Supporting Information

Accurate disulfide-bonding network predictions improve *ab initio* structure prediction of cysteine-rich proteins

Jing Yang, Bao-Ji He, Richard Jang, Yang Zhang, and Hong-Bin Shen

Application to 3D structure modeling of cysteine-rich proteins

Supplementary Figure S2 presents a scatter plot of the TM-score and RMSD of QUARK with Cyscon versus that without Cyscon predictions for the 158 proteins. We can see that most models are improved in terms of both TM-score and RMSD. For the criterion of TM-score (RMSD), 117 (126) out of 158 proteins have better results with the connectivity pattern restraints, which indicates that Cyscon predictions can guide the modeling of disulfide-rich proteins.

In supplementary Figure S3, we present the TM-score and RMSD improvements versus the difference in the number of the disulfide bonds satisfied by the QUARK and QUARK-Cyscon models, i.e. $(N_{mod}\text{-}N_{ori})/N_{pre}$, where N_{mod} and N_{ori} are the number of disulfide bonds satisfied in the QUARK-Cyscon and the original QUARK models, respectively; and N_{pre} is the predicted number of disulfide bonds in the sequence. It is clear that the model quality improves as the number of satisfied disulfide bonds increases. This demonstrates that the integration of the Cyscon bond prediction directly contributed to the 3D model improvement of the QUARK modeling.

Table S1. Benchmark datasets used to build and assess the prediction model

Dataset		B=2	B=3	B=4	B=5	B>5	B=2~5
SPX [a]	numP [b]	160	165	65	38	36	428
	numB [c]	320	495	260	190	250	1265
PDBCYS	numP [b]	100	85	41	37	51	263
	numB [c]	200	255	164	185	379	804

[a] Training dataset of Cyscon prediction model.

[b] Number of proteins.

[c] Number of disulfide bonds.

Table S2. Detailed results of structure modeling of 158 proteins by QUARK with or without disulfide bonds predicted by Cyscon as restraints

PDBID	nA [a]	TM-Q [b]	RMSD-Q [c]	TM-C [d]	RMSD-C [e]	nB [f]	nB-Q [g]	nB-C [h]
1ahoA	64	0.358	8.0	0.323	8.1	4	0	4
1ajjA	37	0.314	7.4	0.430	3.5	3	1	3
1b9wA	91	0.242	12.1	0.273	12.1	5	2	5
1bteA	92	0.337	7.5	0.273	10.1	5	4	5
1bx7A	51	0.270	10.9	0.321	10.3	5	2	5
1clvI	32	0.192	5.2	0.185	7.1	3	1	3
1d0dA	60	0.279	9.6	0.235	9.4	3	1	3
1d0dB	58	0.245	9.1	0.286	7.4	3	1	3
1dtdB	61	0.273	11.3	0.406	4.8	4	1	4
1eaiC	61	0.243	11.3	0.254	9.3	5	1	4
1edmB	39	0.494	3.5	0.509	3.7	3	2	3
1en2A	85	0.257	8.8	0.280	9.1	8	3	7
1ezgA	82	0.183	11.6	0.204	11.1	8	6	7
1fd3A	41	0.342	8.0	0.318	7.4	3	1	3
1fk5A	93	0.636	3.7	0.632	3.6	4	4	4
1fleI	47	0.171	9.4	0.174	10.5	4	2	4
1fltV	95	0.218	12.6	0.225	11.8	3	0	3
1g6xA	58	0.249	9.4	0.269	7.7	3	1	3
1h59A	54	0.401	5.6	0.416	5.7	3	2	3
1h9hI	30	0.309	7.8	0.297	7.3	3	2	3
1hiaI	48	0.251	10.5	0.296	8.9	5	2	4
1hypA	75	0.491	4.1	0.577	3.4	4	2	4
1i71A	83	0.221	13.5	0.248	11.1	3	0	3
1icfI	65	0.416	5.3	0.394	5.3	3	2	3
1jk4A	79	0.253	10.8	0.280	8.0	7	2	5
1kliL	61	0.204	11.2	0.296	10.8	3	1	3
1kp6A	79	0.324	9.0	0.260	10.9	4	2	3
1l0sA	87	0.205	11.1	0.252	10.2	4	2	4
1lpbA	85	0.171	11.8	0.219	11.9	5	0	5
1lr7A	73	0.273	12.3	0.282	11.2	5	2	5
1lu0A	29	0.157	7.9	0.267	7.1	3	1	3
1moxC	49	0.352	10.1	0.460	6.0	3	1	3
1mwpA	96	0.314	12.9	0.300	11.4	3	1	3
1n69A	78	0.364	11.3	0.564	3.4	3	1	3
1oc0B	37	0.213	8.0	0.227	6.6	4	2	4
1p9gA	40	0.206	6.7	0.508	2.9	5	2	4
1q9bA	43	0.433	3.6	0.393	4.5	4	3	4
1r0rI	51	0.261	8.2	0.291	8.2	3	0	3
1rfxA	89	0.225	17.0	0.261	17.8	5	1	5

1tejA	62	0.188	12.3	0.180	9.8	4	2	4
1tgrA	52	0.443	3.7	0.458	3.4	3	3	3
1tgsI	55	0.278	7.2	0.354	5.2	3	2	3
1tukA	67	0.323	10.7	0.422	4.0	4	1	4
1uoyA	64	0.312	6.7	0.366	6.1	4	1	4
1v6pA	62	0.239	10.8	0.281	10.3	4	3	4
1wqjB	80	0.208	11.6	0.228	11.2	6	0	6
1wqjI	62	0.365	7.4	0.439	7.4	3	3	3
1xu1R	38	0.206	7.9	0.189	8.5	3	1	3
1xu2R	36	0.193	7.7	0.182	8.0	3	1	3
1zlhB	74	0.202	12.5	0.209	9.8	6	1	5
1zmiA	28	0.301	6.4	0.378	5.7	3	2	3
1zmmA	31	0.359	7.4	0.380	5.2	3	1	3
1zr0B	59	0.254	10.8	0.314	7.3	3	0	3
1zt3A	80	0.389	6.7	0.406	5.5	3	2	3
2aibA	98	0.338	13.0	0.358	7.0	3	0	2
2b97A	70	0.265	9.8	0.357	9.2	4	1	4
2cg7A	90	0.202	15.6	0.247	15.2	4	2	4
2dspB	91	0.215	12.4	0.231	11.8	6	0	5
2dspI	57	0.284	11.3	0.425	6.0	3	1	3
2erlA	40	0.270	8.4	0.279	9.1	3	1	2
2erwA	53	0.272	9.3	0.273	7.7	3	1	3
2f3cI	46	0.297	8.9	0.308	7.8	3	1	3
2fcwB	78	0.254	9.5	0.303	10.0	6	0	5
2fklA	64	0.321	6.2	0.323	6.0	3	1	3
2fmaA	59	0.348	6.6	0.354	5.8	3	1	3
2g81I	56	0.217	11.0	0.272	6.0	7	4	6
2h5fA	75	0.285	11.6	0.253	10.5	5	1	5
2h7zA	75	0.247	12.4	0.260	13.3	5	0	4
2h7zB	77	0.263	12.5	0.294	11.9	5	1	4
2h9eC	52	0.235	10.3	0.263	9.5	3	0	3
2hlrA	67	0.193	10.2	0.211	11.1	5	0	4
2ijoI	56	0.243	9.1	0.264	7.1	3	0	3
2ilnI	53	0.199	12.5	0.208	7.6	6	4	5
2j8bA	78	0.386	6.0	0.387	5.5	5	3	5
2nlsA	36	0.197	7.6	0.201	7.8	3	1	3
2ottX	94	0.295	11.2	0.333	8.7	4	1	4
2p28A	96	0.201	16.0	0.280	12.1	3	1	3
2posA	94	0.362	11.3	0.518	4.9	3	0	3
2pw8I	61	0.240	15.9	0.264	14.9	3	1	3
2qskA	95	0.220	13.6	0.242	10.1	5	2	5
2rjiA	84	0.317	9.5	0.596	9.4	4	0	4
2rknA	77	0.290	11.1	0.453	4.6	4	3	4

2rkyA	90	0.236	12.0	0.273	13.0	4	2	4
2rkzA	89	0.275	15.8	0.258	12.7	4	1	3
2uuyB	52	0.231	9.6	0.297	7.8	4	2	4
2v33A	91	0.212	13.2	0.227	10.4	3	1	2
2vu8I	33	0.541	5.6	0.538	2.0	3	2	3
2w8xA	72	0.232	12.8	0.236	11.9	4	0	4
2xdgA	89	0.232	12.3	0.355	9.1	3	1	3
2xttA	35	0.579	6.5	0.551	5.3	3	2	3
2y5fL	54	0.276	9.5	0.267	8.3	3	1	3
2z7fI	50	0.196	11.0	0.198	10.6	4	3	4
3bg4D	46	0.264	10.8	0.277	8.8	5	0	5
3bpsE	41	0.317	7.8	0.244	5.6	3	1	3
3bqpA	80	0.613	3.3	0.622	3.2	3	3	3
3bt2B	40	0.228	7.9	0.233	7.9	4	2	4
3bt4A	85	0.162	11.6	0.198	10.2	6	2	5
3c05A	59	0.194	9.6	0.198	10.7	4	0	4
3c05B	59	0.236	10.7	0.165	10.7	4	2	4
3ca7A	50	0.263	8.9	0.289	8.9	3	1	3
3e4hA	29	0.291	6.0	0.226	5.8	3	3	3
3e7rL	40	0.312	6.8	0.361	4.8	3	1	3
3e8yX	30	0.162	10.9	0.242	9.8	3	0	3
3ejhA	93	0.337	12.8	0.337	14.2	4	3	3
3fjuB	65	0.209	10.3	0.311	8.1	5	0	5
3fprA	85	0.279	10.9	0.245	8.3	4	0	4
3h0tC	22	0.119	7.2	0.101	5.1	4	1	3
3i5wA	32	0.194	7.5	0.240	6.4	3	1	3
3iolA	100	0.304	13.7	0.360	11.0	3	0	3
3ix0A	87	0.239	12.4	0.256	11.4	5	2	4
3k9xA	86	0.224	12.7	0.363	13.9	6	4	5
3kl6B	50	0.229	8.0	0.269	7.6	3	1	3
3n44A	54	0.362	8.3	0.371	10.1	3	2	3
3n7sC	91	0.545	7.0	0.597	6.0	3	1	3
3nggA	46	0.206	8.9	0.230	6.6	4	2	4
3nirA	48	0.351	9.4	0.378	5.6	3	1	3
3odvA	38	0.324	6.9	0.314	8.8	3	0	3
3op8A	85	0.253	12.8	0.202	11.5	3	0	3
3psmA	47	0.220	9.4	0.245	8.7	4	2	4
3q8jA	36	0.166	7.6	0.183	7.4	3	0	3
3qteA	32	0.166	8.0	0.572	1.9	3	1	3
3ru4B	61	0.193	9.2	0.186	11.4	7	3	7
3s64A	81	0.444	8.7	0.668	2.9	3	0	3
3tvjI	35	0.485	3.0	0.521	2.6	3	3	3
3u1jE	56	0.241	8.9	0.267	7.1	3	0	3

3uciA	72	0.180	13.6	0.213	11.7	6	3	6
3zrzA	89	0.267	15.7	0.212	15.3	4	2	4
3zxcA	71	0.225	9.6	0.195	11.5	6	2	5
3zzoA	93	0.255	11.9	0.225	11.4	5	1	4
4a94C	51	0.299	8.4	0.405	4.3	3	1	3
4aorD	34	0.173	7.7	0.254	6.4	3	2	3
4ay9A	88	0.195	16.5	0.175	11.8	5	1	4
4bdxA	83	0.375	7.4	0.391	6.2	5	4	4
4bfhA	30	0.137	10.3	0.245	5.2	3	0	3
4bnrI	59	0.248	9.6	0.281	7.4	3	1	3
4bqdA	78	0.218	10.3	0.246	9.4	3	2	3
4bvwA	79	0.310	9.2	0.235	10.2	3	0	3
4cpaI	37	0.152	8.2	0.159	7.8	3	2	2
4gi3C	57	0.256	9.6	0.300	9.0	4	0	3
4guxD	34	0.269	7.4	0.268	6.9	3	2	3
4gv5A	42	0.264	8.8	0.244	8.3	3	1	3
4gvbA	77	0.342	8.3	0.257	9.9	4	1	3
4gvbB	74	0.273	10.3	0.262	10.3	3	2	3
4hcsA	67	0.441	9.2	0.437	8.5	3	2	3
4he7A	53	0.248	9.4	0.215	8.2	4	2	4
4i6oA	68	0.286	13.5	0.676	3.5	3	0	3
4kt1E	90	0.204	11.5	0.230	12.4	7	3	4
4lfsA	35	0.227	11.7	0.229	6.1	3	0	2
4lftA	64	0.187	11.7	0.274	7.2	5	0	4
4ndsA	94	0.293	15.3	0.304	10.3	3	0	3
4oijA	71	0.343	11.6	0.468	8.1	3	0	2
4ozkA	49	0.200	10.7	0.202	9.4	3	1	3
4p39A	69	0.531	10.8	0.597	10.8	3	3	3
4p3aA	70	0.573	3.2	0.676	2.5	3	2	3
4sgbI	51	0.277	8.8	0.243	7.3	4	1	4
4ttnA	29	0.291	5.2	0.233	6.8	3	3	3
4u5hA	85	0.320	9.6	0.487	5.2	5	2	5
4wp4A	43	0.427	3.5	0.386	4.4	4	4	4

[a] Number of amino acids.

[b] TM-score of QUARK prediction without Cyscon predictions.

[c] RMSD (Å) of QUARK prediction without Cyscon predictions.

[d] TM-score of QUARK prediction with Cyscon predictions.

[e] RMSD (Å) of QUARK prediction with Cyscon predictions.

[f] Number of disulfide bonds in protein sequence.

[g] Number of disulfide bonds in original QUARK model.

[h] Number of disulfide bonds in modified QUARK model.

Table S3. Performance evaluation on the SPX dataset with different sequence identity thresholds

Method	B=2		B=3		B=4		B=5		B=2~5	
	Q_P	Q_C	Q_P	Q_C	Q_P	Q_C	Q_P	Q_C	Q_P	Q_C
SVR+OR (SPX) 15 [a]	72.8	72.8	61.1	70.0	44.3	57.8	11.5	45.2	59.1	64.7
SVR+OR (SPX) 20 [a]	74.1	74.1	63.7	72.3	47.1	59.5	19.0	50.7	61.3	67.0
SVR+OR (SPX) 25 [a]	75.8	75.8	64.8	73.3	56.7	66.7	36.4	62.4	65.4	71.2
SVR+OR (bDD) 15 [b]	73.3	73.3	62.0	70.8	45.5	58.0	19.8	52.1	60.5	65.9
SVR+OR (bDD) 20 [b]	78.0	78.0	63.1	72.1	54.3	66.7	23.6	54.1	64.1	69.7
SVR+OR (bDD) 25 [b]	77.0	77.0	67.8	75.6	55.4	70.2	36.3	61.9	66.3	72.9
SVR+OR (bDD) 30 [b]	80.9	80.9	74.7	80.6	61.5	73.9	37.6	62.9	71.8	77.0

[a] Confident bond detections were derived from 10-fold cross-validation at the different identity levels (15%, 20% and 25%).

[b] Annotated bDD database was used as the searching pool and the searched sequence shares no more than the certain identities with the query sequence (15%, 20%, 25% and 30%).

Table S4. Performance evaluation on the PDBCYS dataset with different sequence identity thresholds

Method	B=2		B=3		B=4		B=5		B=2~5	
	Q_P	Q_C	Q_P	Q_C	Q_P	Q_C	Q_P	Q_C	Q_P	Q_C
SVR+ OR (PDBCYS) 15 [a]	80.4	80.4	55.1	64.2	54.6	67.7	31.4	50.7	60.7	66.0
SVR+ OR (PDBCYS) 20 [a]	82.9	82.9	58.1	66.9	54.2	67.4	40.2	59.4	63.7	69.2
SVR+ OR (PDBCYS) 25 [a]	83.4	83.4	61.6	69.3	55.6	68.4	39.2	59.0	65.3	70.4
SVR+ OR (bDD) 15 [b]	83.1	83.1	60.4	69.3	52.8	66.4	34.3	53.6	63.1	68.6
SVR+ OR (bDD) 20 [b]	83.5	83.5	68.5	75.8	54.2	67.0	43.1	62.3	67.2	72.7
SVR+ OR (bDD) 25 [b]	84.4	84.4	76.5	82.5	57.4	69.6	55.4	69.9	72.3	77.0
SVR+ OR (bDD) 30 [b]	87.9	87.9	82.8	86.2	59.3	70.5	55.9	71.9	76.1	79.8

[a] Confident bond detections were derived from 20-fold cross-validation at the different identity levels (15%, 20% and 25%).

[b] Annotated bDD database was used as the searching pool and the searched sequence shares no more than the certain identities with the query sequence (15%, 20%, 25% and 30%).

Table S5. Bond-based coverage of confident bond detector at the different sequence identity thresholds

Identity	SPX (%)	PDBCYS (%)
CV 15 [a]	5.9	10.6
CV 20 [a]	12.6	21.6
CV 25 [a]	25.5	26.7
bDD 15 [b]	8.4	11.7
bDD 20 [b]	23.6	25.9
bDD 25 [b]	36.0	35.4

[a] Confident bond detections were derived from cross-validation at the different identity levels (15%, 20% and 25%).

[b] Annotated bDD database was used as the searching pool and the searched sequence shares no more than the certain identities with the query sequence (15%, 20% and 25%).

Table S6. Performance evaluation after the sequences with known structures are removed from the bDD pool.

Dataset	B=2		B=3		B=4		B=5		B=2~5	
	Q_P	Q_C	Q_P	Q_C	Q_P	Q_C	Q_P	Q_C	Q_P	Q_C
SPX (bDD') [a]	74.0	74.0	62.8	71.2	53.3	66.8	27.6	58.2	62.9	69.5
SPX (bDD) [b]	77.0	77.0	67.8	75.6	55.4	70.2	36.3	61.9	66.3	72.9
PDBCYS (bDD') [a]	79.1	79.1	65.9	73.2	56.0	68.9	46.1	66.7	65.9	72.4
PDBCYS (bDD) [b]	84.4	84.4	76.5	82.5	57.4	69.6	55.4	69.9	72.3	77.0

[a] A small bDD' was used, which contains 1,355 sequences, where 2,121 protein sequences with known 3D structures are removed from the original bDD.

[b] A large bDD was used, which contains 3,476 sequences. Refer to the text for details.

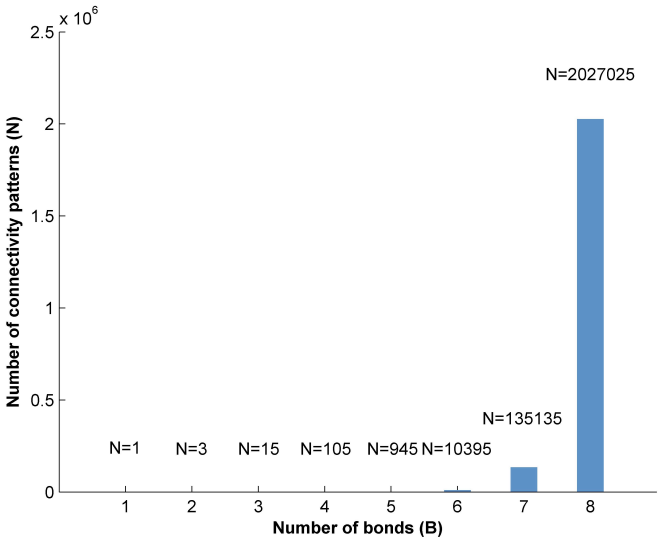

Figure S1. The number of the disulfide connectivity patterns (combinations of all possible bonds) is an exponential function of the number of disulfide bonds.

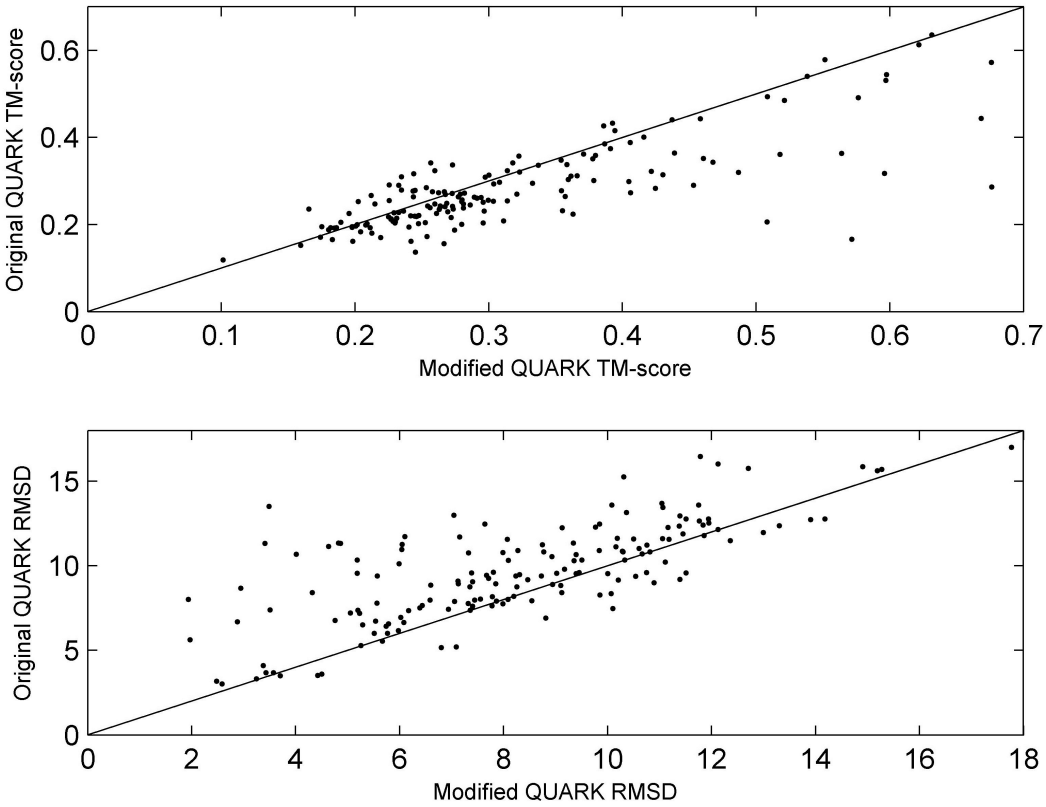

Figure S2. Comparison of TM-score and RMSD between original QUARK and modified QUARK simulations by the Cyscon prediction.

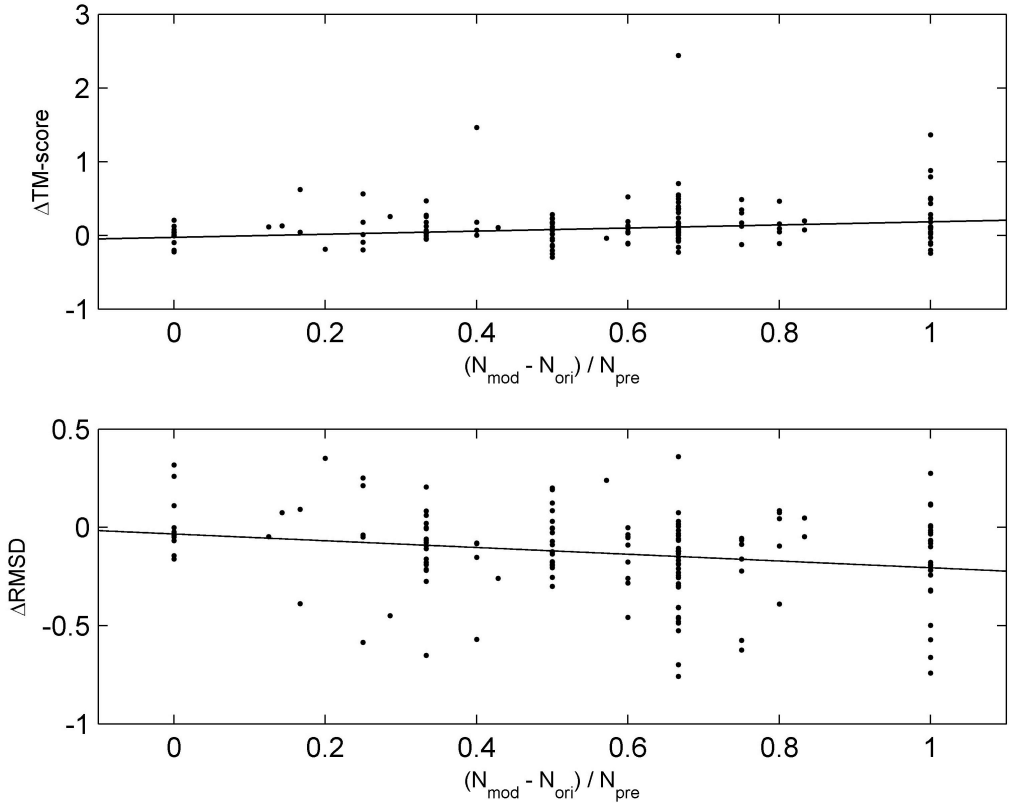

Figure. S3 The correlation of TM-score (RMSD) change (ΔTM-score = TM-score$_{mod}$-TM-score$_{orig}$) and the difference in satisfaction rate between original QUARK and modified QUARK by Cyscon prediction. N_{mod} and N_{ori} are the number of satisfied bonds in the modified and original model, respectively, and N_{pre} is the predicted number of bonds in the sequence.