Biochimie (2012) 94(9):2025-34

Progressive structure-based alignment of homologous proteins: Adopting sequence comparison strategies

Agnel Praveen Joseph [1,2,3,+], Narayanaswamy Srinivasan [4] & Alexandre G. de Brevern [1,2,3,*]

[1] INSERM, UMR-S 665, Dynamique des Structures et Interactions des Macromolécules Biologiques (DSIMB), 6, rue Alexandre Cabanel, 75739 Paris cedex 15, France.
[2] Univ Paris Diderot, Sorbonne Paris Cité, UMR-S665, Paris, F-75739, France.
[3] Institut National de la Transfusion Sanguine (INTS), 6, rue Alexandre Cabanel, 75739 Paris cedex 15, France.
[4] Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560012, India.

Short title: MULtiple Protein Block Alignment

* Corresponding author:
mailing address: de Brevern A.G., INSERM UMR-S 665, Dynamique des Structures et Interactions des Macromolécules Biologiques (DSIMB), Université Denis Diderot - Paris 7, INTS, 6, rue Alexandre Cabanel, 75739 Paris cedex 15, France
E-mail: alexandre.debrevern@univ-paris-diderot.fr
Tel: +33(1) 44 49 31 14
Fax: +33(1) 47 34 74 31

+ Current address: National Centre for Biological Sciences, TIFR, Bellary Road, Bangalore 560065, India

Key words : protein structure comparison; structural alphabet; Protein Blocks; anchor-based alignment; protein structure mining; Protein Data Bank; structure conservation.

Abstract

Comparison of multiple protein structures has a broad range of applications in the analysis of protein structure, function and evolution. Multiple structure alignment tools (MSTAs) are necessary to obtain a simultaneous comparison of a family of related folds. In this study, we have developed a method for multiple structure comparison largely based on sequence alignment techniques. A widely used Structural Alphabet named Protein Blocks (PBs) was used to transform the information on 3D protein backbone conformation as a 1D sequence string.

A progressive alignment strategy similar to CLUSTALW was adopted for multiple PB sequence alignment (mulPBA). Highly similar stretches identified by the pairwise alignments are given higher weights during the alignment. The residue equivalences from PB based alignments are used to obtain a three dimensional fit of the structures followed by an iterative refinement of the structural superposition.

Systematic comparisons using benchmark datasets of MSTAs underlines that the alignment quality is better than MULTIPROT, MUSTANG and the alignments in HOMSTRAD, in more than 85% of the cases. Comparison with other rigid-body and flexible MSTAs also indicate that mulPBA alignments are superior to most of the rigid-body MSTAs and highly comparable to the flexible alignment methods.

1. Introduction

The three dimensional structure of a protein provides tremendous insights on its function [1]. It has been an essential requirement to compare protein structures for the interpretation of functional, dynamic and evolutionary properties. To study the relative structural variations among a group of structures, a simultaneous comparison is required for which multiple structural alignments (MSTA) are relevant. Multiple structure comparisons are also essential components of many modeling and threading procedures [2-4].

Superposition of 3D protein structures can be solely obtained by global translational and rotational searches, however it is not trivial. To achieve this goal, one set of methods uses the representation of structures as contact patterns or distance matrices. These matrices are then compared to obtain the 3D equivalences from which an alignment can be generated [5-7]. Another commonly used strategy is to identify an initial conformational equivalence and then carry out refinements to generate an alignment. Earlier methods used simple sequence alignment algorithms to obtain the initial structural equivalence [8], arbitrary equivalences were also used as starting points [9]. Nonetheless, these alignments tend to be faulty when the sequence similarity is low. Thus the newer approaches are mostly structure based and they derive the initial equivalences by detecting similarities in the local structural regions. The description of 3D structures as a series of secondary structures (helix/strand/coil) provides one such means for comparison. [10-12].

Another set of powerful methods not relying on the secondary structure representation, are based on comparison of local backbone fragments. The length of these fragments is either predefined [13-16] or constrained by a measure for structural similarity between fragments [17, 18]. The most common and efficient technique for sequence order dependant comparisons, is dynamic programming [9, 13, 15, 17, 19]. Few other approaches are not dependant on the order of protein fragments, the structural similarity and the relative

orientation of these fragments being the major constraint for comparison. They are mainly based on algorithms like geometric hashing [10, 18, 20, 21], Monte Carlo Optimization [5], graph-matching [11] etc. These methods detect relationships based on sequence permutations and recombinations. However consistent results are not often obtained in the detection of new relationships and for the assignment of equivalent structural regions.

The inherent flexibility of protein 3D structures supports its biological function by accommodating conformational variations. Hence the structural alignment techniques that cannot detect these flexible movements tend to misinterpret the extent of similarity. Methods capable of performing flexible structure comparison are also developed [15, 19]. They work by identifying fragment similarities followed by the detection of twists and hinge movements. However the discrimination of true hinge movements from acquired structural changes is sometimes difficult and subtle movements can be left unrecognized.

The fragment based approaches mentioned above do not require *a priori* knowledge of the conformation of the fragments. A more recent group of methods attempt to classify local protein structures into a limited set of local backbone conformations before carrying out comparisons. These methods are based on libraries of local backbone structures that represent the frequently occurring regular backbone conformations. With the premise that the secondary structure description in terms of α-helix and β-strands covers only about 50% of all local conformations, several studies attempted to characterize most or all of the backbone structure [22, 23]. A library of local backbone conformations that can be used to abstract a complete protein backbone is called as a Structural Alphabet (SA). Abstraction of structures in terms of SA helps to encode 3D information into a 1D sequence. Hence the comparison of 3D structures can be performed using an alignment of sequence of SAs. Classical amino acid sequence alignment strategies can be adopted for this. A few methods have been developed for comparing protein structures based on structural alphabets (*e.g.*, [24-29]). When compared

to the methods based on similarity of 3D structural measures, these approaches are significantly faster.

Unlike pairwise comparisons involving large database searches which require both speed and efficiency, MSTAs rather focus on accuracy. Multiple structural alignments are sensitive to the number of structures compared and their relative similarity. Most of the MSTAs use the residue equivalences from the pairwise alignments to obtain a fit of the structures. One simple approach is a center-star method where one structure is used as the reference and others are aligned to the reference based on the pairwise alignments [2]. To avoid the loss of information on the relative structural similarities depending on the choice of reference, an average or consensus template is used as reference [16, 30]. Majority of MSTA methods use a progressive alignment strategy to derive a multiple alignment [7, 8, 15]. A guide tree generated based on a relative similarity measure determines the order in which structures are added to the alignment. To reduce the bias dependant on this order of fit, iterative refinements are carried out [14, 16, 17, 31]. Consistency of residue equivalences among the pairwise alignments is also learned to refine the multiple alignment [12, 32]. Another set of methods compute a simultaneous comparison of all structures without using a hierarchical procedure or a reference structure [10, 18, 20]. The latter group works by using techniques like geometric hashing [33] to identify the set of structural regions common for all structures. This helps to overcome the inherent limitation of progressive fit of structures where the optimal alignment could not be derived from multiple pairwise comparisons.

A widely used SA, named Protein Blocks (PBs) [22, 34-37], was used to develop an efficient method for comparing two protein structures [38]. The structures were translated into PB sequences followed by the alignment of the PB sequences. Classical Needleman-Wunsch dynamic programming [39] was used and a dedicated PB substitution matrix was generated for scoring the alignment [40]. Significant improvement in the alignment quality could be

achieved with the use of an anchor-based dynamic programming algorithm. It first identifies all high scoring and structurally favorable local alignments (anchors) and then aligns the segments between them to obtain a global alignment. This improved PB based structure alignment approach (iPBA) outperformed other established methods when tested on benchmark datasets [41, 42].

In this study we extend the iPBA approach to the comparison of multiple structures. A progressive strategy similar to that used in CLUSTALW [43] was adopted. PB sequence alignment determines the residue equivalences for the 3D structural fit and the fitted structures are optimized by structure based iterative refinements. To assess the performance of our approach, the alignments were compared to that generated with other popular methods.

2. Methods

2.1. Protein Blocks

Protein Blocks (PBs) correspond to a set of 16 local prototypes, labeled from *a* to *p*, of 5 residues length described based on the Φ, Ψ dihedral angles. They were obtained by an unsupervised classifier similar to Kohonen Maps [44] and hidden Markov models [45]. This structural alphabet allows a reasonable approximation of local protein 3D structures [34] with a root mean square deviation (*rmsd*) recently evaluated to be 0.42 Å [46]. PBs [46] have been assigned using in-house Python software as in the previous studies [41, 42].

2.2. mulPBA Methodology

Figure 1 gives an outline of the steps involved in mulPBA alignment approach.

2.2.1. Pairwise alignments: The protein structures to be aligned are first translated into PB sequences. The pairwise alignments are obtained using iPBA which performs an

anchor based alignment by finding structurally conserved regions, identified as local alignments [41, 42]. A combination of local [47] and global [39] dynamic programming algorithms is used for the alignment. A set of local alignments (anchors) associated with these two sequences is derived using a modified version of SIM algorithm [42, 47]. The remaining segments between anchors (linkers) are then aligned with relaxed gap penalties, using Needleman-Wunsch algorithm. The PB substitution matrix was generated using substitution frequencies obtained from alignments of domain pairs in PALI [48] with no more than 40% sequence identity [41, 42].

2.2.2. Structure relatedness: A progressive multiple sequence alignment strategy similar to CLUSTALW [43] was used. The PB identities calculated from pairwise alignments were translated into a distance matrix (see Figure 1b). The matrix was then used to generate a guide tree (Figure 1c) [49]. The tree root was identified by mid-point rooting method [43]. Each sequence was assigned a weight depending on the distance from the root. It reduces the bias due to variation in the extent of similarity between the sequences.

2.2.3. Progressive alignment: The tree was used to guide the assembly of sequences based on the degree of similarity, to form the multiple alignment. The alignment of two sequences (or groups of sequences) is carried out using dynamic programming. The average of pairwise PB substitution scores (from the substitution matrix) was used to calculate the score for aligning an element (alignment column) of a sequence group against an element (alignment column) of another. These scores S were weighted using sequence weights obtained from the guide tree. While aligning two profiles P1 and P2 of sizes *k* and *l*, the score for substituting a column *i* of P1 with column *j of* P2 is given by:

$$S_{i,j} = \frac{\sum_{p=1}^{k}\sum_{q=1}^{l}[(sub(PB_{i,p}, PB_{j,q}) * seq\_weight_p * seq\_weight_q) + anchor\_weight]}{k*l}$$

seq_weight_p and seq_weight_q indicates the sequence weights assigned based on the guide tree, to the sequence corresponding to p and q respectively.

From each pairwise alignment used for obtaining the guide tree, the positions corresponding to the alignments in the structurally similar regions (anchors), were stored. These positions were then assigned a weight, namely *anchor_weight*, which is calculated as:

$$anchor\_weight = 250*(1+cov)$$

where *cov* indicates the percentage coverage of the anchor with respect to the alignment length. The value 250 was optimized from an assessment of alignments generated (see below).

2.3. Benchmark datasets of structural alignments.

(i) HOMSTRAD database of structural alignments is commonly used as a benchmark for MSTAs [50]. The structures grouped as a family are aligned using Comparer [51, 52], Mnyfit [53] or Stamp [8] and the results are curated manually. 330 protein families with more than 2 members were used for parameter optimization and assessment of mulPBA.

(ii) The recent version of PALI dataset V 2.8a [48, 54] consists of 1,922 domain families comprising of 231,022 alignment pairs. Structural alignments in this version are generated using MUSTANG [17]. A subset of 200 domain families was chosen randomly from the dataset for optimizing parameters and assessment.

2.4. Multiple alignment scores: Different kinds of scores mainly derived from earlier works were employed, as it is not simple to design an optimal score and a universally accepted measure is not available. These alignment quality measures are computed on the 3D

fit of the structures (by PROFIT) based on the PB sequence alignment.

(i) *Quantifying the quality of alignment core*: The preliminary criteria for considering an alignment column as part of the core is that it should have less than 30% of elements as gaps. In addition, two different definitions were employed for the alignment core:

(a) The maximum distance between any two residues at an aligned position should be less than a given cut-off. The number of alignment columns where the residues are within this cut-off distance, are counted. A weighted average of the number of columns associated with the distance cut-offs of 3.0Å, 4.0Å, 5.0Å and 6.0Å was calculated in a similar way as that of GDT score [55, 56]. This measure is termed as N_{dist}. A fixed distance cut-off of 4Å was used to assess MAMMOTH-MULT [31].

(b) The *rmsd* of an aligned position (column) should be less than 3.0Å. The number of columns with less than 30% gaps and rmsd within 3.0Å, thus constitutes the core. This score is termed as N_{rms}. A similar score was used to assess the performance of MULTIPROT [18]

(ii) *Quantifying the global alignment quality*: For each combination of pairwise alignment extracted from the multiple alignment, the number of aligned residue pairs that are within a distance of 3.5Å was counted. An average was then calculated for all the pairwise alignments, this score was named as $N_{3.5}$. A similar score was used to quantify the quality of SALIGN [57] alignments.

2.5. 3D structural alignment. PROFIT (version 3.1) [58] performs least squares fit of protein structures based on the residue equivalences in a given sequence alignment. The multiple PB sequence alignment is translated to amino acid sequence alignment which is given as input for PROFIT. The structure that has the least overall *rmsd* with the other

structures is chosen as the reference and the most similar structures are fitted in a progressive manner. At each step, the reference template is updated with the averaged coordinates of the superposed structures followed by the superposition of a new structure from the list. PROFIT can also perform a refinement of the fit based on an iterative update of the aligned residues within a given distance (5.0Å).

2.6. Structure Based Sequence Alignment. The 3D superposition obtained using PROFIT [58] was translated to a multiple sequence alignment using dynamic programming scored based on inter Cα distances. A low additional weight is also added based on the substitution score for residue type substitution. The amino acid substitution matrix is generated from alignments in PALI [48] database in a manner similar to that used for obtaining PB substitution matrix.

3. Results

This study is aimed at developing a method for comparison of multiple protein structures based on the 1D representation of backbone conformation. The backbone conformation of the protein chains are first abstracted in terms of PBs. Pairwise comparisons were carried out using the iPBA approach, which uses anchor-based dynamic programming for alignment. A progressive alignment strategy comparable to CLUSTALW [43] was then adopted for generating multiple PB sequence alignments. Three different scores namely N_{rms}, N_{dist} and $N_{3.5}$ were used to evaluate the quality of both the alignment core and the global superposition. A crucial parameter is the choice of the gap opening and extension penalties required for the progressive alignment. The performance with different penalty values were analysed on a dataset of 100 families from the PALI database [48] (Supplementary data 1). The N_{dist} score was used to select the best alignment. The best alignments were spread across

different gap penalty values and no single set of penalties clearly outperforms others. At a gap opening penalty of -800 and extension penalty of -400, maximum number of alignments had high scores. With this set of penalties, about 80% of the alignments had scores close to (difference less than 3) the maximum N_{dist} score for the alignment. The structurally conserved regions are obtained as anchors in the pairwise alignments. In the course of the progressive alignment, the anchor regions were weighed to improve the accuracy. With the above set of penalties and an anchor weight of 250, best performance could be observed (Supplementary data 1).

The performance of our approach (mulPBA) was extensively assessed against the alignments in HOMSTRAD [50], the reference dataset used to assess the performance of different alignment methods [18, 31]. The results of this comparison clearly show that the alignments are improved significantly with mulPBA approach. Out of the 332 alignments compared, 80.1 % had a higher N_{rms} score, 86.4% had a higher N_{dist} score and 87.7% were better in terms of the $N_{3.5}$ (Figure 2A). On an average, about 84.7% gain in alignment quality was obtained across the different measures. On the same dataset of 332 alignments, the alignment quality was compared with that obtained with MULTIPROT [18]. STACCATO [59] was used to generate sequence alignments corresponding to the structure superposition obtained with MULTIPROT. STACCATO considers amino acid substitution weights while generating sequence alignments based on superposed structures and it is often used to obtain the equivalences from MULTIPROT alignments [31]. With respect to MULTIPROT, the average gain based on the different scores was about 86.6% (N_{rms}: 82.4, N_{dist}: 87.3, $N_{3.5}$: 90.1) (Figure 2B). However in most cases, the extent of improvement is not as high as that observed in comparison with HOMSTRAD. The increase in quality with respect to HOMSTRAD and MULTIPROT is largely noticed as improvement in matching residue equivalences and better backbone fit (Supplementary data 2).

The recent version of PALI [48] holds multiple alignments of domain families classified based on SCOP [5] definitions. These alignments are generated using MUSTANG [17], a highly popular MSTA. A subset of 200 family alignments from PALI database was used to compare the quality of alignments with that of MUSTANG. An average gain of about 86.7% ($N_{rms:}$ 78.5, N_{dist}: 89.8 and $N_{3.5}$: 91.9) could be achieved. For comparing the alignment quality on proteins with more than one domain, multi-domain proteins in PALI with more than 2 members sharing less than 40% sequence identity were chosen. Out of 12 families obtained, mulPBA was better than MUSTANG in 9 cases in terms of all the three scores while 8 cases were better when compared to MULTIPROT.

The alignment quality was also compared with other common MSTAs [10, 15, 57] implemented as web-servers (Table 1). The test is performed on a more limited dataset comprising of 12 protein families which were either studied in previous works [8, 10, 15, 17] or chosen from the SABmark dataset [60]. The alignment quality is quantified in terms of N_{rms}, N_{dist} and $N_{3.5}$. The servers used for comparison involve both rigid-body alignment tools and flexible MSTAs. As highlighted in Table 1, for 8 out of the 12 families mulPBA gives the best quality alignments when compared to the rigid-body alignment methods SALIGN [57], MAMMOTH [31] and MASS [10]. Both SALIGN and MASS had the top scores for the alignment of two families. The quality scores of SALIGN alignments for these families (Serine protease and Mettalo-hydrolase) were however close to mulPBA alignment scores.

Comparison with flexible alignments also gave convincing results. Among the set of MSTAs used for comparison, POSA [15] and MATT [19] are flexible aligners. They detect hinge regions involving flexible movements and introduce bends at these points to maximize the extent of superposition. As expected, for cases with flexible movements in the structures, these methods produce alignments with higher number of structural equivalence and they are supposed to give always better results than rigid-body approaches. Hence a direct comparison

between flexible and rigid-body aligners cannot be deciphered easily. Overall, POSA [15] generates the best quality alignments, but in 5 cases on 12 (Cupins, Globins, Serine Protease, Rossmann and Gamma Crystallin), mulPBA was more efficient, which was quite striking. It was also noted that POSA introduced false bends and twists in 2 alignments (Cupins and β Superhelix families, See *Discussions*), where only the average pairwise alignment score ($N_{3.5}$) was better. Also, mulPBA was always better than MATT in the majority of alignments.

4. Discussion

The PB based structure approximation enable a 'sequence like' approach for structure comparison. Combination of local and global dynamic programming algorithms (anchor based) led to the development of an efficient structure comparison tool superior to many other popular methods. Substitutions corresponding to the anchor regions are given higher weights in the process of progressive alignment. A similar strategy is also applied in DBCLUSTAL [61] which is a major improvement over CLUSTALW [43]. Addition of anchor weights as a soft constraint results in a significant increase in the alignment quality (Supplementary data 1).

As seen above, the quality of alignments generated by mulPBA is much better than many other established MSTAs. Figure 3 also shows the comparison of quality of alignments of 5 structures from the Cupin family. Cupin fold is known for its functional diversity marked with variations in the active site residues [62]. The alignments generated by mulPBA (Figure 3A) and MAMMOTH had the best quality scores (Table 1). The histidine residues that interact with metal ion, which is characteristic of many cupins, occur at equivalent positions in the mulPBA alignment. SALIGN [57] fails in superposing one of the cupin structures in the correct orientation (Figure 3B) while MASS [10] misaligns two structures (Figure 3C). The alignment generated by MATT [19] is similar to that obtained with mulPBA. However, the

quality of the fit is lower (Figure 3D) and no flexible movements were considered. The POSA superposition has the highest $N_{3.5}$ score (which gives the average number of equivalences from the underlying pairwise alignments) and several hinge regions that mediate flexible movements were detected (Figure 3E). However, the alignment generated by fitting the detected flexible regions, is marked with distortions in the largely stable beta sheet core of the crystal structures and also with the superposition of non equivalent structural regions (Supplementary data 3). Hence the alignment is quite unrealistic. Also, one of the cupin structures (shown in blue in the figure) is not fitted in the correct orientation (Figure 3E), the equivalent beta strands were not identified.

Representation of backbone conformation in terms of Protein Blocks enables the use of sequence alignment approaches for structure comparisons. In cases where the structures involve large flexible movements, Protein Block alignments were found to detect structural equivalences involving rigid body movements. These are cases where the sequence alignment holds more relevant information that a structure based fit. However, the 3D structure superposition generated by mulPBA is derived by carrying out structure based refinements on the PB based alignment. The result is a rigid-body fit of the structures and hence mulPBA could fall behind the flexible alignment methods like POSA [15] (Table 1).

One such case is the alignment of tRNA synthetase structures (Table 1 & Figure 4A) which involves flexible movements between two domains. In a trial involving the comparison of four tRNA synthetase structures, the performance of mulPBA is behind POSA and MASS. One of the four structures involves a conformational shift in relative orientation of the domains (Figure 4A). The equivalent residues were correctly matched in the PB sequence alignment (Figure 4B). Nevertheless, the residues cannot be matched simultaneously in 3D which results in an alignment where both domains of this structure (with flexible movements) are not fitted well with respect to the other structures (Figure 4C). Hence the final quality

scores are also lower. MASS is successful in matching one of the two domains (Figure 5A). Though MATT [19] is a flexible alignment method, an optimal superposition was not obtained (Figure 5B) while POSA introduces a twist at the flexible loop to generate a good fit of both domains (Figure 5C).

In the example above, the inherent flexibility of the structures resulted in a poor 3D fit. However in the more usual scenario, the larger region of equivalence is fitted upon refinement of the structural fit. When compared to PALI and HOMSTRAD datasets, only 2% of the alignments generated with mulPBA had large and significant decline in the alignment quality. In a few of these cases, the structures involve long and multiple helices. Hence the PB sequences are characterized by long stretches of low complexity (series of PB '*m*') and this can result in wrong residue equivalences in the alignment. Though this problem is largely taken care of with the addition of amino acid substitution weights [42], a few cases of bad equivalences are still encountered at very low sequence identities. Comparison of proteins in the Ferritin family is a one such case where the N_{dist} score is quite low (Table 1), even though the global fit is similar to the alignment generated by MATT. Rarely, wrong anchors are also chosen in a pairwise alignment, which can result in a poor multiple alignment. In the future, we will be assessing more strict constraints for the choice of anchor regions for alignment.

5. Conclusions

The approximation of backbone conformation of protein chain as a PB sequence, enable a sequence-like mode of structure comparison. The progressive mode of alignment coupled with anchor weights and iterative refinements of structural fit result in high quality alignment. The performance of mulPBA is also better than many other popular MSTAs. The sequence-like mode of alignment further adds interest in comparing flexible regions in the

structure. The efficiency of mulPBA also reflects the high relevance in backbone approximation using PBs.

6. Acknowledgements

These works were supported by grants from the Ministry of Research (France), University of Paris Diderot, National Institute for Blood Transfusion (INTS, France), Institute for Health and Medical Research (INSERM, France) and Indian Department of Biotechnology. APJ is supported by CEFIPRA number 3903-E. NS and AdB also acknowledge to CEFIPRA for collaborative grant (number 3903-E).

7. References

[1] D. Baker, A. Sali, Protein structure prediction and structural genomics, Science 294 (2001) 93-96.
[2] T. Akutsu, K.L. Sim, Protein Threading Based on Multiple Protein Structure Alignment, Genome Inform Ser Workshop Genome Inform 10 (1999) 23-29.
[3] R.L. Dunbrack, Jr., Sequence comparison and protein structure prediction, Curr Opin Struct Biol 16 (2006) 374-384.
[4] A. Panchenko, A. Marchler-Bauer, S.H. Bryant, Threading with explicit models for evolutionary conservation of structure and sequence, Proteins Suppl 3 (1999) 133-140.
[5] L. Holm, C. Sander, Protein structure comparison by alignment of distance matrices, J Mol Biol 233 (1993) 123-138.
[6] F. Birzele, J.E. Gewehr, G. Csaba, R. Zimmer, Vorolign--fast structural alignment using Voronoi contacts, Bioinformatics 23 (2007) e205-211.
[7] W.R. Taylor, T.P. Flores, C.A. Orengo, Multiple protein structure alignment, Protein Sci 3 (1994) 1858-1870.
[8] R.B. Russell, G.J. Barton, Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels, Proteins 14 (1992) 309-323.
[9] S. Subbiah, D.V. Laurents, M. Levitt, Structural similarity of DNA-binding domains of bacteriophage repressors and the globin core, Curr Biol 3 (1993) 141-148.
[10] O. Dror, H. Benyamini, R. Nussinov, H. Wolfson, MASS: multiple structural alignment by secondary structures, Bioinformatics 19 Suppl 1 (2003) i95-104.
[11] E. Krissinel, K. Henrick, Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions, Acta Crystallogr D Biol Crystallogr 60 (2004) 2256-2268.
[12] J. Ebert, D. Brutlag, Development and validation of a consistency based multiple structure alignment algorithm, Bioinformatics 22 (2006) 1080-1087.
[13] A.R. Ortiz, C.E. Strauss, O. Olmea, MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison, Protein Sci 11 (2002) 2606-2621.
[14] I.N. Shindyalov, P.E. Bourne, Protein structure alignment by incremental combinatorial extension (CE) of the optimal path, Protein Eng 11 (1998) 739-747.
[15] Y. Ye, A. Godzik, Multiple flexible structure alignment using partial order graphs, Bioinformatics 21 (2005) 2362-2369.
[16] I. Ilinkin, J. Ye, R. Janardan, Multiple structure alignment and consensus identification for proteins, BMC Bioinformatics 11 (2010) 71.
[17] A.S. Konagurthu, J.C. Whisstock, P.J. Stuckey, A.M. Lesk, MUSTANG: a multiple structural alignment algorithm, Proteins 64 (2006) 559-574.
[18] M. Shatsky, R. Nussinov, H.J. Wolfson, A method for simultaneous alignment of multiple protein

structures, Proteins 56 (2004) 143-156.
[19] M. Menke, B. Berger, L. Cowen, Matt: local flexibility aids protein multiple structure alignment, PLoS Comput Biol 4 (2008) e10.
[20] N. Leibowitz, R. Nussinov, H.J. Wolfson, MUSTA--a general, efficient, automated method for multiple structure alignment and detection of common motifs: application to proteins, J Comput Biol 8 (2001) 93-121.
[21] V.A. Ilyin, A. Abyzov, C.M. Leslin, Structural alignment of proteins by a novel TOPOFIT method, as a superimposition of common volumes at a topomax point, Protein Sci 13 (2004) 1865-1874.
[22] A.P. Joseph, G. Agarwal, S. Mahajan, J.-C. Gelly, L.S. Swapna, B. Offmann, F. Cadet, A. Bornot, M. Tyagi, H. Valadié, B. Schneider, F. Cadet, N. Srinivasan, A.G. de Brevern, A short survey on Protein Blocks, Biophysical Reviews 2 (2010) 137-145.
[23] B. Offmann, M. Tyagi, A.G. de Brevern, Local Protein Structures, Current Bioinformatics 3 (2007) 165-202.
[24] I. Friedberg, T. Harder, R. Kolodny, E. Sitbon, Z. Li, A. Godzik, Using an alignment of fragment strings for comparing protein structures, Bioinformatics 23 (2007) e219-224.
[25] F. Guyon, A.C. Camproux, J. Hochez, P. Tuffery, SA-Search: a web tool for protein structure mining based on a Structural Alphabet, Nucleic Acids Res 32 (2004) W545-548.
[26] S.Y. Ku, Y.J. Hu, Protein structure search and local structure characterization, BMC Bioinformatics 9 (2008) 349.
[27] C.H. Tung, J.W. Huang, J.M. Yang, Kappa-alpha plot derived structural alphabet and BLOSUM-like substitution matrix for rapid search of protein structure database, Genome Biol 8 (2007) R31.
[28] S. Wang, W.M. Zheng, CLePAPS: fast pair alignment of protein structures based on conformational letters, J Bioinform Comput Biol 6 (2008) 347-366.
[29] J. Yang, Comprehensive description of protein structures using protein folding shape code, Proteins 71 (2008) 1497-1518.
[30] M. Gerstein, M. Levitt, Using iterative dynamic programming to obtain accurate pairwise and multiple alignments of protein structures, Proc Int Conf Intell Syst Mol Biol 4 (1996) 59-67.
[31] D. Lupyan, A. Leo-Macias, A.R. Ortiz, A new progressive-iterative algorithm for multiple structure alignment, Bioinformatics 21 (2005) 3255-3263.
[32] E. Sandelin, Extracting multiple structural alignments from pairwise alignments: a comparison of a rigorous and a heuristic approach, Bioinformatics 21 (2005) 1002-1009.
[33] R. Nussinov, H.J. Wolfson, Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques, Proc Natl Acad Sci U S A 88 (1991) 10495-10499.
[34] A.G. de Brevern, C. Etchebest, S. Hazout, Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks, Proteins 41 (2000) 271-287.
[35] O. Zimmermann, U.H. Hansmann, LOCUSTRA: accurate prediction of local protein structure using a two-layer support vector machine approach, J Chem Inf Model 48 (2008) 1903-1908.
[36] M. Dudev, C. Lim, Discovering structural motifs using a structural alphabet: application to magnesium-binding sites, BMC Bioinformatics 8 (2007) 106.
[37] H. Rangwala, C. Kauffman, G. Karypis, svmPRAT: SVM-based protein residue annotation toolkit, BMC Bioinformatics 10 (2009) 439.
[38] M. Tyagi, A.G. de Brevern, N. Srinivasan, B. Offmann, Protein structure mining using a structural alphabet, Proteins 71 (2008) 920-937.
[39] S.B. Needleman, C.D. Wunsch, A general method applicable to the search for similarities in the amino acid sequence of two proteins, J Mol Biol 48 (1970) 443-453.
[40] M. Tyagi, V.S. Gowri, N. Srinivasan, A.G. de Brevern, B. Offmann, A substitution matrix for structural alphabet based on structural alignment of homologous proteins and its applications, Proteins 65 (2006) 32-39.
[41] J.C. Gelly, A.P. Joseph, N. Srinivasan, A.G. de Brevern, iPBA: a tool for protein structure comparison using sequence alignment strategies, Nucleic Acids Res 39 (2011) W18-23.
[42] A.P. Joseph, N. Srinivasan, A.G. de Brevern, Improvement of protein structure comparison using a structural alphabet, Biochimie 93 (2011) 1434-1445.
[43] J.D. Thompson, D.G. Higgins, T.J. Gibson, CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, Nucleic Acids Res 22 (1994) 4673-4680.
[44] T. Kohonen, Self-Organizing Maps (3rd edition), Springer, 2001, 501 p.
[45] L.R. Rabiner, A tutorial on hidden Markov models and selected application in speech recognition, Proceedings of the IEEE 77 (1989) 257-286.
[46] A.G. de Brevern, New assessment of a structural alphabet, In Silico Biol 5 (2005) 283-289.
[47] X. Huang, Miller, W, A time-efficient linear-space local similarity algorithm, Advances in Applied Mathematics 12 (1991) 337 - 357.
[48] S. Balaji, S. Sujatha, S.S. Kumar, N. Srinivasan, PALI-a database of Phylogeny and ALIgnment of

homologous protein structures, Nucleic Acids Res 29 (2001) 61-65.
[49] N. Saitou, M. Nei, The neighbor-joining method: a new method for reconstructing phylogenetic trees, Mol Biol Evol 4 (1987) 406-425.
[50] K. Mizuguchi, C.M. Deane, T.L. Blundell, J.P. Overington, HOMSTRAD: a database of protein structure alignments for homologous families, Protein Sci 7 (1998) 2469-2471.
[51] A. Sali, T.L. Blundell, Definition of general topological equivalence in protein structures. A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming, J Mol Biol 212 (1990) 403-428.
[52] Z.Y. Zhu, A. Sali, T.L. Blundell, A variable gap penalty function and feature weights for protein 3-D structure comparisons, Protein Eng 5 (1992) 43-51.
[53] M.J. Sutcliffe, I. Haneef, D. Carney, T.L. Blundell, Knowledge based modelling of homologous proteins, Part I: Three-dimensional frameworks derived from the simultaneous superposition of multiple structures, Protein Eng 1 (1987) 377-384.
[54] V.S. Gowri, S.B. Pandit, P.S. Karthik, N. Srinivasan, S. Balaji, Integration of related sequences with protein three-dimensional structural families in an updated version of PALI database, Nucleic Acids Res 31 (2003) 486-488.
[55] A. Zemla, LGA: A method for finding 3D similarities in protein structures, Nucleic Acids Res 31 (2003) 3370-3374.
[56] A. Zemla, B. Geisbrecht, J. Smith, M. Lam, B. Kirkpatrick, M. Wagner, T. Slezak, C.E. Zhou, STRALCP--structure alignment-based clustering of proteins, Nucleic Acids Res 35 (2007) e150.
[57] M.S. Madhusudhan, B.M. Webb, M.A. Marti-Renom, N. Eswar, A. Sali, Alignment of multiple protein structures based on sequence and structure features, Protein Eng Des Sel 22 (2009) 569-574.
[58] A. Martin, C. Porter, http://www.bioinf.org.uk/software/profit/, 2010.
[59] M. Shatsky, R. Nussinov, H.J. Wolfson, Optimization of multiple-sequence alignment based on multiple-structure alignment, Proteins 62 (2006) 209-217.
[60] I. Van Walle, I. Lasters, L. Wyns, SABmark--a benchmark for sequence alignment that covers the entire known fold space, Bioinformatics 21 (2005) 1267-1268.
[61] J.D. Thompson, F. Plewniak, J. Thierry, O. Poch, DbClustal: rapid and reliable global multiple alignments of protein sequences detected by database searches, Nucleic Acids Res 28 (2000) 2919-2926.
[62] G. Agarwal, M. Rajavel, B. Gopal, N. Srinivasan, Structure-based phylogeny as a diagnostic for functional characterization of proteins with a cupin fold, PLoS One 4 (2009) e5736.
[63] The PyMOL Molecular Graphics System, Schrödinger, LLC, p. Version 1.2r3pre.

Figure legends

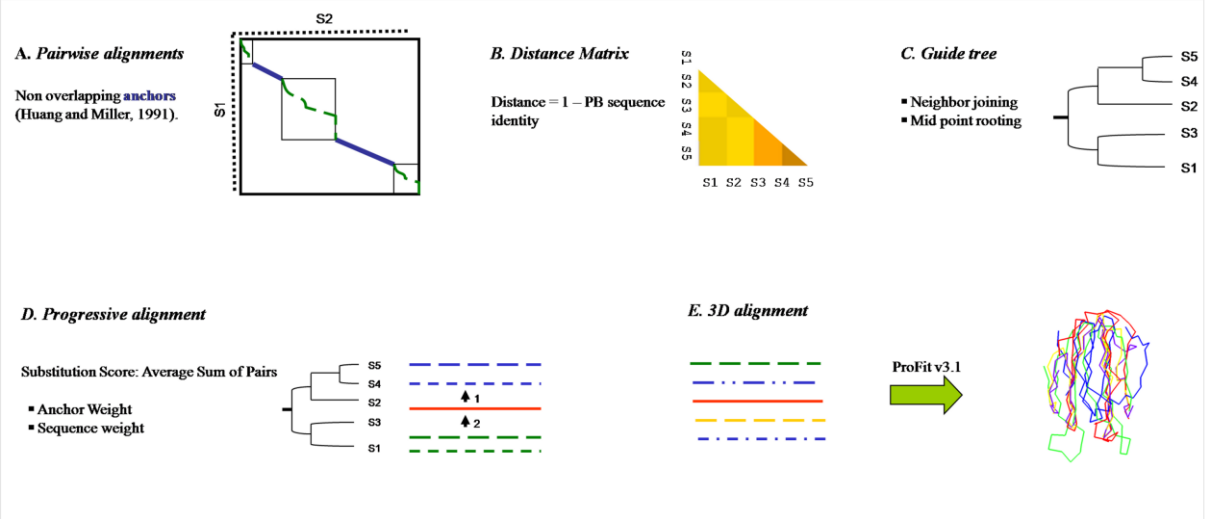

Figure 1. *The alignment approach behind mulPBA.* (A) An optimal set of anchors are identified using SIM algorithm [47]. The intervening segments are aligned using Needleman-Wunsch algorithm [39] (B) The PB sequence identities calculated from pairwise alignments, are used to generate a distance matrix (C) A guide tree is obtained from the distance matrix using the Neighbour Joining algorithm [49] (D) The guide tree determines the progressive manner of alignment of PB sequences (E) The residue equivalences from the multiple PB sequence alignment is translated into a 3D fit using ProFit [53].

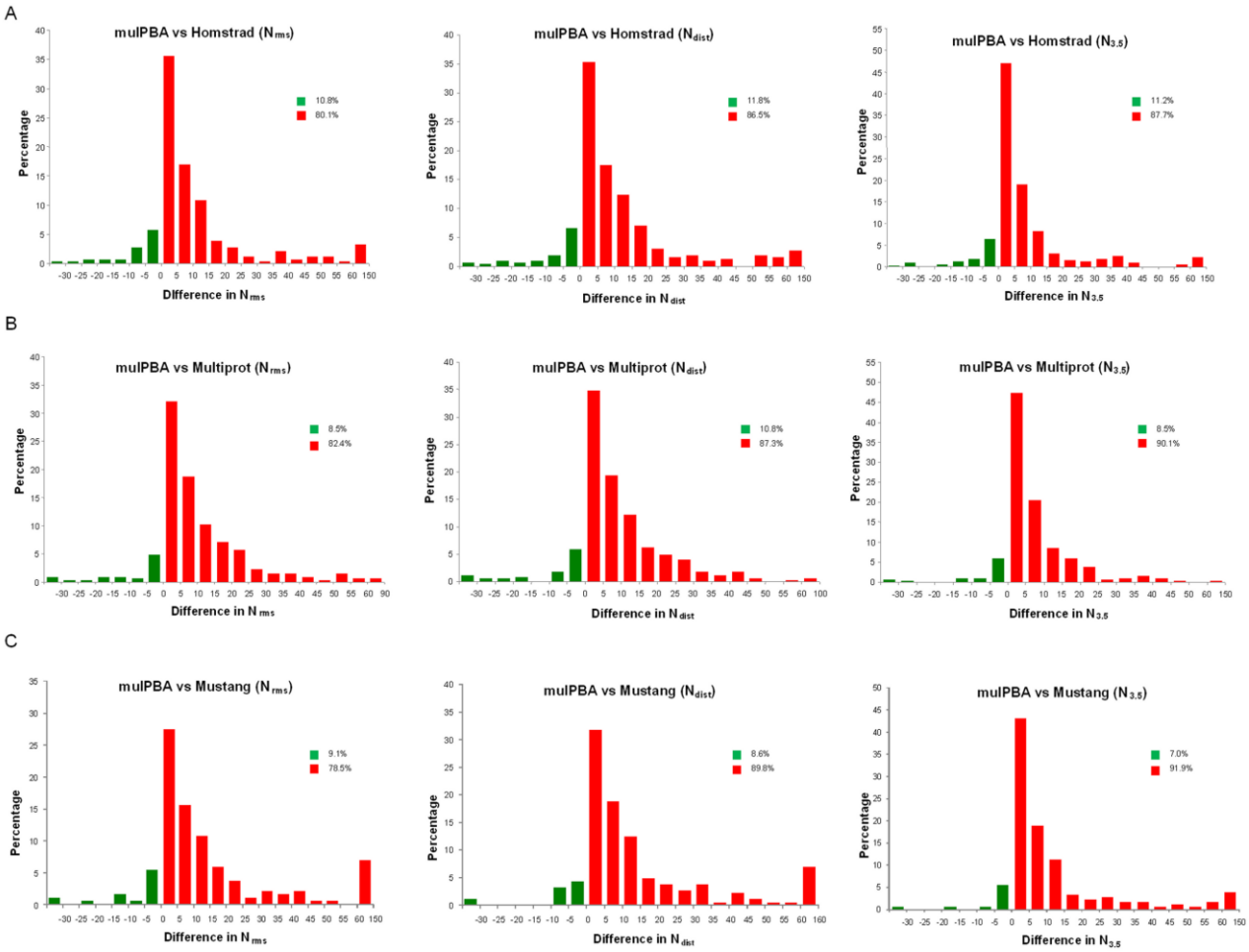

Figure 2. *Comparison of mulPBA alignments with HOMSTRAD, MULTIPROT and MUSTANG*. The alignment quality is quantified in terms of N_{rms}, N_{dist} and $N_{3.5}$ (*see Methods*). The difference in these scores with respect to the three alignment methods and the corresponding percentage of alignments are plotted. The alignments with better scores (positive difference) are highlighted in red while the negative cases are in green. The total percentage of positive and negative cases is also indicated. Panels A, B and C give the results of comparison with HOMSTRAD, MULTIPROT and MUSTANG respectively.

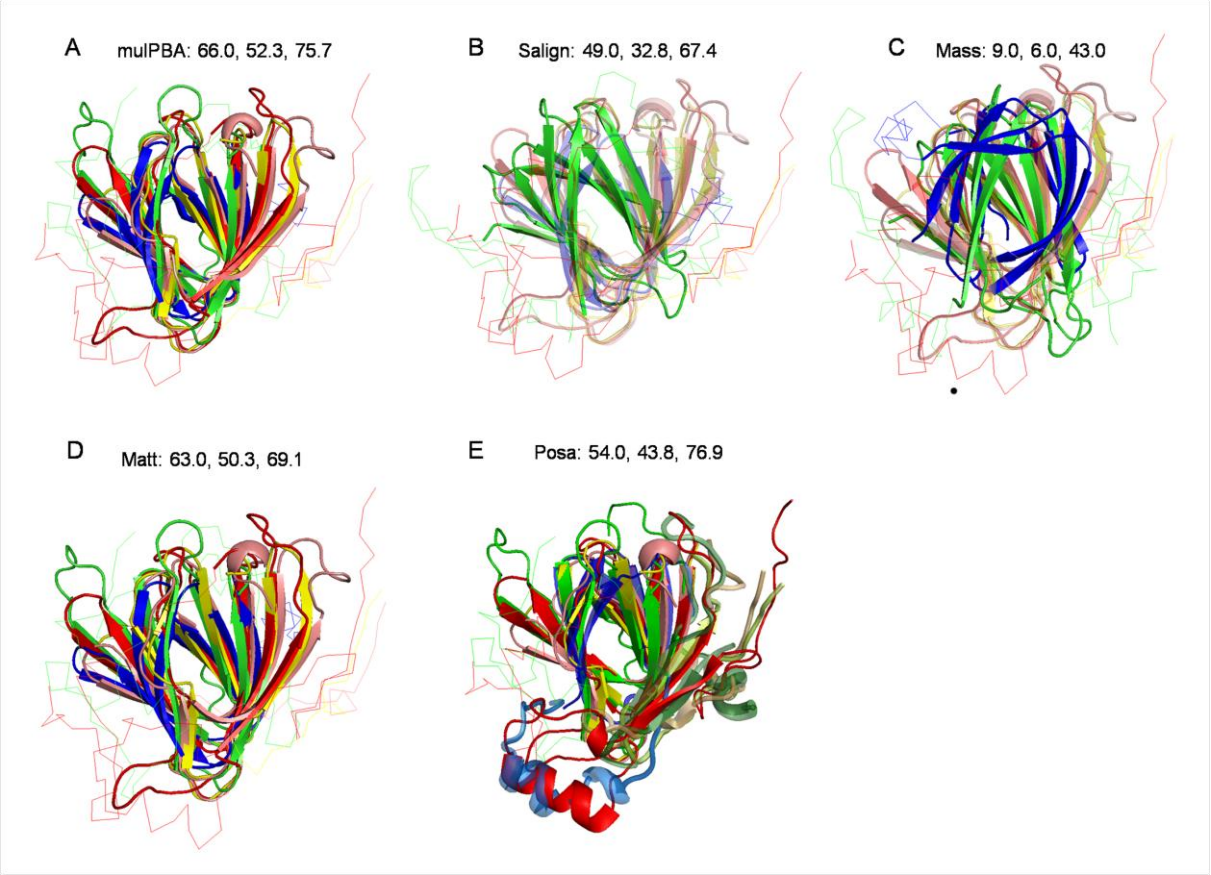

Figure 3. *Alignment of proteins with cupin fold.* Alignment of 5 structures with cupin fold (PDB ID+chain: 1DZRa (*green*), 1O5Ua (*blue*), 1QXRa (*red*), 1V70a (*pink*) and 1VJ2a (*orange*)) using mulPBA (A), SALIGN (B), MASS (C), MATT (D) and POSA (E). The quality scores in terms of N_{rms}, N_{dist} and $N_{3.5}$ are also given. In (B) and (C) involving mis-alignments, the structures fitted in the optimal orientation are presented as partially transparent. In the panel (E) corresponding to Posa alignment, the structural regions altered based on detected flexibilities are indicated with a thick backbone. Figures are rendered in PyMol [63]

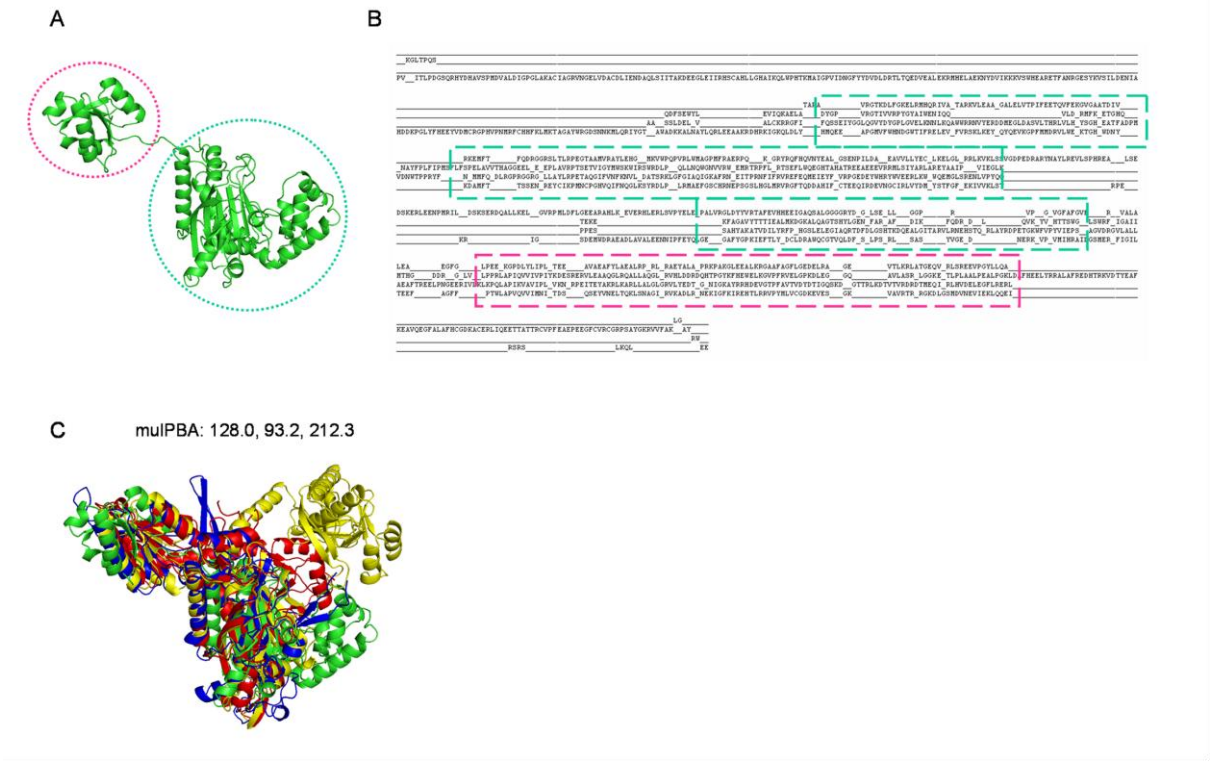

Figure 4. *Alignment of tRNA synthetases*. Four tRNA synthetase structures (PDB ID+chain: 1ADJa (*green*), 1ATIa (*blue*), 1HC7a (*red*) and 1QF6a (*yellow*)) are compared (A) One of the structures showing the two domains (highlighted in *pink* and *light blue*), joined by a flexible loop. (B) Alignment of PB sequences corresponding to the four structures. The regions of alignment corresponding to the two domains are highlighted. The 3D fit of the four tRNA synthetase structures using mulPBA (C). The quality scores in terms of N_{rms}, N_{dist} and $N_{3.5}$ are also given. Figures are rendered in PyMol [63]

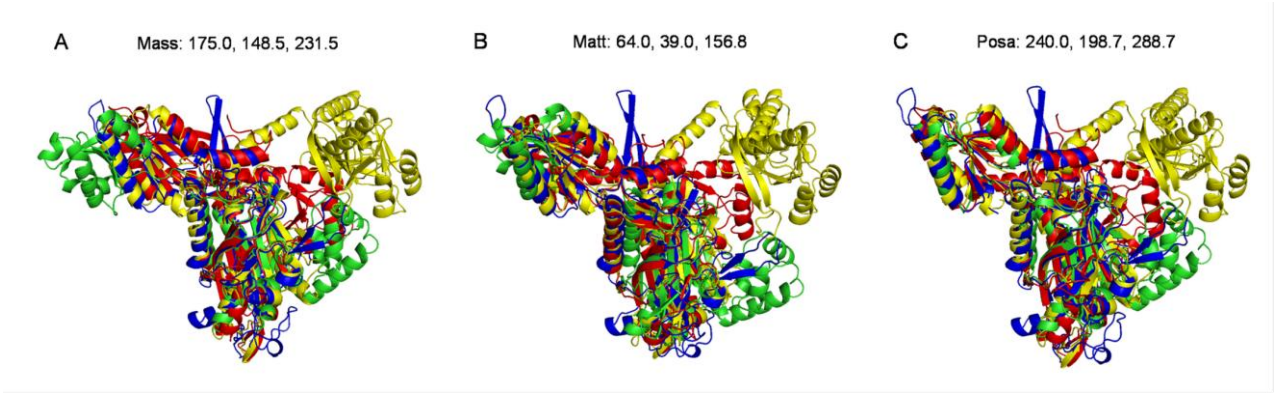

Figure 5. *Alignment of tRNA synthetases*. Four tRNA synthetase structures (PDB ID+chain: 1ADJa (*green*), 1ATIa (*blue*), 1HC7a (*red*) and 1QF6a (*yellow*)) are compared using MASS (A), MATT (B) and POSA (C). The quality scores in terms of N_{rms}, N_{dist} and $N_{3.5}$ are also given. Figures are rendered in PyMol [63].

Table 1. *Comparison of mulPBA with different MSTAs*. The protein families used for comparison are given, the average length is indicated within parentheses. The protein chains used for alignment are also listed using PDB ID followed by chain identifier. The alignment quality is indicated in terms of N_{rms}, N_{dist} and $N_{3.5}$, listed in order. The top score among the methods is highlighted in *red*. The second best scores are in *green*. The best scores among the first four methods (mulPBA, SALIGN, MAMMOTH and MASS) which are rigid-body MSTAs, are highlighted in bold. 'NA' indicates that the method fails in generating an alignment.