Supplemental Information

Recognizing Protein Substructure Similarity Using Segmental Threading

Sitao Wu and Yang Zhang

Table S1. Average results of SEGMER threading in different categories of segments

	Segment type	# of RSSEs	First			Best in Top 5		
			TM-score	RMSD(Å)	Cov	TM-score	RMSD(Å)	Cov
144 hard targets	Cont*	2	0.414	3.62	0.99	0.462	3.38	0.99
		3	0.420	5.76	0.98	0.462	5.33	0.97
		4	0.482	6.21	0.97	0.520	5.66	0.97
	Disco†	2	0.350	8.02	0.99	0.386	8.02	0.99
		3	0.339	10.86	0.98	0.372	10.50	0.98
		4	0.444	9.62	0.97	0.469	9.18	0.97
150 easy targets	Cont*	2	0.526	1.88	0.99	0.586	1.76	0.99
		3	0.602	2.67	0.98	0.647	2.44	0.97
		4	0.663	3.10	0.97	0.700	2.77	0.97
	Disco†	2	0.428	4.03	0.99	0.474	4.30	0.99
		3	0.438	6.02	0.98	0.481	5.85	0.97
		4	0.498	6.42	0.98	0.539	6.00	0.97
12 CASP8 FM targets	Cont*	2	0.376	4.44	0.96	0.427	3.90	0.96
		3	0.336	7.96	0.95	0.374	7.59	0.95
		4	0.307	9.29	0.93	0.339	8.66	0.92
	Disco†	2	0.380	10.16	0.90	0.419	7.91	0.90
		3	0.327	13.94	0.84	0.367	13.02	0.84
		4	0.308	15.22	0.81	0.348	12.45	0.82

*Cont: continuous segments

†Disco: discontinuous segments

Table S2. Average TM-score of the substructures for continuous 2-RSSE segments predicted by SEGMER, MUSTER and HHpred (bold numbers show the best result in each category)

	Segments*	Threading program	First	Best in Top 5
144 hard targets	Common	SEGMER	**0.424**	**0.472**
		MUSTER	0.371	0.435
		HHpred	0.356	0.407
	Unaligned	SEGMER	0.433	0.482
150 easy targets	Common	SEGMER	**0.546**	**0.606**
		MUSTER	0.507	0.575
		HHpred	0.495	0.552
	Unaligned	SEGMER	0.520	0.589
12 CASP8 FM targets	Common	SEGMER	**0.428**	**0.485**
		MUSTER	0.314	0.418
		HHpred	0.296	0.330
	Unaligned	SEGMER	0.323	0.373

*'Common': segments that have alignments by all three algorithms; 'Unaligned': segments that have no alignments by the whole-chain threading algorithms (MUSTER)

Table S3. Average TM-score of the substructures for continuous 3-RSSE segments predicted by SEGMER, MUSTER and HHpred (bold numbers show the best result in each category)

	Segments*	Threading program	First	Best in Top 5
144 hard targets	Common	SEGMER	**0.426**	**0.469**
		MUSTER	0.388	0.449
		HHpred	0.373	0.421
	Unaligned	SEGMER	0.452	0.489
150 easy targets	Common	SEGMER	**0.614**	**0.661**
		MUSTER	0.572	0.631
		HHpred	0.557	0.607
	Unaligned	SEGMER	0.548	0.584
12 CASP8 FM targets	Common	SEGMER	**0.379**	**0.402**
		MUSTER	0.295	0.349
		HHpred	0.264	0.292
	Unaligned	SEGMER	0.284	0.354

*'Common': segments that have alignments by all three algorithms; 'Unaligned': segments that have no alignments by the whole-chain threading algorithms (MUSTER)

Table S4. Average TM-score of the substructures for continuous 4-RSSE segments predicted by SEGMER, MUSTER and HHpred (bold numbers show the best result in each category)

	Segments*	Threading program	First	Best in Top 5
144 hard targets	Common	SEGMER	**0.549**	**0.584**
		MUSTER	0.492	0.555
		HHpred	0.461	0.505
	Unaligned	SEGMER	0.618	0.656
150 easy targets	Common	SEGMER	**0.682**	**0.720**
		MUSTER	0.640	0.693
		HHpred	0.622	0.668
	Unaligned	SEGMER	0.706	0.730
12 CASP8 FM targets	Common	SEGMER	**0.354**	**0.372**
		MUSTER	0.273	0.303
		HHpred	0.220	0.251
	Unaligned	SEGMER	0.244	0.294

*'Common': segments that have alignments by all three algorithms; 'Unaligned': segments that have no alignments by the whole-chain threading algorithms (MUSTER)

Table S5. Average TM-score of the substructures for discontinuous 2-RSSE segments predicted by SEGMER, MUSTER and HHpred (bold numbers show the best result in each category)

	Segments*	Threading program	First	Best in Top 5
144 hard targets	Common	SEGMER	**0.351**	**0.388**
		MUSTER	0.301	0.361
		HHpred	0.302	0.348
	Unaligned	SEGMER	0.359	0.396
150 easy targets	Common	SEGMER	**0.459**	**0.508**
		MUSTER	0.433	0.492
		HHpred	0.428	0.478
	Unaligned	SEGMER	0.366	0.410
12 CASP8 FM targets	Common	SEGMER	**0.438**	**0.486**
		MUSTER	0.337	0.393
		HHpred	0.219	0.285
	Unaligned	SEGMER	0.378	0.415

*‘Common’: segments that have alignments by all three algorithms; ‘Unaligned’: segments that have no alignments by the whole-chain threading algorithms(MUSTER)

Table S6. Average TM-score of the substructures for discontinuous 3-RSSE segments predicted by SEGMER, MUSTER and HHpred (bold numbers show the best result in each category)

	Segments*	Threading program	First	Best in Top 5
144 hard targets	Common	SEGMER	**0.305**	**0.341**
		MUSTER	0.257	0.315
		HHpred	0.271	0.313
	Unaligned	SEGMER	0.422	0.450
150 easy targets	Common	SEGMER	**0.466**	**0.511**
		MUSTER	0.435	0.497
		HHpred	0.443	0.486
	Unaligned	SEGMER	0.475	0.514
12 CASP8 FM targets	Common	SEGMER	**0.381**	**0.419**
		MUSTER	0.335	0.394
		HHpred	0.267	0.323
	Unaligned	SEGMER	0.335	0.386

*'Common': segments that have alignments by all three algorithms; 'Unaligned': segments that have no alignments by the whole-chain threading algorithms (MUSTER)

Table S7. Average TM-score of the substructures for discontinuous 4-RSSE segments predicted by SEGMER, MUSTER and HHpred (bold numbers show the best result in each category)

	Segments*	Threading program	First	Best in Top 5
144 hard targets	Common	SEGMER	**0.405**	**0.433**
		MUSTER	0.293	0.358
		HHpred	0.305	0.352
	Unaligned	SEGMER	0.689	0.706
150 easy targets	Common	SEGMER	**0.548**	**0.592**
		MUSTER	0.519	0.582
		HHpred	0.524	0.573
	Unaligned	SEGMER	0.631	0.667
12 CASP8 FM targets	Common	SEGMER	**0.356**	**0.394**
		MUSTER	0.321	0.377
		HHpred	0.248	0.308
	Unaligned	SEGMER	0.276	0.339

*'Common': segments that have alignments by all three algorithms; 'Unaligned': segments that have no alignments by the whole-chain threading algorithms (MUSTER)

Table S8. Comparison of the accuracies of short/medium/long-range contact predictions extracted from MUSTER, SEGMER, and the combination of SEGMER+MUSTER on testing proteins (bold numbers show the best result in each category)

		MUSTER	SEGMER	SEGMER+MUSTER
144 hard targets	$ACC_{C\alpha\_short}$*	0.247	**0.348**	0.334
	$ACC_{C\alpha\_medium}$*	0.224	**0.326**	0.310
	$ACC_{C\alpha\_long}$*	0.274	0.239	**0.310**
	ACC_{SG_short}†	0.358	0.421	**0.460**
	ACC_{SG_medium}†	0.307	0.387	**0.416**
	ACC_{SG_long}†	0.347	0.331	**0.419**
150 easy targets	$ACC_{C\alpha\_short}$*	0.404	**0.488**	0.472
	$ACC_{C\alpha\_medium}$*	0.475	**0.554**	0.516
	$ACC_{C\alpha\_long}$*	0.684	0.713	**0.756**
	ACC_{SG_short}†	0.584	0.647	**0.664**
	ACC_{SG_medium}†	0.637	0.690	**0.721**
	ACC_{SG_long}†	0.788	0.783	**0.841**
12 CASP8 FM targets	$ACC_{C\alpha\_short}$*	0.120	0.190	**0.211**
	$ACC_{C\alpha\_medium}$*	**0.170**	0.125	0.116
	$ACC_{C\alpha\_long}$*	0.025	0.019	**0.027**
	ACC_{SG_short}†	0.208	0.199	**0.269**
	ACC_{SG_medium}†	0.123	**0.208**	0.192
	ACC_{SG_long}†	0.096	0.048	**0.098**

*Average accuracy for C_α contact prediction for short-range ($6\leq|i-j|<12$), medium-range ($12\leq|i-j|<24$) and long-range ($|i-j|\geq24$) with top L/5 predictions (L is protein length)

†Average accuracy for side-chain center contact prediction for short-range ($6\leq|i-j|<12$), medium-range ($12\leq|i-j|<24$) and long-range ($|i-j|\geq24$) with top L/5 predictions (L is protein length)