Supplementary Material:
GrAlign: Fast and Flexible Alignment of Protein 3D Structures Using Graphlet Degree Similarity

Noël Malod-Dognin and Nataša Pržulj

January 13, 2014

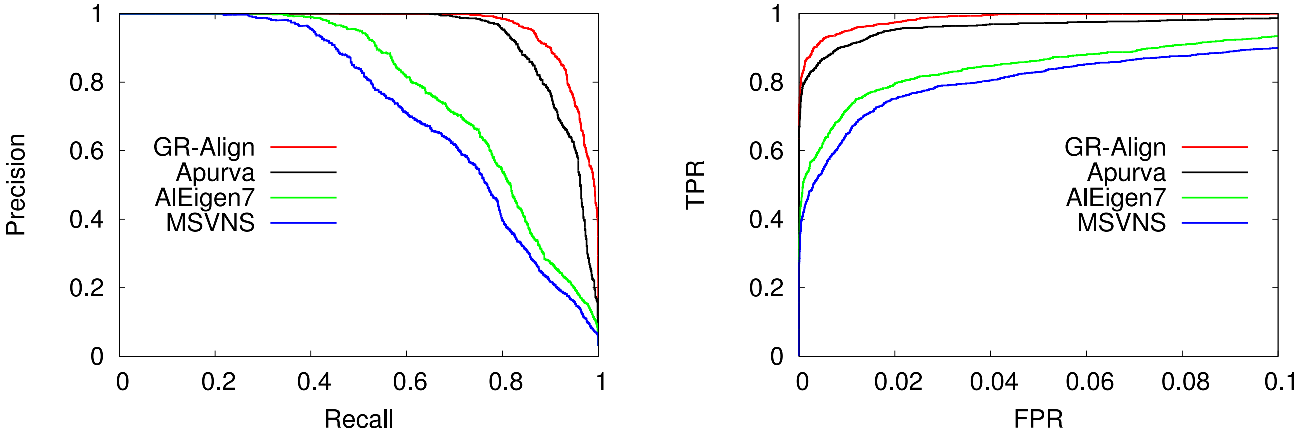
1 Additional method

For a given alignment $(a_1 \leftrightarrow b_1, a_2 \leftrightarrow b_2, \ldots, a_n \leftrightarrow b_n)$, the *distance difference matrix* is a symmetric square $n \times n$ matrix with entries $|d_{a_i a_j} - d_{b_i b_j}|$ at position (i, j), where $d_{a_i a_j}$ (resp. $d_{b_i b_j}$) is the Euclidean distance between the α-carbons of residues a_i and a_j (resp. b_i and b_j). Distance difference matrices are used in [1] for identifying flexible regions and hinges. In a distance difference matrix, one is interested in the blocks of low distance differences appearing along the diagonal, which correspond to common rigid substructures.

2 Additional results

2.1 Comparing CMO methods

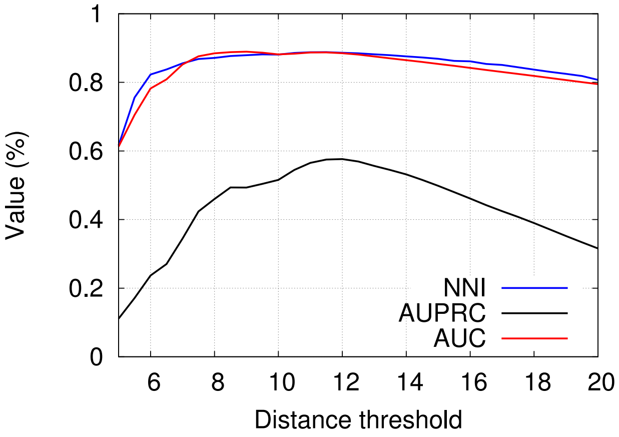
In the main document, we compared the classification performance of GR-Align, A_purva, AlEigen7 and MSVNS on the Proteus_300. Supplementary Figure 1 presents the Precision-Recall curves and the ROC curves that are obtained when the edge-correctness of the alignments are compared with the SCOP classification at Family level.



Supplementary Figure 1. Classification performance comparison on the Proteus_300 dataset, when using the SCOP classification at family level as reference. Left: For each distance threshold ϵ, the Precision-Recall curves plot the Precision versus the Recall that are obtained by each method. Right: The corresponding ROC curves plot the True Positive Rate (TPR) versus the False Positive Rate (FPR).

2.2 Effect of the distance threshold

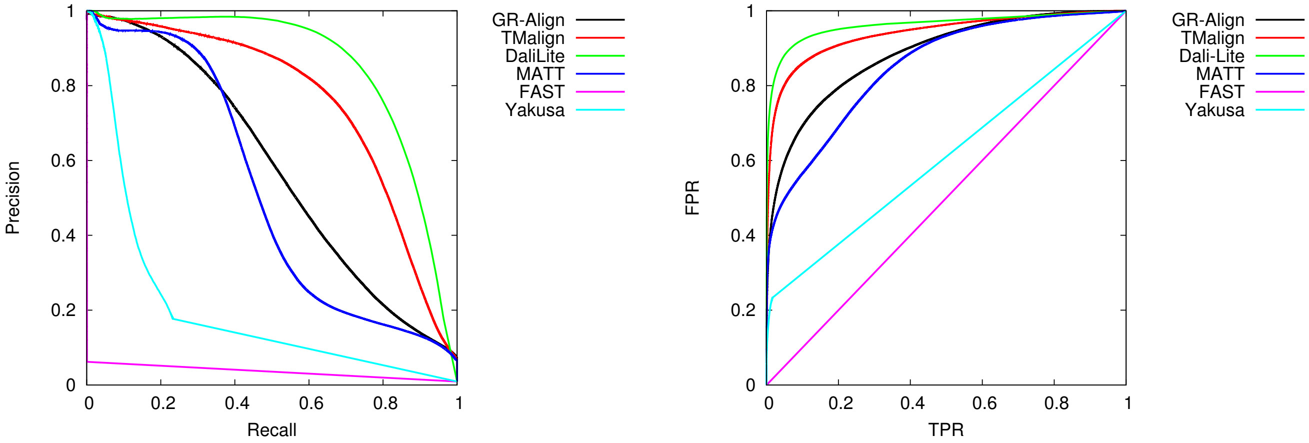
In the main document, we measured on the Gold-standard benchmark dataset the classification performance of GR-Align when the distance threshold ϵ for defining the contact edges in the contact maps is varied from 5Å to 20Å, in increments of 0.5Å. Supplementary Figure 2 presents the corresponding nearest neighbour identification rates (NNI), area under the precision-recall curves (AUPRC), and area under the ROC curves (AUC).



Supplementary Figure 2. Effect of the distance threshold on GR-Align's classification performance. On the Gold-standard benchmark dataset, the lines present the nearest neighbour identification rates (NNI), the area under the precision-recall curves (AUPRC), and the area under the ROC curves (AUC) that are obtained by GR-Align when the distance threshold for generating contact maps varies from 5Å to 20Å, in increments of 0.5Å.

2.3 Large scale comparison

In the main document, we compared the classification performances of GR-Align, DaliLite, TM-Align, MATT, Yakusa, and FAST on the Gold-standard benchmark. Supplementary Figure 3 presents side by side the corresponding precision-recall curves (already presented in the main document) and the ROC curves.



Supplementary Figure 3. Classification performance comparison on the Gold-standard benchmark, when using the CATH classification at Topology level as reference. Left: For each distance threshold ϵ, Precision-Recall curves plot the Precision versus the Recall that are obtained by each method. Right: The corresponding ROC curves plot the True Positive Rate (TPR) versus the False Positive Rate (FPR).

2.4 Aligning flexible proteins

In the main document, we used GR-Align, DaliLite, MATT and TM-Align to retrieve from Astral-40 database the protein domains that are most similar to a human calmodulin (SCOP id d1clla_). Supplementary Tables 1 and 2 present the 10 domains that are the most similar to d1clla_ according to GR-Align (Supplementary Table 1:Left), DaliLite (Supplementary Table 1:Right), MATT (Supplementary Table 2:Left), and TM-Align (Supplementary Table 2:Right). As already mentioned in the main document, GR-Align's top scoring alignments are in better agreement with the SCOP classification, and map more residues than the ones produced by the other methods.

GR-Align				DaliLite			
SCOP id	EC (%)	Class	Cov. (%)	SCOP id	Z-score	Class	Cov. (%)
d1exra_	97.3	a.39.1.5	100	d1exra_	19.1	a.39.1.5	100
d3fwba_	78.6	a.39.1.5	98.6	d3fwba_	13.4	a.39.1.5	88.2
d1wdcb_	75.9	a.39.1.5	95.1	d1oqpa_	12.4	a.39.1.5	50.7
d2mysb_	74.1	a.39.1.5	93.8	d1wdcb_	10.7	a.39.1.5	65.6
d1m45a_	66.5	a.39.1.5	93.1	d1s6ca_	10.6	a.39.1.5	63.2
d1auib_	64.2	a.39.1.5	93.8	d1xo5a_	9.8	a.39.1.5	65.6
d3jtdc_	63.1	a.39.1.5	95.8	d2zfda1	9.8	a.39.1.5	50.0
d1s6ia_	62.6	a.39.1.5	95.8	d3d10a_	9.5	a.39.1.2	56.2
d1hqva_	61.1	a.39.1.8	93.8	d2pvba_	9.2	a.39.1.4	47.2
d1s6ca_	59.8	a.39.1.5	95.8	d1auib_	8.9	a.39.1.5	91.0

Supplementary Table 1. The ten best-ranking protein domains found by GR-Align and DaliLite. Columns 1 to 4 show, for each of the top 10 ranking protein alignments returned by GrAlign, the SCOP id of the mapped proteins, the edge-correctness of the alignments, the SCOP classification of the mapped protein (d1clla_ classification is a.39.1.5), and the percentage of d1clla_'s residues that are covered by the alignment, respectively. Columns 5 to 8 show the same for the top 10 ranking alignments returned by DaliLite, except that the similarity is expressed in terms of Z-Score.

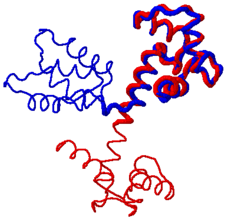
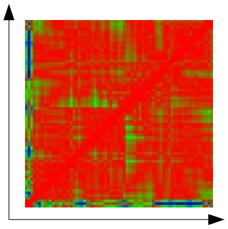
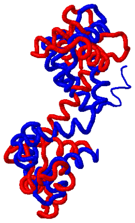
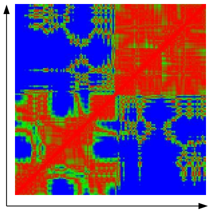
MATT				TM-Align			
SCOP id	Raw score	Class	Cov. (%)	SCOP id	TM-score	Class	Cov. (%)
d1exra_	424.7	a.39.1.5	99.3	d1exra_	0.974	a.39.1.5	100
d1y1xa_	181.1	a.39.1.8	61.8	d1oqpa_	0.613	a.39.1.5	51.4
d1alva_	177.0	a.39.1.8	61.1	d3d10a_	0.542	a.39.1.2	56.9
d1hqva_	176.2	a.39.1.8	58.3	d1qlsa_	0.527	a.39.1.2	59.0
d1k94a_	172.5	a.39.1.8	61.8	d1wdcb_	0.507	a.39.1.5	72.2
d1ij5a_	155.5	a.39.1.9	63.2	d1a4pa_	0.504	a.39.1.2	56.9
d1wdcb_	155.0	a.39.1.5	79.2	d2nxqa_	0.497	a.39.1.5	45.1
d1oqpa_	152.8	a.39.1.5	50.7	d1xk4a1	0.493	a.39.1.2	51.4
d3fwba_	149.6	a.39.1.2	50.7	d1xk4c1	0.492	a.39.1.2	50.7
d3d10a_	149.2	a.25.1.5	54.9	d2egdb_	0.486	a.39.1.2	53.5

Supplementary Table 2. The ten best-ranking protein domains found by MATT and TM-Align. Columns 1 to 4 show, for each of the top 10 ranking protein alignments returned by MATT, the SCOP id of the mapped proteins, the raw score of the alignments, the SCOP classification of the mapped protein (d1clla_ classification is a.39.1.5), and the percentage of d1clla_'s residues that are covered by the alignment, respectively. Columns 5 to 8 show the same for the top 10 ranking alignments returned by TM-Align, except that the similarity is expressed in terms of TM-Score.

Supplementary Table 3 details the alignments of the Human calmodulin (SCOP id: d1clla_) and of the Backer's Yeast calmodulin (SCOP id: d3fwba_) that are obtained using the rigid-body superimposition based method TM-Align, and the flexible one of GR-Align. TM-Align maps together the second EF-hand unit of each protein, for a total of 71 mapped residues. The mapped structure can be well superimposed with a root mean square deviation of superimposed coordinates (RMSD) of 1.86Å. This is also highlighted by the 71x71 distance difference matrix of the alignment, where most distance differences are smaller than 2.5Å. GrAlign maps the whole structure together, for a total of 143 mapped residues. The mapped structure cannot be well superimposed, as highlighted by a RMSD of 7.86Å. The 143x143 distance difference matrix of the alignment shows that the alignment consists of two rigid regions (each corresponding to one EF-hand unit) that can be well superimposed individually, as highlighted by the low-distance-difference blocks (in red, i.e., having distance differences smaller than 2.5Å) that have between each others large distance differences (the blue blocks, i.e., having distance differences larger than 5Å). The bordering of the two rigid regions indicates the location of the hinge (residue 73 of d1clla_).

References

[1] U. Emekli, D. Schneidman-Duhovny, H.J. Wolfson, R. Nussinov, and T. Haliloglu. Hingeprot: automated prediction of hinges in protein structures. *Proteins*, 70(4):1219–1227, 2008.

Method	# Residues	RMSD	Superimposition	dist.diff. matrices
TM-Align	71	1.86Å		 71 × 71
GR-Align	143	7.98Å		 143 × 143

Supplementary Table 3. Alignment between d1clla_ and d3fwba_. For each alignment, column 2 presents the number of mapped residues, column 3 presents the Root Mean Squared Deviation of superimposed coordinates, column 4 presents the corresponding superimposition (with mapped structures in bold), and column 5 presents the distance difference matrix of the mapping. The distances differences are colour coded: distances differences smaller than 2.5Å are in red, distance differences between 2.5Å and 5Å are in green, and distance difference larger than 5Å are in blue.