BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 15 2005, pages 3255–3263
doi:10.1093/bioinformatics/bti527

*Structural bioinformatics*

# A new progressive-iterative algorithm for multiple structure alignment

Dmitry Lupyan[1], Alejandra Leo-Macias[2] and Angel R. Ortiz[2,*]

[1]Department of Physiology and Biophysics, Mount Sinai School of Medicine, One Gustave Levy Place, Box 1218, New York, 10029 NY USA and [2]Bioinformatics Unit, Centro de Biología Molecular 'Severo Ochoa' (CSIC-UAM), Universidad Autónoma de Madrid, Cantoblanco, 28049 Madrid, Spain

Received on April 28, 2005; revised on May 29, 2005; accepted on June 2, 2005
Advance Access publication June 7, 2005

**ABSTRACT**

**Motivation:** Multiple structure alignments are becoming important tools in many aspects of structural bioinformatics. The current explosion in the number of available protein structures demands multiple structural alignment algorithms with an adequate balance of accuracy and speed, for large scale applications in structural genomics, protein structure prediction and protein classification.

**Results:** A new multiple structural alignment program, MAMMOTH-mult, is described. It is demonstrated that the alignments obtained with the new method are an improvement over previous manual or automatic alignments available in several widely used databases at all structural levels. Detailed analysis of the structural alignments for a few representative cases indicates that MAMMOTH-mult delivers biologically meaningful trees and conservation at the sequence and structural levels of functional motifs in the alignments. An important improvement over previous methods is the reduction in computational cost. Typical alignments take only a median time of 5 CPU seconds in a single R12000 processor. MAMMOTH-mult is particularly useful for large scale applications.

**Availability:** http://ub.cbm.uam.es/mammoth/mult

**Contact:** aro@cbm.uam.es

## 1 INTRODUCTION

Multiple structural alignments (MStA) are important tools in a large number of applications in structural bioinformatics. They are crucial to provide benchmarks to improve sequence alignment algorithms, a cornerstone of bioinformatics research (Lassmann and Sonnhammer, 2002; Raghava *et al*., 2003). They are also helpful in protein structure classification and structure-based function prediction, highlighting structurally conserved regions of functional significance (May, 2002), as well as selectivity determinants (Al-Lazikani *et al*., 2001; Sheinerman *et al*., 2003). Similarly, they are extensively employed in protein structure prediction, providing high quality sequence profiles in fold recognition (Kelley *et al*., 2000; Petrey *et al*., 2003; Shi *et al*., 2001; Tang *et al*., 2003), aiding in the preparation of template–target alignments in comparative modeling (Burke *et al*., 1999), and helping to define low dimensional search subspaces for structural refinement. Finally, MStAs can also deepen our understanding of protein evolution, providing starting points for the analysis of structural and sequence variations in homologous proteins (Balaji and Srinivasan, 2001; Mizuguchi and Blundell, 2000). As a result, a variety of multiple structure alignment algorithms have been reported over the years (Godzik and Skolnick, 1994; Guda *et al*., 2004; Leibowitz *et al*., 2001b; Ochagavia and Wodak, 2004; Orengo and Taylor, 1996; Russell and Barton, 1992; Shatsky *et al*., 2004; Taylor *et al*., 1994; Yang and Honig, 2000). But despite all this previous research, structural alignment algorithms with an adequate balance of accuracy and speed are still required for large scale applications in protein structure prediction and protein structural classification. Here, we present a new algorithm to compute MStAs that addresses this problem. Our goal was to develop a deterministic but fast algorithm, adapted to large scale studies, but able to deliver high quality alignments and extensive structural cores. To this end, we have devised a hybrid algorithm, with a progressive-type layout but with two modifications to limit the deficiencies of progressive techniques. The first one is a correction step at each node of the tree to allow a dynamic reassignment of residue correspondences, similar to that of (Yang and Honig, 2000). The second is the introduction of an iterative refinement step at each node that helps to bring in register the most divergent members of the set. This refinement bears similarity with the iterative method proposed by Barton and Sternberg for multiple sequence alignment (Barton and Sternberg, 1987). The actual structural comparisons use the basic structure of MAMMOTH (Ortiz *et al*., 2002). Within MAMMOTH extensive use is made of C$\alpha$–C$\alpha$ vectors to assign correspondences, yielding a fast structure comparison program. We have made efforts to preserve this inherent speed in the multiple alignment version. This has allowed us to carry out an extensive optimization of the parameter space using manually curated databases of multiple structural alignments, such as HOMSTRAD (Mizuguchi *et al*., 1998) and CAMPASS (Sowdhamini *et al*., 1998). The result is a high quality of structurally implied sequence alignments.

## 2 METHODS

### 2.1 Multiple structural alignment (MStA) algorithm

The algorithm uses a standard progressive layout (steps 1 to 3) with two additional steps at each node (step 3.3) to minimize the greediness of the progressive algorithm (step 3.3) and to ensure a well defined core (3.4).

(1) Perform an all-against-all pairwise comparison using the standard MAMMOTH algorithm. A $N \times N$ similarity matrix is obtained, where

*To whom correspondence should be addressed.

for every pair $i$ and $j$ of structures, a similarity score is computed as $s_{ij} = -\ln(P_{ij})$, where $s_{ij}$ is the MAMMOTH score for pairwise alignment. $P_{ij}$ is the probability that the set of aligned residues could have been obtained by a random match of two different folds in the dabatase (Ortiz *et al*., 2002).

(2) Create a dendrogram by applying average linkage (Johnson and Wichern, 1998) to the matrix derived in step 1.

(3) Follow the nodes of the tree, from leaves to root. Let $A$ and $B$ be the two branches at a given node, and $n_A$ and $n_B$ the number of structures forming part of each branch. At each node we carry out the following steps:

(3.1) Assign correspondences between both subgroups based on their average C$\alpha$–C$\alpha$ vectors. We compute a similarity matrix **S** of the structures in $A$ with the structures in $B$ at each pair of positions $i$ and $j$ of the partial alignments already accumulated in each branch by averaging the pairwise C$\alpha$–C$\alpha$ vector similarities over the $n_A$ and $n_B$ elements of the branches:

$$S_{ij}^{AB} = \frac{1}{(n_A n_B)} \sum_{k=1}^{n_A} \sum_{l=1}^{n_B} u_{ijkl}, \qquad (1)$$

where $u_{ijkl}$ corresponds to the backbone similarity of the $i$ and $j$ positions for proteins $k$ and $l$, measured by their *URMS* similarity (Kedem *et al*., 1999), using the values previously stored in step 1. Gap penalties are the same employed in step 1. The alignment path along the **S** matrix is then obtained by a local–global dynamic programming step (Needleman and Wunsch, 1970), in order to yield correspondences between the $n_A$ and $n_B$ structures.

(3.2) Compute the 3D superimposition based with the MaxSub (Siew *et al*., 2000) routine implemented in MAMMOTH.

(3.3) Reassign correspondences based on the 3D superimposition. We compute a new similarity matrix following (Rossmann and Argos, 1976):

$$S_{ij}^{AB} = \frac{1}{(n_A n_B)} \sum_{k=1}^{n_A} \sum_{l=1}^{n_B} (w_b u_{ijkl} + w_d \mathrm{e}^{-\alpha d_{ijkl}^2}), \qquad (2)$$

where $d_{ijkl}$ is the euclidean distance between the C$\alpha$ atoms of residues in positions $i$ and $j$ for proteins $k$ and $l$, $\alpha$ is a parameter controlling the gaussian width, $w_b$ is the weight of the backbone similarity to the overall similarity and $w_d$ is the weight distance of the cartesian distance component. Once the matrix is filled, a dynamic programming step is taken to obtain the reassignments. This allows correction of possible misassignments in step 3.1 by introducing tertiary information based on the initial superimposition.

(3.4) Minimize the RMS fluctuation of the core by using a SIMPLEX optimization. For this last step, a procedure similar to that reported by (Barton and Sternberg, 1987) is employed. Each one of the structures in the family is displaced and rotated in turn, minimizing the RMS with respect to all other members, which are kept fixed. The process is iterated until convergence in the error function is achieved. The function to be minimized is:

$$\varepsilon_{\mathrm{core}} = \sum_{m=1}^{n_{\mathrm{core}}} \sum_{k}^{n_A+n_B} \sum_{k \neq l}^{n_A+n_B} d(\vec{r}_{mk}, \vec{r}_{ml})^2, \qquad (3)$$

where $n_{\mathrm{core}}$ is the number of core residues (see below), $n_A$ and $n_B$ are the number of elements of subfamilies $A$ and $B$, and $d^2$ is the squared cartesian distance between the C$\alpha$ atoms of proteins $k$ and $l$ in position $m$ of the core in the structural alignment. The rationale is that because neither step 3.2 nor 3.3 force the structural core to be optimally superimposed, an additional step is required to ensure it.

## 2.2 Parameter optimization

Gap opening and extension penalties for steps 1 and 3.1 were taken directly from the pairwise version of MAMMOTH (Ortiz *et al*., 2002). The assignment correction step 3.3 needs three additional parameters: gap initiation, gap penalty and weight of the distance C$\alpha$ matching. These were optimized empirically with a Monte Carlo based simulated annealing algorithm and a training set of 105 families (see next paragraph) using the quality scores defined in the next section with a set of HOMSTRAD alignments (see below). Simulations were started from an initial set of parameters guessed by trial and error. 500 simulation steps were found to be required for convergence. Jackknife testing was used to assess the reliability of the final parameter set (gap opening, 7.0; gap extension, 0.15; distance weight, 8.0; and threshold, 4.0 Å). No additional parameters are required.

## 2.3 Performance criteria

We monitored the quality of the alignments with three different parameters: (1) Core extension (%core), measured as percentage of residues in the core (with respect to the shortest protein model). We defined two types of cores, 'strict core' and 'loose core'. The 'strict core' is formed by the set of positions with 100% conservation, and within 4.0 Å of each other in the final structural alignment in 3D. The 'loose core' comprises those positions with at least 66% conservation, and within 3.0 Å from their average for that position in the alignment in 3D. This last core definition is used as the target function in the parameter optimization step described above. On the other hand, the 'strict core' is used in all comparisons in this article; (2) mean RMS fluctuation ($\langle \mathrm{RMS}_{\mathrm{core}} \rangle$) of the strict core residues; and (3) Quality of the implied multiple sequence alignment, as computed with norMD (norMD score) (Thompson *et al*., 2001). Note that this value is used only to measure the quality of the output; sequence information is not used at any point to build the multiple structural alignment.

## 2.4 Training and testing sets

(1) HOMSTRAD set. A set of 105 structural alignments were compiled from the HOMSTRAD database (Mizuguchi *et al*., 1998), a database of multiple structure alignments for homologous families (http://www-cryst.bioc.cam.ac.uk/homstrad/). Since they are manually curated, HOMSTRAD alignments are adopted as gold standard at the family level.

(2) CAMPASS set. This set comprises 551 manual alignments at the superfamility level (Sowdhamini *et al*., 1998), derived in a similar manner to those of HOMSTRAD. They were downloaded from http://www-cryst.bioc.cam.ac.uk/~campass/. CAMPASS alignments are adopted here as gold standard at the superfamily level.

(3) Multiprot set. This set corresponds to the set of structures in the CAMPASS set discussed above, but with the alignments generated using the MultiProt program (Shatsky *et al*., 2004). Structurally implied sequence alignments were generated from the MultiProt output using the program Stacatto (M. Shatsky, personal communication).

(4) FSSP set. FSSP is a collection of structure alignments generated with Dali, which provides multiple alignments based on a 'pileup' alignment of structural neighbors from 'master–slave' pairwise alignments. Although they are not strictly multiple structural alignments, we have included FSSP alignments, as the FSSP database is a very popular resource for multiple structure comparisons. To build this set, an all-against-all comparison was performed with MAMMOTH on a set of downloaded FSSP alignments (Holm and Sander, 1997, 1998). A clique-detection algorithm was used to select sets of structures that could be aligned, both with MAMMOTH-mult and FSSP. 1385 cliques were obtained, where all members in the clique were above the threshold of structural similarity, according to both MAMMOTH and FSSP. The results of applying MAMMOTH-mult to this set were compared to the alignments processed by DaliLite (Holm and Park, 2000), using the performance criteria previously described.

**Table 1.** Comparison of structural alignment quality for different sets

| Alignment method | Structural level | $Z$-score (%core) | $Z$-score ($\langle RMS_{core} \rangle$) | $Z$-score (norMD) |
|---|---|---|---|---|
| (A) Semi-manual | | | | |
| HOMSTRAD | Family | 1.03 | 0.31 | 0.10 |
| CAMPASS | Superfamily | 9.17 | 8.78 | 4.81 |
| (B) Automatic | | | | |
| DaliLite | Family | 4.97 | 3.39 | 5.41 |
| | Superfamily | 7.04 | 4.72 | 4.20 |
| | Fold | 2.73 | 2.37 | 2.27 |
| MultiProt | Superfamily | 11.05 | 5.22 | 7.60 |

The three parameters describing alignment quality (%core, $\langle RMS_{core} \rangle$ and norMD, see Methods), are computed for four datasets (FSSP, MultiProt, HOMSTRAD and CAMPASS, see Methods) using both MAMMOTH-mult and a reference method (DaliLite and MultiProt for the case of automatic alignments, and alignments downloaded from the corresponding web servers for HOMSTRAD and COMPASS). A Wilcoxon sum of ranks test was applied to the two-paired groups. The $Z$-score of the resulting statistic is shown. Positive values indicate improvement of MAMMOTH-mult over the alternative method, negative is the opposite. Only positive values were obtained. $Z$-scores $>3.0$ are indicative of statistically significant differences.

(5) Superfolds set. This set is formed by alignment of proteins belonging to two different superfolds (immunoglobulins and globins), whose structural alignments and structural classification have been studied in detail by different authors. The immunoglobulin set corresponds to a set of 26 domains analyzed by Bork *et al*. (1994), who manually classified the structures into four distinct groups. We also studied the globin fold, as this is a classical fold analyzed in most investigations in multiple structure alignment. We used the SCOP classification as gold standard.

## 3 RESULTS

### 3.1 Quality of the MStA

The performance of the MStA algorithm is summarized in Table 1. We first discuss the results with respect to the gold standards. For the 105 HOMSTRAD families, the average values of the three scoring parameters used to evaluate the success of the method (%core, $\langle RMS_{core} \rangle$ and norMD score, see Methods) take values, respectively, of 71%, 0.80 Å and 0.63 for MAMMOTH mult, and 67%, 0.78 Å and 0.63 for HOMSTRAD. MAMMOTH-mult results are slightly better than those of the HOMSTRAD, but rank differences are not statistically significant (Table 1). However, as structures within the family begin to diverge, MAMMOTH-mult tends to arrive at better structural alignments than HOMSTRAD (Fig. 1). An extreme example is shown in Figure 2A and B, which shows the superimpositions obtained by HOMSTRAD and MAMMOTH-mult for the eight members of the C-type lectin family. Here, the HOMSTRAD alignment has 19.82% of the residues in the core and 1.01 Å $\langle RMS_{core} \rangle$, while the MAMMOTH-mult alignment has 59.50% residues in the core and 0.86 Å $\langle RMS_{core} \rangle$. The tendency of MAMMOTH-mult to generate improved alignments, as compared with manual methods, with increasingly divergent structures is confirmed with the CAMPASS set. In this case (Fig. 1) MAMMOTH-mult arrives at better alignments, with strongly significant statistical differences (Table 1), particularly regarding the size of the core and the RMS fluctuation of the core residues. Figure 2C and D shows an example of the differences in alignment for the 8-member Sm-like ribonucleoproteins superfamily. CAMPASS arrives at a final alignment with 6.12% in the core and 1.84 Å $\langle RMS_{core} \rangle$, whereas MAMMOTH-mult provides 40.62% residues in the core and 1.44 Å $\langle RMS_{core} \rangle$.

Next, we compared MAMMOTH-mult alignments with other automatic methods. In the comparison with MultiProt (Shatsky *et al*., 2004) using the CAMPASS set of alignments, we also observe a significant improvement in the alignments with MAMMOTH-mult (Table 1). Improvements are particularly significant in the size of the core detected and, in this case, in the quality of the implied sequence alignment. We also made a comparison with FSSP alignments, as obtained with DaliLite (Holm and Park, 2000). We first ensured compatibility between both programs (see Methods). For the 1385 alignments selected, the average values of the three scoring parameters take the values of 37%, 1.01 Å and $-0.89$ for MAMMOTH-mult, and 33%, 1.18 Å and $-1.19$ for FSSP. When alignments are divided into structural classes (family, superfamily and fold, see Table 1), we observe that improvements are large at the family and particularly superfamily levels, but more modest at the fold level. It must be noted that in many of the FSSP as well as in some MAMMOTH alignments, conventional core was not detected (0%); these cases were not considered in the calculation of averages. Finally, we carried out multiple structural alignments with two sets of structures used by (Ochagavia and Wodak, 2004) in their validation of MALECON. The first one is a globin set, formed by the following structures: 1ash, 1eca, 1gdj, 1hlm, 1mba, 1babA, 1ew6A, 1h97A, 1ithA, 1sctA, 1dlwA, 1flp, 1hbg, 1lhs and 1vhbA. Although a direct comparison of the results should be done with caution, owing to the slightly different core definitions, the core detected by MAMMOTH-mult comprises 131 residues with a $\langle RMS_{core} \rangle$ of 1.56 Å, while 59 residues with a $\langle RMS_{core} \rangle$ of 1.73 Å were reported with MALECON (Ochagavia and Wodak, 2004). For the second case, the OB fold set (1afp, 1b9nA3, 1ckmA2, 1esfA1, 1fr3A, 1jic, 1tiiD, 2tmp, 1b7yB2, 1bovA, 1eif02, 1fjgQ, 1htp, 1sro and 2sns), MALECON and MAMMOTH-mult show a similar behavior—both failing to find an alignment with the complete set. When the set is reduced to 10 structures (1sro, 1b7yB2, 1tiiD, 1bovA, 2sns, 1esfA1, 1eif02 1fjgQ 1b9nA3 and 1fr3A), MAMMOTH-mult and MALECON produce similar results (not shown).

### 3.2 Analysis of representative cases

In what follows, we discuss the alignments automatically derived by MAMMOTH-mult for two well-known test cases. Extensive comparisons have been done in our group for a large number of cases, but for the sake of brevity, we provide here results only for two

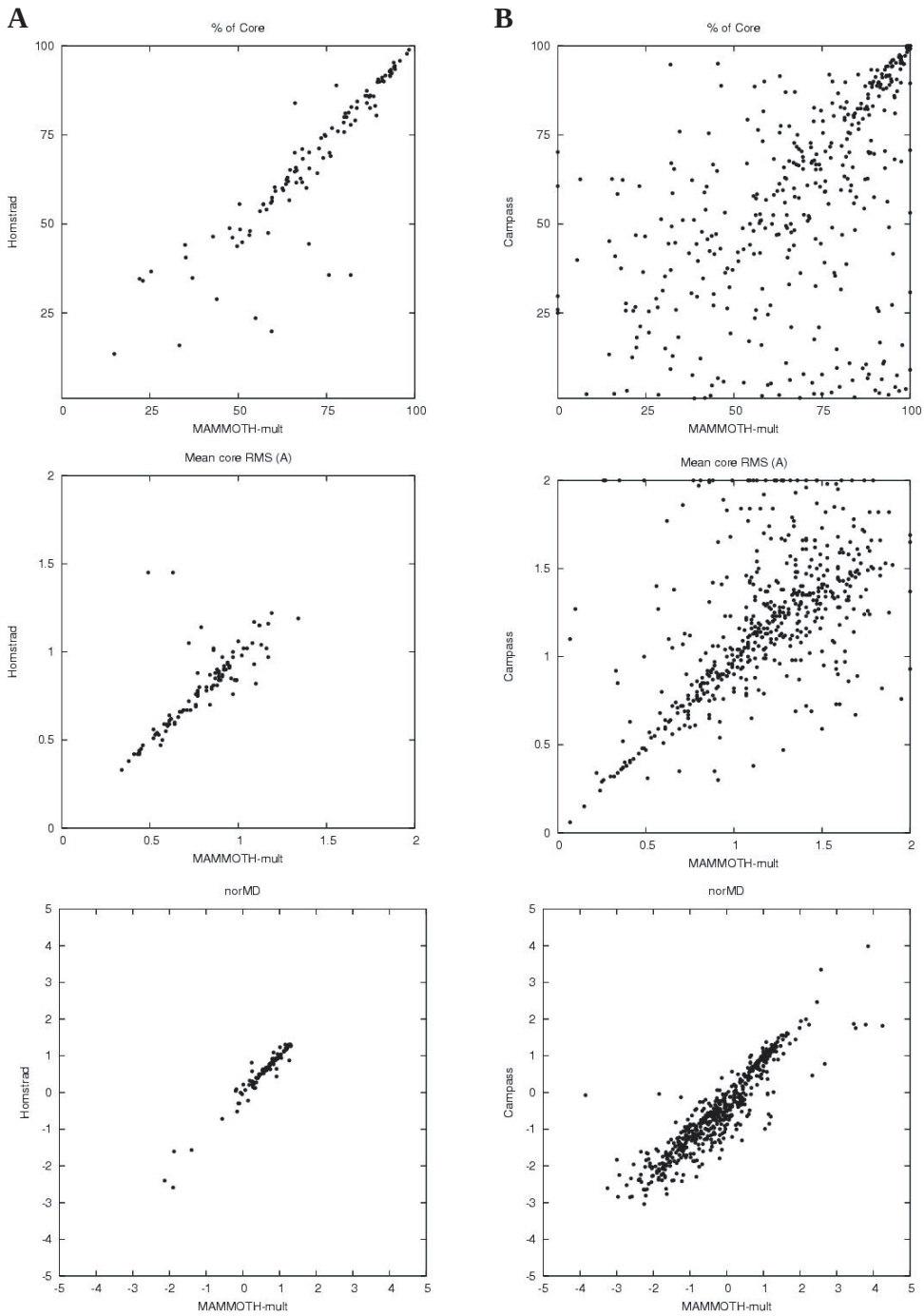


**Fig. 1.** Structural alignments generated by MAMMOTH-mult in comparison with other methods. *x*-axis refers to the MAMMOTH results, while the *y*-axis refers to the reference method. Each point corresponds to one protein family, with an average number of members per family is seven. Plots for percentage of core (%core), mean RMS fluctuation of the residues in the core ($\langle RMS_{core} \rangle$), and norMD score (norMD) are shown. (**A**) HOMSTRAD; (**B**) CAMPASS.

representative examples (described as Superfold Set in Methods): immunoglobulins and globins.

*3.2.1 Immunoglobulins* A set of 26 different immunoglobulin domains were semi-manually classified in four different subtypes (v-type, h-type, s-type and c-type) by Bork *et al*. (1994). Their study grouped the structures on the basis of the number, connection and variations in the position of the edge strands relative to a common core of four beta strands. The common core found by MAMMOTH-mult can be observed in Figure 3, using the notation employed by Bork *et al*. MAMMOTH-mult also finds a common core of well conserved four central strands (B, C, E and F), and variations in the positioning

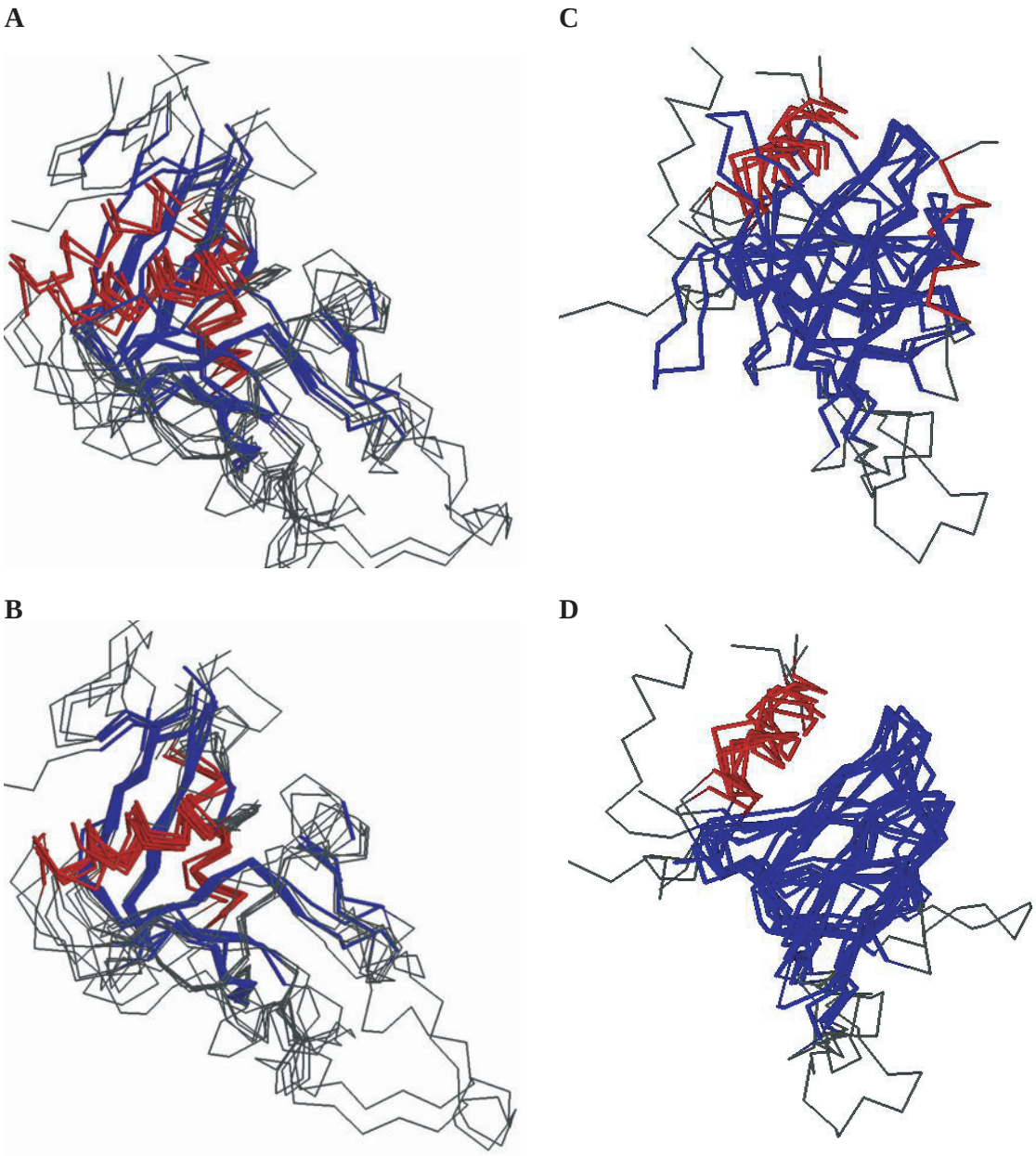


**Fig. 2.** Structural alignments for the eight members from the C-type lectin family (**A**, HOMSTRAD; **B**, MAMMOTH-mult), and the eight members of the Sm-like ribonucleoproteins superfamily (**C**, CAMPASS; **D**, MAMMOTH-mult). The colored regions (blue for strands and red for helices) highlight structurally conserved core regions, as defined by MAMMOTH-mult.

of the edge strands A, C′ and G (Fig. 3). In agreement with their study, we also find considerable variation in the A, C″, D strands, which do not form part of the MAMMOTH-mult evolutionary core.

The fourth group (h-type), however, only consists of one structure (1gof), with the other two members (1cgt and 1clc) distributed among the other c- and s-types. This is not surprising, as the h-type is considered a hybrid between them (Bork *et al*., 1994). Regarding the alignment, a clear conservation of aromatic residues in the core strands, particularly for strands C and F, can be noted (Fig. 3), a feature also observed by Bork and coworkers. Similarly, the increase of the length of the CE segment in going from the s- to c- to v-type is also apparent, as already noted (Bork *et al*., 1994).

*3.2.2 Globins* Globins are organized in the so-called three-on-three $\alpha$-helical fold in SCOP (Andreeva *et al*., 2004; Murzin *et al*., 1995). A set of 23 globins (superfamily SCOP entry 46 458) was selected from ASTRAL (Brenner *et al*., 2000; Chandonia *et al*., 2004) so that all pairwise sequence identities were $<40\%$. The set comprises three different SCOP families (46 459, 74 660 and 46 463). The tree produced by MAMMOTH-mult reproduces the SCOP

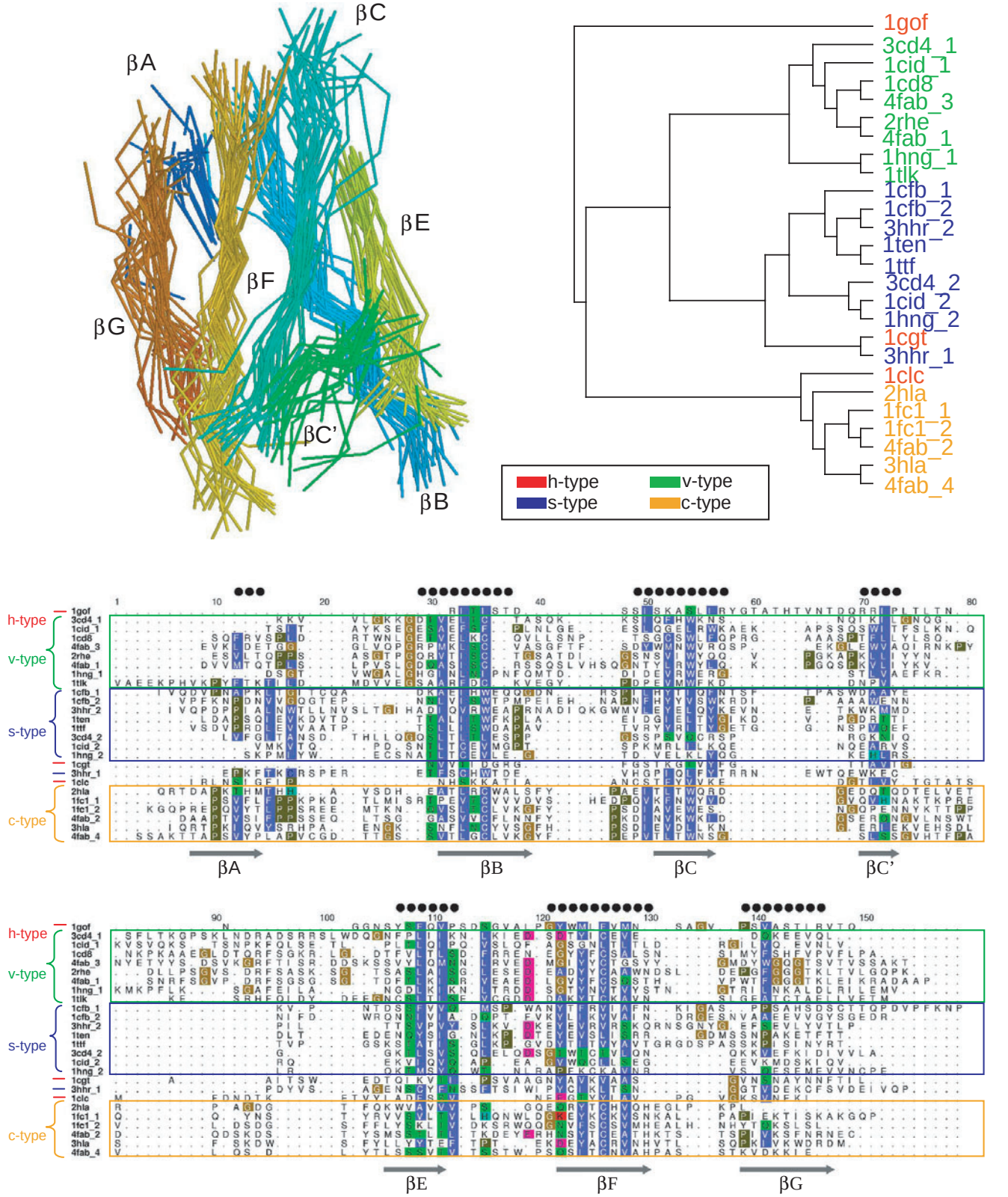


**Fig. 3.** MAMMOTH-mult structural alignments for immunoglobulins. The implied multiple sequence alignment, the structural alignment for the evolutionary core ('large core', see Methods) detected by MAMMOTH-mult and the dendrogram corresponding to the displayed structural alignment are shown. Structure IDs in the dendrogram are colored according to reference manual classifications (see Methods).

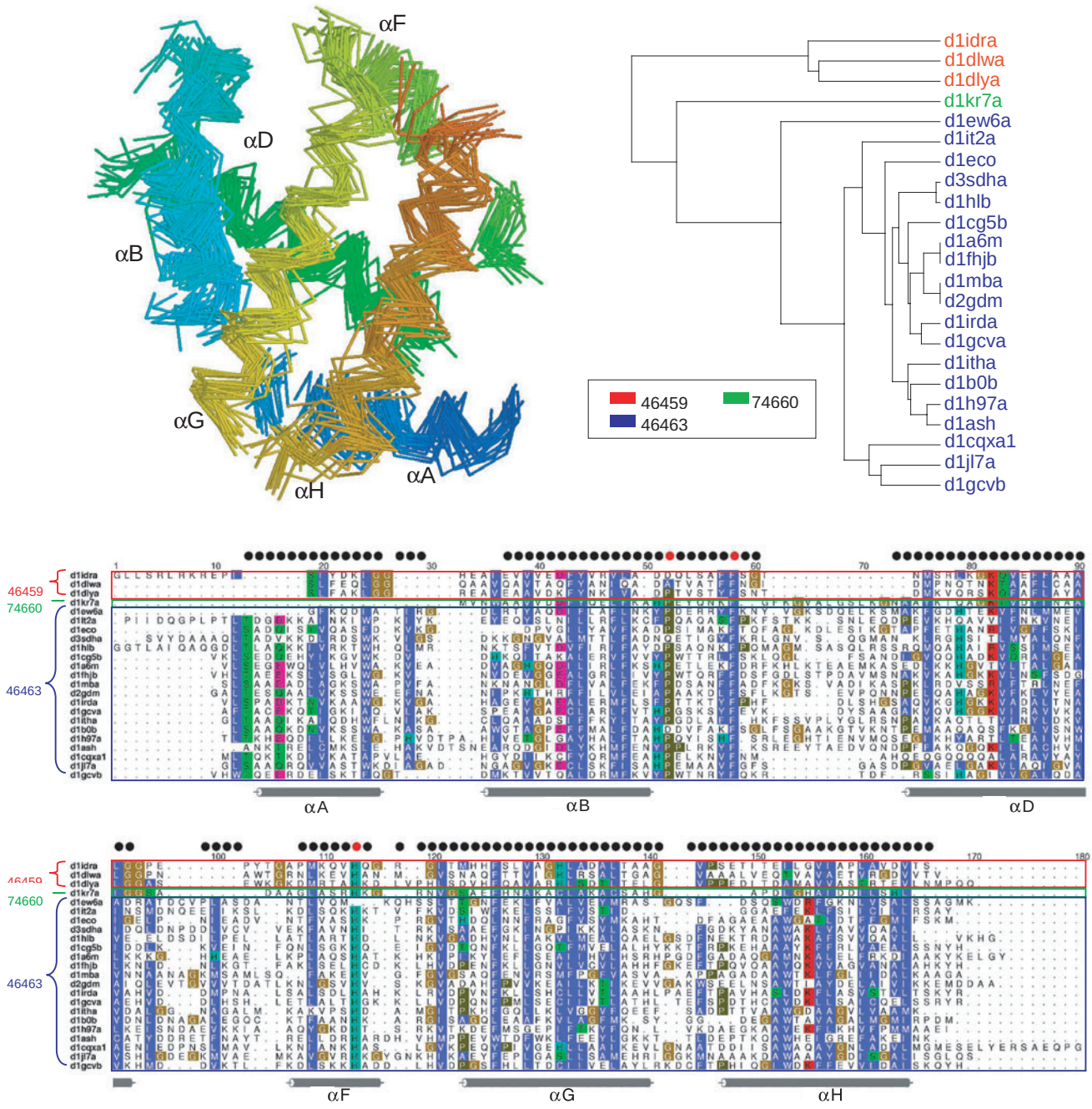


**Fig. 4.** MAMMOTH-mult structural alignments for globins. The implied multiple sequence alignment, the structural alignment for the evolutionary core ('large core', see Methods) detected by MAMMOTH-mult and the dendrogram corresponding to the displayed structural alignment are shown. Structure IDs in the dendrogram are colored according to reference manual classifications (see Methods).

classification exactly, separating canonical heme-binding globins (46 463) from protozoan/bacterial truncated hemoglobins (46 459) and neural hemoglobins (74 660) (Fig. 4). In the alignment there is a clear conservation of the key proximal histidine in the active site (position 113, HisF8 following Perutz's notation). This histidine establishes a Fe–N bond with the Fe atom in the heme group, and its conservation is one of the distinct features of a globin profile (Kapp *et al*., 1995). The only protein failing to align a histidine in this position is d1ew6a_ (LaCount, 2000), from Amphitrite ornate. Interestingly, this is the only dehaloperoxidase in the set. In this case, the proximal histidine is shifted by three positions. This displacement probably forces a 60° rotation in the imidazol moiety. Previous studies have suggested that this rotation helps to establish a stronger Fe–N bond that contributes to the electron push needed by the peroxidase activity (LaCount, 2000). There is also an absolute conservation of the phenylalanine PheCD1 (position 58), considered

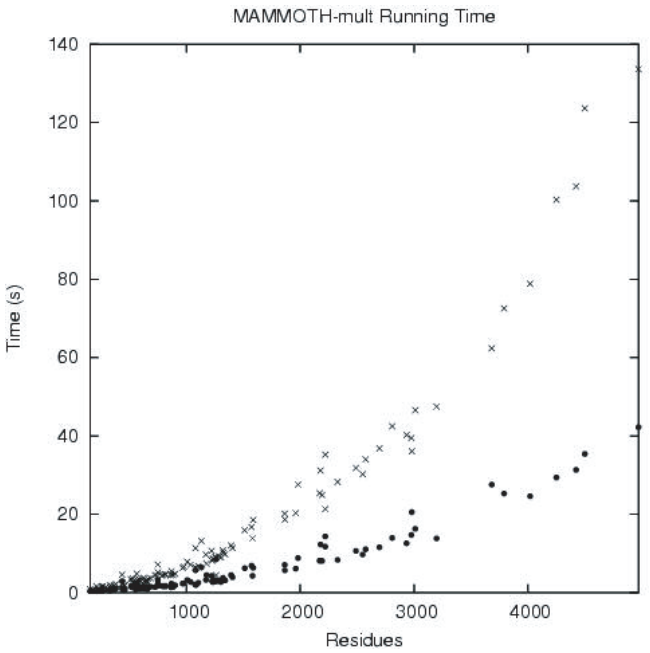


**Fig. 5.** MAMMOTH-mult timings. Total number of residues aligned in the family versus running time (in seconds). Results are shown for computations on a R12000 processor (crosses) and a Pentium IV PC at 2 GHz running Red-Hat Linux (dots).

a key residue in the interaction of the protein with the heme group, and also fully conserved in globin alignments (Kapp *et al*., 1995). Finally, ProC2 (position 52), in the immediate neighborhood of the heme binding pocket, is almost completely conserved, as noted in previous analysis (Ptitsyn and Ting, 1999).

### 3.3 Timings

Computational times are crucial for large scale applications. The dependence of the computing time with respect to the number of residues aligned is shown in Figure 5. A typical alignment of 15 structures with 150 residues each takes ∼5 s using a Pentium IV PC at 2 GHz and ∼25 s in a R12000 (Fig. 5). Aligning the complete set of 105 structural families from HOMSTRAD requires 27 min of CPU time on a single processor R12000. For comparison, the method of Nussinov *et al*. (Leibowitz *et al*., 2001a) requires ∼10 h of CPU time only to structurally align a set of 10 TIM barrels. The same set of structures can be aligned with MAMMOTH-mult in 28 s in R12000 processor. The reasons for these short computing times are that the underlying pairwise alignment algorithm is intrinsically very fast, and that at each node in steps 3.1 and 3.2 the URMS information is obtained from look-up tables filled during the pairwise comparisons at step 1.

### 3.4 MAMMOTH-mult server

A web server enabling the use of the program has been established at http://ub.cbm.uam.es/mammoth/mult. The server can be used in two different ways: it can either multiple align a target protein against a given SCOP superfamily, or align among them a set of input proteins. In the first case, a form is presented to the user where a protein structure in PDB format can be uploaded and a tag of the superfamily can be requested; the user then can select a subset of the domains to perform the alignment. In the second case, the input file to be uploaded consists of a single file with all the proteins to align in PDB format. In both cases the server performs the alignment and e-mails the results to the user.

## 4 CONCLUSIONS

A new MStA algorithm is described. The alignments produced with the new method show improved quality when compared with other methods, and are at least of the same quality of available manual alignments in several widely used databases. Detailed analysis of the alignments for two well characterized cases indicates that MAMMOTH-mult produces biologically meaningful trees, and preserves conservation of functional and structural motifs in the alignments. Typical alignments take an average of ∼5 CPU seconds in a standard desktop workstation. Overall, these results show that MAMMOTH-mult can be particularly useful for large scale applications in protein structure classification, protein structure prediction and in structural genomics applications. A web server enabling the use of the program is available at http://ub.cbm.uam.es/mammoth/mult

## ACKNOWLEDGEMENTS

We thank Dr M. Shatsky for generously providing us with the program Stacatto. This work has been partly funded by grant BIO2001-3745 from the Spanish MCYT. A.L.M. is a FPI predoctoral fellow. Research at Centro de Biología Molecular 'Severo Ochoa' is facilitated by an institutional grant from Fundación Ramón Areces.

*Conflict of Interest:* none declared.

## REFERENCES

Al-Lazikani,B. *et al*. (2001) Combining multiple structure and sequence alignments to improve sequence detection and alignment: application to the SH2 domains of Janus kinases. *Proc. Natl Acad. Sci. USA*, **98**, 14796–14801.

Andreeva,A. *et al*. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res.*, **32**, D226–D229.

Balaji,S. and Srinivasan,N. (2001) Use of a database of structural alignments and phylogenetic trees in investigating the relationship between sequence and structural variability among homologous proteins. *Protein Eng.*, **14**, 219–226.

Barton,G.J. and Sternberg,M.J. (1987) A strategy for the rapid multiple alignment of protein sequences. Confidence levels from tertiary structure comparisons. *J. Mol. Biol.*, **198**, 327–337.

Bork,P. *et al*. (1994) The immunoglobulin fold. Structural classification, sequence patterns and common core. *J. Mol. Biol.*, **242**, 309–320.

Brenner,S.E. *et al*. (2000) The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res.*, **28**, 254–256.

Burke,D.F. *et al*. (1999) An iterative structure-assisted approach to sequence alignment and comparative modeling. *Proteins*, (Suppl. 3), 55–60.

Chandonia,J.M. *et al*. (2004) The ASTRAL compendium in 2004. *Nucleic Acids Res.*, **32**(Database issue), D189–D192.

Godzik,A. and Skolnick,J. (1994) Flexible algorithm for direct multiple alignment of protein structures and sequences. *Comput. Appl. Biosci.*, **10**, 587–596.

Guda,C. *et al*. (2004) CE-MC: a multiple protein structure alignment server. *Nucleic Acids Res.*, **32**, W100–W103.

Holm,L. and Park,J. (2000) DaliLite workbench for protein structure comparison. *Bioinformatics*, **16**, 566–567.

Holm,L. and Sander,C. (1997) Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids Res.*, **25**, 231–234.

Holm,L. and Sander,C. (1998) Touring protein fold space with Dali/FSSP. *Nucleic Acids Res.*, **26**, 316–319.

Johnson,R. and Wichern,D. (1998) *Applied Multivariate Statistical Analysis*. Prentice Hall, Upper Saddle City, New Jersey.

Kapp,O.H. *et al*. (1995) Alignment of 700 globin sequences: extent of amino acid substitution and its correlation with variation in volume. *Protein Sci.*, **4**, 2179–2190.

Kedem,K. *et al*. (1999) Unit-vector RMS (URMS) as a tool to analyze molecular dynamics trajectories. *Proteins*, **37**, 554–564.

Kelley,L.A. *et al*. (2000) Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J. Mol. Biol.*, **299**, 499–520.

LaCount,M.W. *et al*. (2000) The crystal structure and amino acid sequence of dehaloperoxidase from Amphitrite ornata indicate common ancestry with globins. *J. Biol. Chem.*, **275**, 18712–18716.

Lassmann,T. and Sonnhammer,E.L. (2002) Quality assessment of multiple alignment programs. *FEBS Lett.*, **529**, 126–130.

Leibowitz,N. *et al*. (2001a) Automated multiple structure alignment and detection of a common substructural motif. *Proteins*, **43**, 235–245.

Leibowitz,N. *et al*. (2001b) MUSTA–a general, efficient, automated method for multiple structure alignment and detection of common motifs: application to proteins. *J. Comput. Biol.*, **8**, 93–121.

May,A.C. (2002) Definition of the tempo of sequence diversity across an alignment and automatic identification of sequence motifs: application to protein homologous families and superfamilies. *Protein Sci.*, **11**, 2825–2835.

Mizuguchi,K. and Blundell,T. (2000) Analysis of conservation and substitutions of secondary structure elements within protein superfamilies. *Bioinformatics*, **16**, 1111–1119.

Mizuguchi,K. *et al*. (1998) HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Sci.*, **7**, 2469–2471.

Murzin,A.G. *et al*. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.

Needleman,S. and Wunsch,C. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.

Ochagavia,M.E. and Wodak,S. (2004) Progressive combinatorial algorithm for multiple structural alignments: application to distantly related proteins. *Proteins*, **55**, 436–454.

Orengo,C.A. and Taylor,W.R. (1996) SSAP: sequential structure alignment program for protein structure comparison. *Meth. Enzymol.*, **266**, 617–635.

Ortiz,A.R. *et al*. (2002) MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci.*, **11**, 2606–2621.

Petrey,D. *et al*. (2003) Using multiple structure alignments, fast model building, and energetic analysis in fold recognition and homology modeling. *Proteins*, **53**(Suppl. 6), 430–435.

Ptitsyn,O.B. and Ting,K.L. (1999) Non-functional conserved residues in globins and their possible role as a folding nucleus. *J. Mol. Biol.*, **291**, 671–682.

Raghava,G.P. *et al*. (2003) OXBench: a benchmark for evaluation of protein multiple sequence alignment accuracy. *BMC Bioinformatics*, **4**, 47.

Rossmann,M.G. and Argos,P. (1976) Exploring structural homology of proteins. *J. Mol. Biol.*, **105**, 75–95.

Russell,R.B. and Barton,G.J. (1992) Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins*, **14**, 309–323.

Shatsky,M. *et al*. (2004) A method for simultaneous alignment of multiple protein structures. *Proteins*, **56**, 143–156.

Sheinerman,F.B. *et al*. (2003) Sequence, structure and energetic determinants of phosphopeptide selectivity of SH2 domains. *J. Mol. Biol.*, **334**, 823–841.

Shi,J. *et al*. (2001) FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J. Mol. Biol.*, **310**, 243–257.

Siew,N. *et al*. (2000) MaxSub: an automated measure for the assessment of protein structure prediction quality. *Bioinformatics*, **16**, 776–785.

Sowdhamini,R. *et al*. (1998) CAMPASS: a database of structurally aligned protein superfamilies. *Structure*, **6**, 1087–1094.

Tang,C.L. *et al*. (2003) On the role of structural information in remote homology detection and sequence alignment: new methods using hybrid sequence profiles. *J. Mol. Biol.*, **334**, 1043–1062.

Taylor,W.R. *et al*. (1994) Multiple protein structure alignment. *Protein Sci.*, **3**, 1858–1870.

Thompson,J.D. *et al*. (2001) Towards a reliable objective function for multiple sequence alignments. *J. Mol. Biol.*, **314**, 937–951.

Yang,A.S. and Honig,B. (2000) An integrated approach to the analysis and modeling of protein sequences and structures. III. A comparative study of sequence conservation in protein structural families using multiple structural alignments. *J. Mol. Biol.*, **301**, 691–711.