by a sulfur atom. For Asp, Glu, Lys, Asn, Gln, Arg the last side-chain carbon atom was chosen. Leu was represented by a geometric center of the last three carbon atoms. Using such a representation, MultiProt successfully detected conserved Gln and Arg (Fig. 6), as well as other 11 amino acids. The selected side-chain specific representation was encouraged by this example. We are currently looking for other cases where such a representation would detect a correct structural alignment of conserved functional residues. Interestingly, multiple alignment using $C_\alpha$ atoms produces the same 3D superposition, but as explained above, the functionally conserved arginines are not "detected" to be aligned. This problem could be resolved by applying some post-analysis method. However, using a priori the correct representation is more advantageous, since the chance for false-positive solutions is lower.

### *PTB Domain*

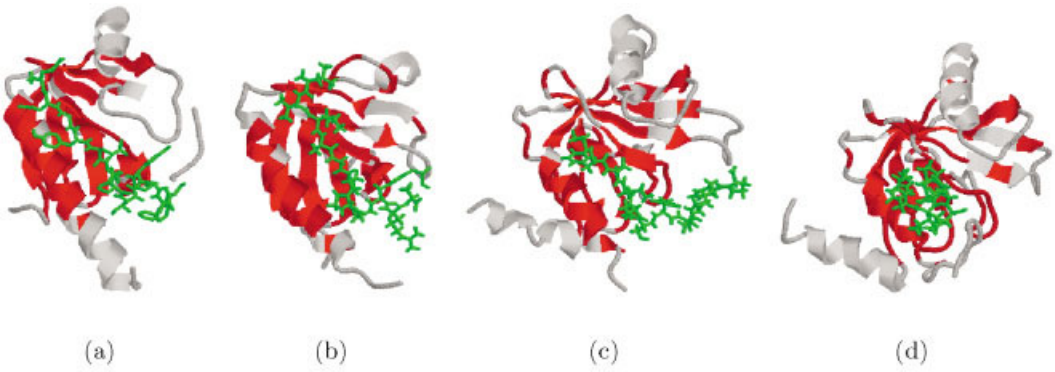
The PTB (phosphotyrosine-binding) domains are particularly interesting. Recent data indicate that these modular domains form part of a superfamily of recognition domains. They can interact with different targets on different parts of their surfaces. While they were originally believed to bind only phosphorylated targets, they are now known to bind also non-phosphorylated targets, some of which lack Tyr altogether. This diverse binding makes them a good case study. The PTB example demonstrates MultiProt's ability to handle protein-binding sites. MultiProt is able to compare all binding sites of available PTB domains, to obtain a consensus, common binding site core.

We selected six proteins (X11 PTB domain 1x11, Irs-1 PTB domain complexed with a Il-4 receptor phosphopeptide 1irs, Shc PTB domain complexed with a trka receptor phosphopeptide shc, Numb PTB domain complexed with a nak peptide 1ddm, Dnumb PTB domain complexed with a phosphotyrosine peptide 2nmb and Evh1 domain in complex with acta peptide 1evh). These molecules were taken from a study by Forman-Kay and Pawson.[36] The first five molecules are from the PTB family, while 1evh is from the EVH family. The detected common structural core contains 66 amino acids. For details see Figure 7.

### *Interface Alignment*[4]

Protein-protein interactions occur on the surface of a protein and are governed by physical-chemical forces. Thus, inspection of protein-protein interfaces should provide clues to the associations between different protein chains. Toward the ultimate goal of understanding how proteins interact, a good starting point is inspection of the interfaces structurally. A protein–protein interface consists of residues that interact with each other across the

[4]These results are contributed by Ozlem Keskin okeskin@ku.edu.tr



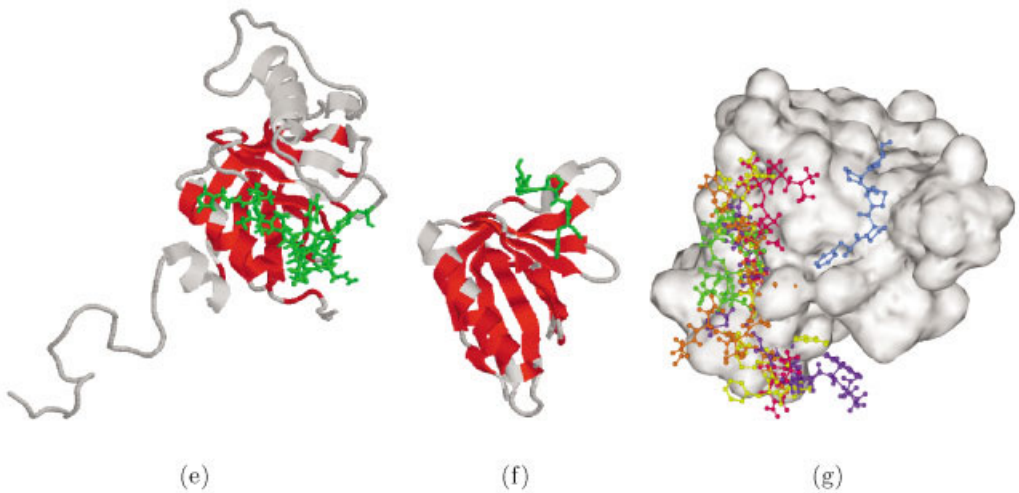
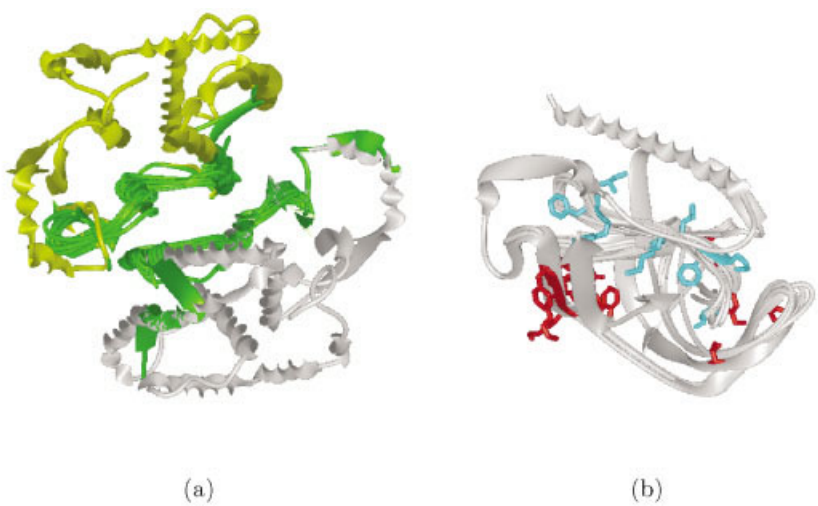




Fig. 7. **a–f**: Proteins (1×11,1irs,1shc,1ddm, 2nmb, 1evh) are displayed in orientation according to the multiple structural alignment by MultiProt. The common structural core is in red, whereas in gray are the structurally mis-aligned parts. The first five molecules are from the PTB family, while 1evh is from EVH. **g**: All peptides/ligands are superimposed according to the same multiple alignment as (a–f). While the first five ligands bind to almost the same surface region, the ligand on the right is taken from 1evh. Comparing structures (a–e) with (f) one can see that the binding site of 1evh is different from others.



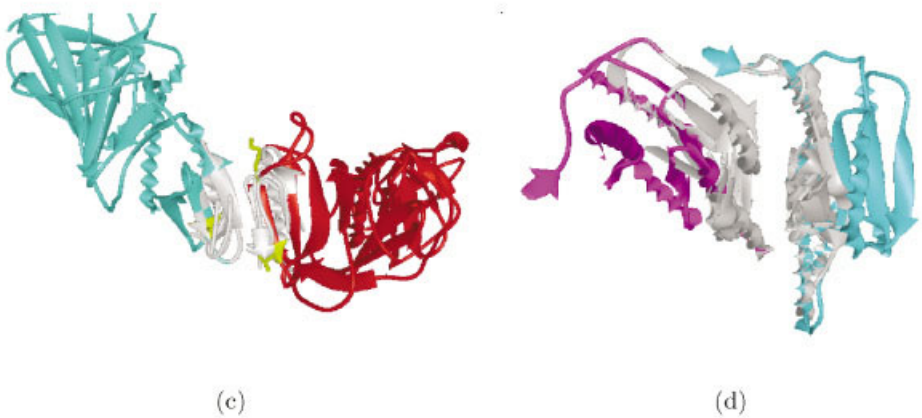




Fig. 8. **a–d**: The structural alignments of four different protein-protein interface families. In each figure, in addition to the aligned interfaces, the ribbon diagrams of the two chains that the interfaces belong to are displayed. These two chains are for the representatives of the family. See discussion in the text.

**TABLE VI. Protein Interface Families**[†]

| Family name | PDB codes | Number of Mols | $S_{al}$ | $S_{bio}$ | SCOP classification (Superfamily) |
|---|---|---|---|---|---|
| Transferases | 10gsAB, 1axdAB, 1b48AB 1c72AB, 1f2eAB, 1gnwAB 1gwcBC, 1jlvAB, 1ljrAB 1pd212 | 10 | 67 | 17 | Glutathione S-transferases, C-terminal domain |
| Ribonuclease, DNA binding proteins | 1a2pBC, 1axcAC, 1axcAE 1b77AB, 1b77AC | 5 | 18 | 0 | DNA clamp |
| Serine proteases | 1antLI, 1as4AB, 1c8oAB 1d5sAB, 1hleAB, 1jjoCE 1paiAB | 7 | 67 | 20 | Serpins |
| Transferase S | 1c2yAE, 1c41AB, 1c41AE 1ejbAB, 1hqkAB, 1hqkAE 1rvv12, 1rvvZ1 | 8 | 41 | 3 | Lumazine synthase |

[†]For the multiple alignment only the interface part was used, not the complete structure.
$S_{al}$ is the number of aligned amino acids, $S_{bio}$ is the size of the bio-core.

binding interface. At least two chains are involved. To be able to investigate the structural features of interfaces, we use our algorithm, which does not take into account the linear sequence information. Because interfaces are composed of at least two chains, and most of the time only discontinuous segments from each chain are involved in the binding, a structural alignment method that is independent of the order of the residues on the chains is essential. There is another very important point one should be careful in aligning interfaces: chain identities should be taken into consideration. Let us say we have two interfaces: Interface1 and Interface2. The first of these two interfaces is composed of chains ChainA and ChainB, and the latter ChainC and ChainD. If a method aligns the interfaces, all the segments from Chain A should be aligned with segments from ChainC (or ChainD) but not a combination of segments from both ChainC and D. MultiProt has the capability of aligning different interfaces simultaneously taking this criterion into account. With MultiProt, one can also detect sequentially conserved residues at a specific position on the alignments. This is an invaluable tool for finding the conserved residues hot spots at the interfaces.

Here, we apply our algorithm to clusters of protein-protein interfaces that are pre-filtered so that they are known to share common motifs at their interfaces. Only the interfaces are compared. The remainder of the chains are not considered. Figure 8(a–d) shows alignment of four different families of interfaces. Table VI gives the results of the interface alignment. These results have been obtained from O. Keskin et al. (Full details will be given elsewhere; Keskin, Tsai, Wolfson and Nussinov, unpublished).

Figure 8(a) is an example of the Glutathione S-transferases. The green part in the figure shows the results of the structural alignments of ten interfaces. The yellow and the gray parts are the ribbon diagrams of the two complete chains of Glutathione S-transferases 1c72 (Chains A and B). There are four α-helices and two β-strands in the interface that are being aligned by MultiProt.

In Figure 8(b), the interfaces of six serpins are displayed. The red and the cyan residues are the conserved residues of these six families. Note that only the conformations of the side chains of antichymotrypsin (1as4) are shown in the figure. Cyan depicts the conserved residues in Chain A and in red are the conserved residues in Chain B.

In Figure 8(c), the gray region displays the aligned five interfaces of *DNA binding* proteins, again the green and red chains are the ribbon diagrams of A and B chains of the DNA clamp 1b77.

Figure 8(d) displays the alignment of eight interfaces for the Lumazine synthase family. The magenta and the cyan regions show the ribbon diagram of the chains A and E of the Lumazine Synthase 1c2y. Gray depicts the interface. All the aligned interfaces have at most 3.5 Å RMSD amongst them. These examples are important, indicating that MultiProt can successfully align interfaces of different sizes and different numbers simultaneously and very efficiently. The interface examples presented here would not be aligned structurally with the other available structural alignment programs since the remainder of the chains would determine the alignments.

## CONCLUSIONS

Here we have presented a powerful tool for a simultaneous alignment of multiple protein structures. The advantage of MultiProt over previous methods is a combination of (i) simultaneous structure superposition (no side effects

of pairwise alignment methods), (ii) solutions are detected for any number of molecules (separation between more similar structures and outliers), (iii) proteins can consist of several chains, and (iv) the final alignments can optionally preserve the sequence order or be sequence-order independent. That is, MultiProt has the ability to detect non-topological similarities. Consequently, if there are at least three consecutive residues in the match, (v) MultiProt can be applied to multiply-align binding sites. The case studies presented here demonstrate these abilities. Despite the complex task, MultiProt is extremely efficient and is suitable for simultaneous comparison of up to tens of proteins.

Here we presented proteins as rigid molecules. A natural extension of multiple structural alignment is a *Flexible Multiple Structural Alignment*, where proteins are allowed to undergo hinge bending movements.[23] This work is now in progress.

## ACKNOWLEDGMENTS

We thank Ozlem Keskin for contribution of the results of the protein–protein interfaces. We thank Hadar Benyamini for biological suggestions and we thank I. Samish for bringing to our attention the paper by M. Kosloff and Z. Selinger. This research has been supported in part by the "Center of Excellence in Geometric Computing and its Applications" funded by the Israel Science Foundation (administered by the Israel Academy of Sciences). The research of H.J. Wolfson is partially supported by the Hermann Minkowski-Minerva Center for Geometry at Tel Aviv University. The research of R. Nussinov has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, under contract number NO1-CO-12400. The content of this publication does not necessarily reflect the view or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organization imply endorsement by the U.S. Government.

## REFERENCES

1. Durbin R, Eddy S, Krogh A, Mitchison G. Biological sequence analysis. New York: Cambridge University Press, 1998.
2. Gribskov M, McLachlan AD, Eisenberg D. Profile analysis: detection of distantly related proteins. Proc Natl Acad Sci USA 1987;84: 4355–4358.
3. Taylor WR, Orengo CA. Protein structure alignment. J Mol Biol 1989;208:1–22.
4. Mitchel EM, Artymiuk PJ, Rice DW, Willet P. Use of techniques derived from graph theory to compare secondary structure motifs in proteins. J Mol Biol 1989;212:151–166.
5. Vriend G, Sander C. Detection of common three-dimensional substructures in proteins. Proteins 1991;11:52–58.
6. Nussinov R, Wolfson HJ. Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques. Proc Natl Acad Sci USA 1991;88:10495–10499.
7. Shindyalov I, Bourne P. Protein structure alignment by incremental combinatorial extension (ce) of the optimal path. Protein Eng 1998;11:739–747.
8. Lemmen C, Lengauer T. Computational methods for the structural alignment of molecules. J Comput Aided Mol Des 2000;14: 215–232.
9. Eidhammer I, Jonassen I, Taylor WR. Structure comparison and structure patterns. J Comput Biol 2000;7:685–716.
10. Akutsu T, Halldorson MM. On the approximation of largest common subtrees and largest common point sets. Theor Comput Sci 2000;233:33–50.
11. Gerstein M, Levitt M. Using iterative dynamic programming to obtain accurate pairwise and multiple alignments of protein structures. In: Proceedings of the fourth international conference on intelligent systems in molecular biology (Menlo Park, CA). AAAI press, 1996. p 59–67.
12. Akutsu T, Sim KL. Protein threading based on multiple protein structure alignment. In: Genome informatics (GIW'99), Universal Academy Press, Tokyo, 1999. p 23–29.
13. Russell RB, Barton GJ. Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. Proteins 1992;14:309–323.
14. Taylor WR, Flores TP, Orengo CA. Multiple protein structure alignment. Protein Sci 1994;3:1858–1870.
15. Leibowitz N, Nussinov R, Wolfson HJ. MUSTA-a general, efficient, automated method for multiple structure alignment and detection of common motifs: application to proteins. J Comput Biol 2001;8:93–121.
16. Leibowitz N, Fligelman ZY, Nussinov R, Wolfson HJ. Automated multiple structure alignment and detection of a common substructural motif. Proteins 2001;43:235–245.
17. Fischer D, Lin SL, Wolfson HJ, Nussinov R. A geometry-based suite of molecular docking processes. J Mol Biol 1995;248:459–477.
18. Sandak B, Nussinov R, Wolfson HJ. An automated robotics-based technique for biomolecular docking and matching allowing hinge-bending motion. Comput Appl Bios (CABIOS) 1995;11:87–99.
19. Shatsky M, Fligelman ZY, Nussinov R, Wolfson HJ. Alignment of flexible protein structures. In: 8th International conference on intelligent systems for molecular biology. Heidleberg, Germany: AAAI press, 2000. p 329–343.
20. Jonassen I, Eidhammer I, Taylor WR. Discovery of local packing motifs in protein structures. Proteins 1999;34:206–219.
21. Dror O, Benyamini H, Nussinov R, Wolfson HJ. MASS: multiple structural alignment by secondary structures. Bioinformatics 2003;19, Suppl. 1:i95–i104
22. Dror O, Benyamini H, Nussinov R, Wolfson HJ. Multiple structural alignment by secondary structures: algorithm and applications. Protein Sci 2003;12(11):2492–2507.
23. Shatsky M, Wolfson HJ, Nussinov R. Flexible protein alignment and hinge detection. Proteins 2002;48:242–256.
24. deBerg M, vanKreveld M, Overmars M, Schwarzkopf O. Computational geometry–algorithms and applications. Berlin: Springer-Verlag, 2000.
25. Shatsky M, Nussinov R, Wolfson HJ. MultiProt–a multiple protein structural alignment algorithm. In: Guigo R, Gusfield D, editors. Workshop on algorithms in bioinformatics (Rome, Italy, 2002). Berlin: Springer Verlag; p 235–250. (Lecture notes in computer science; 2452).
26. Fischer D, Elofsson A, Rice D, Eisenberg D. Assessing the performance of fold recognition methods by means of a comprehensive benchmark. In: Hunter L, Klein T, editors. Proceedings of the Pacific symposium on biocomputing, 1996. Singapore: World Scientific Press, 1996. p 300–318.
27. Madej T, Gibrat J, Bryant S. Threading a database of protein cores. Proteins 1995;23:356–369.
28. Holm L, Sander C. Protein structure comparison by alignment of distance matrices. J Mol Biol 1993;233:123–138.
29. Bachar O, Fischer D, Nussinov R, Wolfson HJ. A computer vision based technique for 3-D sequence independent structural comparison. Protein Eng 1993;6:279–288.
30. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucleic Acids Res 2000;28:235–242
31. Bashford D, Chothia C, Lesk AM. Determinants of a protein fold. Unique features of the globin amino acid sequences. J Mol Biol 1987;196:199–216.
32. Wu TD, Schmidler SC, Hastie T. Regression analysis of multiple protein structures. J Comput Biol 1998;5:585–595.
33. Mizuguchi K, Deane CM, Blundell TL, Overington JP. HOMSTRAD: a database of protein structure alignments for homologous families. Protein Sci 1998;7:2469–2471.
34. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
35. Kosloff M, Selinger Z. Substrate assisted catalysis–application to g proteins. Trends Biochem Sci 2001;26:161–166.
36. Forman-Kay JD, Pawson T. Diversity in protein recognition by ptb domains. Curr Opin Struct Biol 1999;9:690–695.