Supporting information for protein structure alignment beyond spatial proximity

Sheng Wang, Jianzhu Ma, Jian Peng and Jinbo Xu

Supplemental Data

In addition to the metrics described in the main text, we also evaluate structure alignments using several more metrics. In particular, for a given alignment, we count the number of aligned positions with distance deviation less than a given cutoff distance (e.g., 0.5, 1, 2, 4, and 8Å). We also calculate the uGDT of an alignment, which is defined as uGDT=$(n_1+n_2+n_4+n_8)/4$ where n_i is the number of aligned positions with distance deviation less than i Å. uGDT can be interpreted as a variant of alignment length weighted by alignment quality at each aligned position. GDT (Global Distance Test), which is uGDT normalized by protein length, is the official metric used by CASP (Critical Assessment of Structure Prediction) to evaluate the quality of a protein 3D model. Similar to TMscore, uGDT is a measure for geometric similarity, but not for evolutionary relationship. The higher the uGDT of an alignment is, the better. Finally, we also calculate the number of gap openings in an alignment. A desirable alignment of two evolutionarily-related proteins shall not contain too many gap opens since the probability of opening a gap during evolution is very small.

As shown in Table S1, in terms of uGDT, DeepAlign is slightly worse than TMalign on CDD and MALIDUP, but better than DALI, MATT and Formatt. On the more challenging MALISAM, DeepAlign obtains similar uGDT as DALI and TMalign, outperforming MATT and Formatt. DeepAlign has the smallest number of gap openings on all the three benchmarks and TMalign has much larger number of gap openings. Overall, DeepAlign compares favorably to DALI, MATT and TMalign even evaluated by a geometric similarity metric not used in the DeepAlign scoring function. The DeepAlign alignments contain fewer number of gap openings due to evolutionary information used by DeepAlign to generate alignments.

Table S1. Performance of five pairwise structure alignment tools and human experts on three benchmarks CDD, MALIDUP and MALISAM. The numbers shown in this table are the averaged numbers per alignment.

Method	$n_{0.5}$	n_1	n_2	n_4	n_8	uGDT	#gap opens
CDD (3591)							
DeepAlign	18.08	42.58	80.08	114.93	132.07	92.41	11.63
DALI	18.05	42.81	81.32	115.69	130.02	92.46	12.10
MATT	18.00	42.53	80.13	115.49	128.03	91.54	15.97
Formatt	17.67	41.61	75.68	102.09	111.43	82.71	14.39
TMalign	17.69	41.90	81.88	119.48	136.52	94.94	14.76
Manual	14.69	31.08	50.40	60.50	62.44	51.11	13.76
MALIDUP (241)							
DeepAlign	16.10	31.91	53.77	74.72	84.95	61.34	6.77
DALI	15.83	31.49	53.98	74.56	82.75	60.69	8.52
MATT	16.00	31.77	53.44	74.68	82.12	60.50	8.89
Formatt	16.04	31.13	50.34	65.21	70.32	54.25	8.07
TMalign	15.89	31.51	55.25	77.21	86.50	62.62	9.45
Manual	16.14	31.88	53.29	71.06	77.30	58.38	7.03
MALISAM (130)							
DeepAlign	11.08	18.54	33.84	52.41	60.57	41.34	6.73
DALI	10.88	18.46	33.75	52.92	60.46	41.40	9.04
MATT	10.68	18.14	31.86	50.28	56.27	39.14	8.85
Formatt	10.77	17.62	29.22	41.24	44.85	33.23	8.02
TMalign	10.22	17.06	32.94	52.26	60.88	40.78	10.22
Manual	10.85	18.15	31.91	50.16	56.38	39.15	7.27

Specific Examples

Case study 1: d1h99a1 and d1h99a2

[DeepAlign alignment]

>d1h99a1

RRIMJJKLCELC--FCAJHHJIHHHHHHHHHHIKAGCFCAJ--JJHIIHIIHJIJHHHHHHOGBF-EEPIMIHHHIKKPIHHHHHHHJHHHH
HHIKAGEECAJHJHHHHHHHHHHIMMLR
GAMEKFKTLLYD--IPIECMEVSEEIISYAKLQLGKKLNDS--IYVSLTDHINFAIQRNQKGLDI-KNALLWETKRLYKDEFAIGKEALVMV
KNKTGVSLPEDEAGFIALHIVNAELNEE

>d1h99a2

----------MPNIINITKVMEEILSIVKYHFKIEFNEESLHYYRFVTDLKFFAQRLFNGTHMEDDFLLDTVKEKYHRAYECTKKIQTYI
EREYEHKLTSDELLYLTIDIERVVK---
----------RRIHHHHHHHHHHHHHHHHHHIKAGDCAPJGAJHHHIIHHHHHHHHHJIJOGERRRPMHIKIIJIKKMHHHJJHHHHHHHH
HIKKAGEECAHHHHHHHHHHHHIIR---

[DALI alignment]

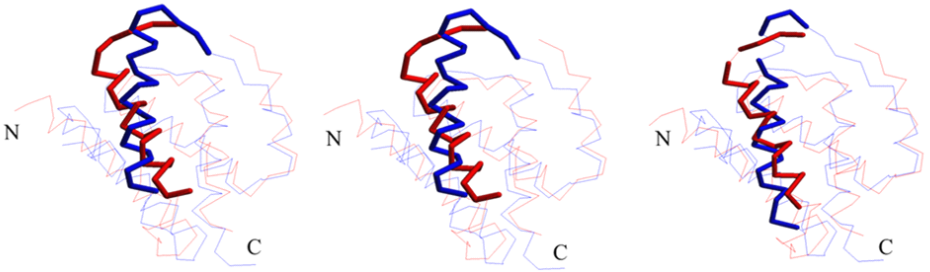
>d1h99a1

RRIMJJKLCELCFCAJHHJIHHHHHHHHHHIKAGCF------CAJJJHIIHIIHJIJHHHHHHOGBFEEPIMIHHHIKKPIHHHHHHHJHHH
HHHIKAGEECAJHJHHHHHHHHHHIMMLR
GAMEKFKTLLYDIPIECMEVSEEIISYAKLQLGKKL------NDSIYVSLTDHINFAIQRNQKGLDIKNALLWETKRLYKDEFAIGKEALVM
VKNKTGVSLPEDEAGFIALHIVNAELNEE

>d1h99a2

----------MPNIINITKVMEEILSIVKYHFKIEFNEESLHYYRFVTDLKFFAQRLFN--GTHM-EDDFLLDTVKEKYHRAYECTKKIQTY
IEREYEHKLTSDELLYLTIDIERVVK---
----------RRIHHHHHHHHHHHHHHHHHHIKAGDCAPJGAJHHHIIHHHHHHHHHJIJ--OGER-RRPMHIKIIJIKKMHHHJJHHHHHHH
HHIKKAGEECAHHHHHHHHHHHHIIR---

(A)



DeepAlign Manual DALI

(B)

Figure S1. The DeepAlign, Maual and DALI alignments for two closely-related domains d1h99a1 and d1h99a2. (A) The alignments in the FASTA format. The lines containing blue fonts are the CLE (conformational letter) strings and the lines containing red fonts are protein sequences. (B) 3D structure superimposition according to the alignments.

Table S2. Evaluation of the structure alignments of d1h99a1 and d1h99a2 generated by different tools. "Blosum1 (clesum1)" is the average mutation score per aligned position and "Bolsum (clesum)" is the mutation score of the whole alignment.

Method	LALI	RMSD	TMscore	RefAcc	Blosum1	Clesum1	Blosum	Clesum
DeepAlign	100	3.40	0.631	99.0	0.600	2.219	60	222
DALI	99	3.09	0.661	74.0	0.172	1.623	17	161
MATT	97	3.20	0.626	96.9	0.598	2.346	58	228
Formatt	89	3.17	0.579	90.6	0.629	2.371	56	211
TMalign	98	2.93	0.667	74.0	0.266	1.682	26	165

Case study 2: d1nekc_ and d1nekd_

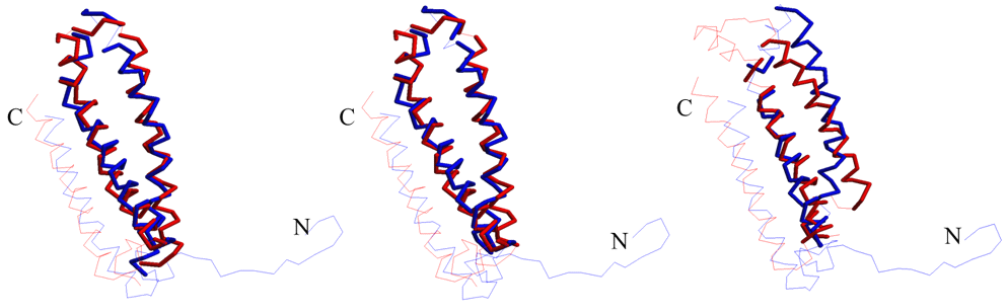
[DeepAlign alignment]

```
>d1nekc_
RRCNGFQECCFDGBPIJGEF-------CAJHHJIIHHHHHJIHHJIJJHHHHHHHJIJKMLAIJHHHHHIIKKMAHHHHHIKIHHHHHHHHH
IIHHIIIJHJKCMLCAGAHHJIH-------HHHIJIJHJHHHHHHHHHIIR--
MIRNVKKQRPVNLDLQTIRF-------PITAIASILHRVSGVITFVAVGILLWLLGTSLSSPEGFEQASAIMGSFFVKFIMWGILTALAYHV
VVGIRHMMMDFGYLEETFEAGKR-------SAKISFVITVVLSLLAGVLVW--
>d1nekd_
--------------------SNASALGRNGVHDFILVRATAIVLTLYIIYMVGFFATS-GELT-YEVWIGFFASAFTKVFTLLALFSILIHA
WIGMWQVLTDY---VK-------PLALRLMLQLVIVVALVVYVIYGFVVVWGV
--------------------RRPIGAGALCIHHHIIHHHHJHHHHHHHHHHIHHJIHIM-PKEE-AJHIIJHIKGAJIHHIHHHHHHHHHIII
JIIIJHHHJIK---MP-------MPIHIHHIIJHIIHHHHHJHHIJHHHIIJR
```

[MATT alignment]

```
>d1nekc_
RRCNGFQECCFDGBPIJGEFCAJ-HHJI-I-HHHHHJIHHJIJJHHHHHHHJIJKMLAIJHHHH----------H--I-IKKMAHHHHH
IKIHHHHHHHHHIIHHIIIJHJKCMLCAGAHHJIHHHHIJIJHJHHHHHHHHHIIR---
MIRNVKKQRPVNLDLQTIRFPIT-AIAS-I-LHRVSGVITFVAVGILLWLLGTSLSSPEGFEQA----------S--A-IMGSFFVKFI
MWGILTALAYHVVVGIRHMMMDFGYLEETFEAGKRSAKISFVITVVLSLLAGVLVW---
>d1nekd_
----------------------SNASALGRNGVHDFILVRATAIVLTLYIIYMVGFF------ATSGELTYEVWIGFFASAFTK-V-FTLL
ALFSILIHAWIGMWQVLTDYV------KPLALRLMLQLVIVVALVVYVIYGFV-VVWGV
----------------------RRPIGAGALCIHHHIIHHHHJHHHHHHHHHHIHHJI------HIMPKEEAJHIIJHIKGAJIH-H-IHHH
HHHHHHIIIJIIIJHHHJIKM------PMPIHIHHIIJHIIHHHHHJHHIJHH-HIIJR
```

(A)



(B)

Figure S2. The DeepAlign, DALI and MATT alignments for two closely-related domains d1nekc_ and d1nekd_. (A) The alignments in the FASTA format. The lines containing blue fonts are the CLE (conformational letter) strings and the lines containing red fonts are protein sequences. (B) 3D structure superimposition according to the alignments.

Table S3. Evaluation of the structure alignments of d1nekc_ and d1nekd_ generated by different tools. "Blosum1 (clesum1)" is the average mutation score per aligned position and "Bolsum (clesum)" is the mutation score of the whole alignment.

Method	LALI	RMSD	TMscore	RefAcc	Blosum1	Clesum1	Blosum2	Clesum2
DeepAlign	97	3.34	0.607	98.4	0.546	1.609	53	156
DALI	101	3.90	0.600	100.0	0.455	1.734	46	175
MATT	91	3.90	0.469	30.6	-0.594	1.207	-54	110
Formatt	81	3.99	0.415	30.6	-0.543	1.309	-44	106
TMalign	98	3.23	0.622	85.5	0.490	1.537	48	151

Case study 3: d1ef5a_ and d1ndda_

[DeepAlign alignment]

>d1ef5a_
RRGBCEEEEPGCBAKMFELBBPGACLQQLNJKJIJIKKKKCLQFQAGNKLDFDEEBQNLL-EEEQCNGPOOMNMN-LDDCQCEEEEECER-
BDTCIIRISVEDNNGNMYKSIMLTSQDKTPAVIQRAMSKHNLESDPAEEYELVQVISEDK-ELVIPDSANVFYAM-NSQVNFDFILRKKN-
>d1ndda_
---MLIKVK--TLTGKEIEID-IEPTDKVERIKERVEEKEGIP---PQQQRLIYS-----GKQMNDEKTAAD---YKILGGSVLHLVLALR
---RREDEF--EAJOLEECEE-EDAHGCAJHHHHHHHJKKAGC---AJHGEEEEN-----OGCCCAHGDAJJ---HOGECNGCECEECEER

[DALI alignment]

>d1ef5a_
RRGBCEEEEPGCBAKMFELBBPGACLQQLNJKJIJIKKKKCLQFQAGNKLDFDEEBQNLLEEEQ-CNGP-OOMNMNLDDC---QCEEEEECER-
BDTCIIRISVEDNNGNMYKSIMLTSQDKTPAVIQRAMSKHNLESDPAEEYELVQVISEDKELVI-PDSA-NVFYAMNSQV---NFDFILRKKN-
>d1ndda_
---MLIKVK--TLTGKEIEID-IEPTDKVERIKER-VEEKEGIPP--QQQRLIYS-----GKQMNDEKTAADYK------ILGGSVLHLVLALR
---RREDEF--EAJOLEECEE-EDAHGCAJHHHHH-HHJKKAGCA--JHGEEEEN-----OGCCCAHGDAJJHO------GECNGCECEECEER

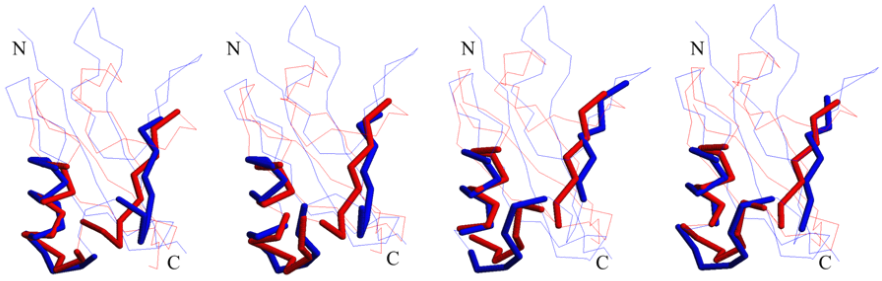
[MATT alignment]

>d1ef5a_
RRGE--CEEEEPGCB---AKMFELBBP-GACLQQLNJKJIJIKKKKCLQFQAGNKLDFDEEBQNLLEEEQCNGPO-OMNMNLDDC-----QCEEEEECER
EDTC--IIRISVEDN---NGNMYKSIM-LTSQDKTPAVIQRAMSKHNLESDPAEEYELVQVISEDKELVIPDSAN-VFYAMNSQV-----NFDFILRKKN
>d1ndda_
----MLIKV------KTLTGKEI-EI-DIEPTDKVERIKERVEE-KEGIPFQ----QQRLIYS--GKQMNDEKTAADYK------ILGGSVLHLVL-ALR
----RREDE------FEAJOLEE-CE-EEDAHGCAJHHHHHHJ-KKAGCAJ----HGEEEEN--OGCCCAHGDAJJHO------GECNGCECEEC-EER

[TMalign alignment]

>d1ef5a_
RRGBCEEEEPGCBAKMFELBBPGACLQQLNJKJIJIKKKKCLQFQAGNKLDFDEEBQNLLEEEQCNGPOOMNMNLDDC--QCEEEEECER
EDTCIIRISVEDNNGNMYKSIMLTSQDKTPAVIQRAMSKHNLESDPAEEYELVQVISEDKELVIPDSANVFYAMNSQV--NFDFILRKKN
>d1ndda_
---MLIKVK--TLTGKEIEID-IEPTDKVERIKERVE-EKEGIPPQ---QQRLIYS---GKQM-NDEKT-AAD-YKILGGSVLHLVLALR
---RREDEF--EAJOLEECEE-EDAHGCAJHHHHHHH-JKKAGCAJ---HGEEEEN---OGCC-CAHGD-AJJ-HOGECNGCECEECEER

(A)



DeepAlign DALI MATT TMalign

(B)

Figure S3. The DeepAlign, DALI, MATT and TMalign alignments for two domains d1ef5a_ and d1ndda_. (A) The alignments in the FASTA format. The lines containing blue fonts are the CLE (conformational letter) strings and the lines containing red fonts are protein sequences. (B) 3D structure superimposition according to the alignments.

Table S4. Evaluation of the structure alignments of d1ef5a_ and d1ndda_ generated by different tools. "Blosum1 (clesum1)" is the average mutation score per aligned position and "Bolsum (clesum)" is the mutation score of the whole alignment.

Method	LALI	RMSD	TMscore	RefAcc	Blosum1	Clesum1	Blosum	Clesum
DeepAlign	71	3.16	0.563	87.5	0.043	1.202	3	85
DALI	68	2.74	0.585	78.1	-0.029	0.854	-2	58
MATT	62	3.02	0.475	34.4	0.000	0.665	0	41
Formatt	53	2.96	0.410	34.4	0.151	0.698	8	37
TMalign	72	3.15	0.575	37.5	-0.194	0.482	-14	35

Table S5. Profile score at each position for the alignments between d1ef5a_ and d1ndda_ generated by DeepAlign, DALI, MATT, Formatt and TMalign. Only the sub-alignment corresponding to the 38-th and the 53rd residues of d1ef5a_ are shown. The last row contains the sum of the profile scores for the sub-alignments.

DeepAlign		DALI		MATT		Formatt		TMalign	
SE	0.12	SE	0.09	SE	0.12	SE	0.12	S-	n/a
KK	0.29	KE	0.17	K-	n/a	K-	n/a	KE	0.17
HE	-0.03	HK	-0.02	HK	-0.02	H-	n/a	HK	-0.02
NG	0.34	NE	-0.07	NE	-0.07	N-	n/a	NE	-0.07
LI	0.33	LG	-0.44	LG	-0.44	LG	-0.44	LG	-0.44
EP	0.16	EI	-0.30	EI	-0.30	EI	-0.30	EI	-0.30
S-	n/a	SP	0.16	SP	0.16	SP	0.16	SP	0.16
D-	n/a	DP	0.03	DP	0.03	DP	0.03	DP	0.03
P-	n/a	P-	n/a	PQ	0.09	PQ	0.09	PQ	0.09
AP	0.21	A-	n/a	A-	n/a	A-	n/a	A-	n/a
EQ	0.26	EQ	0.26	E-	n/a	E-	n/a	E-	n/a
EQ	0.15	EQ	0.15	E-	n/a	E-	n/a	E-	n/a
YQ	0.06	YQ	0.06	Y-	n/a	Y-	n/a	YQ	-0.08
ER	0.03	ER	0.03	EQ	0.07	EQ	0.07	EQ	0.07
LL	0.81	LL	0.81	LQ	0.12	LQ	0.12	LR	-0.14
VI	0.22	VI	0.22	VR	-0.05	VR	-0.05	VL	0.23
	2.93		1.14		-0.29		-0.20		-0.29