STRUM: Structure-based stability change prediction on single-point mutation

Lijun Quan, Qiang Lv, Yang Zhang

Supplemental Information

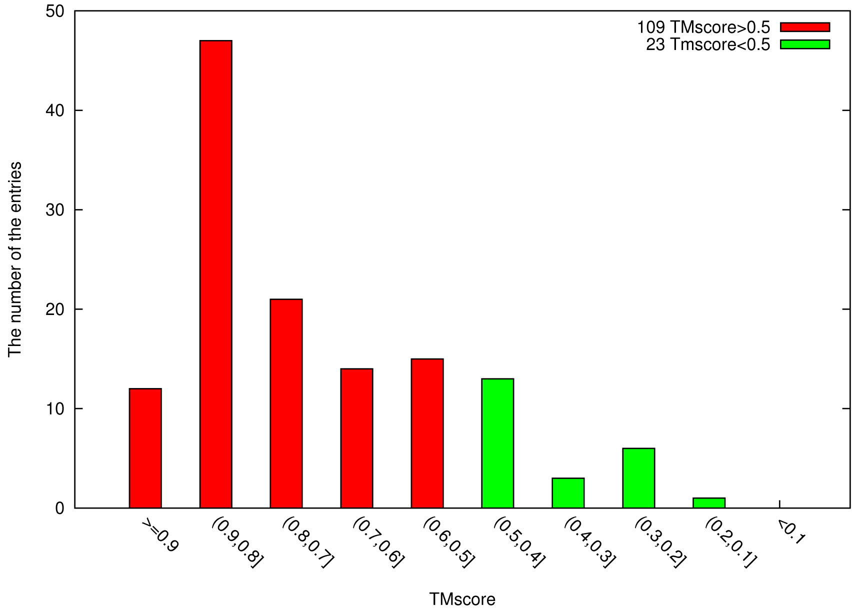

Figure S1. Histogram distribution of TM-score of the I-TASSER models on 131 proteins from the S2648 dataset.

Table S1. Summary of 120 features used in STRUM and their distributions in Q3421.

Feature class	No.	Feature	Median		Mean		p-value	Description
			DM[a]	SM[b]	DM[a]	SM[b]	(M-W Test)[c]	
Sequence-based	*Physicochemical properties*							
	1	A_w	9.00	12.00	9.92	11.97	3.20E-22	Wild-type amino acid
	2	A_m	9.00	9.00	9.39	9.36	0.051	Mutant-type amino acid
	3	V_d	-0.30	-0.05	-0.32	-0.13	4.40E-21	Volume difference
	4	V_w	3.05	2.91	2.94	2.88	4.60E-08	Wild volume
	5	MW_d	-18.05	-14.02	-20.26	-10.48	2.40E-11	molecular weight difference
	6	MW_w	131.18	133.10	133.50	132.62	0.11	Mutant molecular weight
	7	H_d	-0.65	-2.30	-1.48	-2.44	1.20E-10	Hydrophobicity scale difference
	8	H_w	0.80	2.80	2.76	4.43	6.80E-24	Wild hydrophobicity scale
	9	IP_d	0.00	0.04	-0.12	0.10	0.00097	Isoelectric point difference
	10	IP_w	5.97	5.87	6.06	5.85	4.40E-10	Wild isoelectric point
	Conservation score from multiple sequence alignments							
	11	$PSSM_C$	0.12	0.11	0.40	0.38	0.46	Column C in position specific scoring matrix
	12	$PSSM_M$	0.22	0.18	0.50	0.49	0.028	Column M in position specific scoring matrix
	13	$PSSM_F$	0.28	0.22	1.30	0.88	1.60E-06	Column F in position specific scoring matrix
	14	$PSSM_I$	0.60	0.35	1.44	1.06	1.40E-11	Column I in position specific scoring matrix
	15	$PSSM_L$	0.95	0.55	2.05	1.56	1.80E-12	Column L in position specific scoring matrix
	16	$PSSM_V$	0.72	0.45	1.72	1.22	2.00E-13	Column V in position specific scoring matrix
	17	$PSSM_W$	0.09	0.09	0.89	0.42	0.099	Column W in position specific scoring matrix
	18	$PSSM_Y$	0.22	0.19	1.24	1.15	0.038	Column Y in position specific scoring matrix
	19	$PSSM_A$	0.71	0.84	1.46	1.58	0.0072	Column A in position specific scoring matrix
	20	$PSSM_G$	0.26	0.36	1.07	1.15	7.40E-09	Column G in position specific scoring matrix
	21	$PSSM_T$	0.46	0.50	0.98	1.32	0.00023	Column T in position specific scoring matrix
	22	$PSSM_S$	0.51	0.68	0.86	1.24	3.00E-10	Column S in position specific scoring matrix
	23	$PSSM_Q$	0.31	0.41	0.66	0.81	4.80E-07	Column Q in position specific scoring matrix
	24	$PSSM_N$	0.26	0.39	0.75	0.83	5.90E-10	Column N in position specific scoring matrix
	25	$PSSM_E$	0.38	0.64	1.13	1.63	1.80E-08	Column E in position specific scoring matrix
	26	$PSSM_D$	0.28	0.48	0.90	1.70	1.70E-16	Column D in position specific scoring matrix
	27	$PSSM_H$	0.17	0.19	0.58	0.98	0.00077	Column H in position specific scoring matrix
	28	$PSSM_R$	0.41	0.51	1.07	1.01	0.0033	Column R in position specific scoring matrix
	29	$PSSM_K$	0.45	0.61	1.09	1.23	5.60E-06	Column K in position specific scoring matrix
	30	$PSSM_P$	0.14	0.17	0.67	0.70	0.00039	Column P in position specific scoring matrix
	31	R_i	1.26	1.18	1.47	1.44	0.01	Conservation score at mutant position *i*
	Local structure feature derived from sequence							
	32	Coil	0.13	0.21	0.33	0.38	0.00098	Coil probability
	33	Helix	0.06	0.11	0.35	0.38	3.50E-05	Helix probability
	34	Beta	0.07	0.04	0.32	0.24	5.30E-05	Strand probability
	35	SA	0.25	0.35	0.26	0.33	3.10E-15	Solvent accessibility of wild type residue
	36	Phi	-85.30	-83.60	-82.01	-82.39	0.04	Torsion angle
	37	Psi	99.30	-2.70	54.39	44.66	0.32	Torsion angle
Threading template-based	*Conservation scores from multiple template alignments*							
	38	wBLOSUM	0.05	0.03	-0.17	-0.34	2.90E-10	Wide-type LOMETS conservation score
	39	mBLOSUM	0.03	0.03	-0.21	-0.36	0.062	Mutant LOMETS conservation score
	Normal mode analyses							
	40	F_w	0.19	0.19	7.71	4.31	0.057	Wild-type LOMETS fluctuation score
	41	F_m	0.19	0.20	9.05	5.39	0.015	Mutant LOMETS fluctuation score
	42	Rmsip	0.52	0.52	0.51	0.50	0.15	Rmsip between wild-type and mutant templates
I-TASSER model-based	*Knowledge-based statistical potentials*							
	43	RW_w	-20309.64	-20309.64	-20328.70	-22284.99	0.036	Pair-wise distance-dependent energy for wild-type protein
	44	$RWplus_w$	-21637.53	-21637.53	-21667.96	-23732.96	0.045	Side-chain orientation-dependent energy for wild-type protein
	45	$DFIRE_w$	-231.69	-231.69	-236.33	-258.69	0.012	Distance-related energy for wild-type protein
	46	$dDFIRE1_w$	-185.04	-185.04	-183.60	-200.34	0.034	Angle-related energy between Hydrogen-bonded atoms for wild-type protein
	47	$dDFIRE2_w$	-23.84	-25.67	-27.92	-30.66	7.70E-05	Angle-related energy between polar and nonpolar atoms for wild-type protein
	48	$dDFIRE3_w$	-13.03	-14.54	-15.31	-16.95	0.00016	Angle-related energy between polar atoms for wild-type protein
	49	$dDFIRE_w$	-9.01	-9.15	-9.50	-10.74	0.00059	Total dDFIRE energy for wild-type protein
	50	RW_m	-20385.69	-20690.87	-20237.75	-22210.12	0.011	Pair-wise distance-dependent energy for mutant protein
	51	$RWplus_m$	-21848.74	-22112.32	-21682.96	-23777.13	0.015	Side-chain orientation-dependent energy for mutant protein
	52	$DFIRE_m$	-249.74	-252.01	-245.75	-267.68	0.0061	Distance-related energy for mutant protein
	53	$dDFIRE1_m$	-186.39	-189.05	-183.39	-200.23	0.0094	Angle-related energy between Hydrogen-bonded atoms for mutant protein
	54	$dDFIRE2_m$	-33.02	-34.25	-34.04	-36.43	0.0035	Angle-related energy between polar and nonpolar atoms for mutant protein
	55	$dDFIRE3_m$	-18.16	-18.29	-18.31	-19.75	0.0017	Angle-related energy between polar atoms for mutant protein
	56	$dDFIRE_m$	-9.11	-9.23	-10.01	-11.27	0.00042	Total dDFIRE energy for mutant protein

Physics-based energy terms from AMBER							
57	Int_{rw}	76.56	80.73	64.42	69.62	0.0019	Internal potential for wild-type residue
58	VDW_{rw}	-6.46	-6.16	-6.16	-6.25	0.086	Van der Waals energy for wild-type residue
59	EEL_{rw}	-71.84	-72.11	-70.74	-70.87	0.44	Electrostatic energy for wild-type residue
60	EGB_{rw}	-2.68	-4.64	-10.27	-20.98	1.80E-11	Polar solvation energy for wild-type residue
61	$ESURF_{rw}$	0.26	0.40	0.35	0.43	4.20E-10	Non-polar solvation energy for wild-type residue
62	$ATotal_{rw}$	-13.02	-16.51	-22.41	-28.04	7.00E-08	Amber Total energy for wild-type residue
63	Int_{rm}	82.70	82.73	75.28	70.94	0.44	Internal potential for mutant residue
64	VDW_{rm}	-4.85	-4.92	-4.19	-4.13	0.45	Van der Waals energy for mutant residue
65	EEL_{rm}	-74.40	-73.80	-71.73	-70.82	0.24	Electrostatic energy for mutant residue
66	EGB_{rm}	-2.73	-2.90	-7.41	-9.69	0.0078	Polar solvation energy for mutant residue
67	$ESURF_{rm}$	0.18	0.35	0.29	0.41	3.40E-20	Non-polar solvation energy for mutant residue
68	$ATotal_{rm}$	-1.91	-2.27	-7.76	-13.29	0.036	Amber Total energy for mutant residue
69	$BOND_{pw}$	80.16	80.16	105.66	109.58	0.097	Bond energy for wild-type protein
70	$ANGLE_{pw}$	344.20	339.43	330.15	350.28	0.41	Angle energy for wild-type protein
71	$DIHED_{pw}$	1664.18	1664.18	1637.82	1741.30	0.27	Dihedral energy for wild-type protein
72	$VDWAALS_{pw}$	-729.48	-729.48	-730.20	-779.07	0.046	Van der Waals energy for wild-type protein
73	ELE_{pw}	-9501.43	-8867.55	-9255.85	-9775.77	0.26	Electrostatic energy for wild-type protein
74	$1\text{-}4VDW_{pw}$	500.73	473.14	484.98	514.10	0.31	1-4 Van der Waals energy for wild-type protein
75	$1\text{-}4ELE_{pw}$	5336.95	5404.14	5553.33	5778.11	0.12	1-4 Electrostatic energy for wild-type protein
76	EGB_{pw}	-1663.26	-1587.12	-1629.87	-1644.02	0.017	Polar solvation energy for wild-type protein
77	$ESURF_{pw}$	49.09	49.09	49.33	51.02	0.19	Non-polar solvation energy for wild-type protein
78	$Ggas_{pw}$	-1610.02	-1610.02	-1874.11	-2061.47	0.011	Total gas phase free energy for wild-type protein
79	$Gsolv_{pw}$	-1613.03	-1515.65	-1580.53	-1593.00	0.018	Total solvation free energy for wild-type protein
80	$ATotal_{pw}$	-3561.10	-3561.10	-3454.64	-3654.47	0.022	Total energy for wild-type protein
81	$BOND_{pm}$	91.11	91.72	144.18	151.20	0.16	Bond energy for mutant protein
82	$ANGLE_{pm}$	356.20	356.91	371.33	405.87	0.06	Angle energy for mutant protein
83	$DIHED_{pm}$	1572.00	1575.76	1554.64	1657.87	0.18	Dihedral energy for mutant protein
84	$VDWAALS_{pm}$	-636.74	-622.78	-620.93	-640.04	0.26	Van der Waals energy for mutant protein
85	ELE_{pm}	-8478.29	-7963.77	-8450.43	-8944.69	0.27	Electrostatic energy for mutant protein
86	$1\text{-}4VDW_{pm}$	479.53	481.32	485.26	521.71	0.08	1-4 Van der Waals energy for mutant protein
87	$1\text{-}4ELE_{pm}$	5326.98	5311.43	5516.64	5725.47	0.075	1-4 Electrostatic energy for mutant protein
88	EGB_{pm}	-2287.72	-2162.95	-2311.74	-2317.10	0.014	Polar solvation energy for mutant protein
89	$ESURF_{pm}$	54.76	54.65	55.19	56.87	0.48	Non-polar solvation energy for mutant protein
90	$Ggas_{pm}$	-808.20	-967.58	-999.30	-1122.60	0.0019	Total gas phase free energy for mutant protein
91	$Gsolv_{pm}$	-2246.28	-2107.90	-2256.55	-2260.23	0.014	Total solvation free energy for mutant protein
92	$ATotal_{pm}$	-3502.39	-3522.71	-3255.85	-3382.83	0.1	Total energy for mutant protein
Empirical potential from FoldX							
93	$FTotal_{w}$	237.72	234.20	241.85	249.49	0.029	Overall stability for wild-type protein
94	$BBHbond_{w}$	-75.45	-75.45	-72.91	-78.88	0.038	Contribution of backbone Hbonds for wild-type protein
95	$SCHbond_{w}$	-25.86	-24.57	-26.09	-27.51	0.34	Contribution of sidechain-sidechain and sidechain-backbone Honds for wild-type protein
96	F_VDW_{w}	-144.51	-144.51	-141.96	-152.34	0.14	Contribution of the van der Waals for wild-type protein
97	F_ELE_{w}	-9.69	-9.28	-10.66	-10.58	0.0043	Electrostatic interactions for wild-type protein
98	SP_{w}	237.79	227.37	225.90	240.78	0.21	Penalization for burying polar groups for wild-type protein
99	SH_{w}	-183.84	-183.84	-178.69	-191.82	0.084	Contribution of hydrophobic groups for wild-type protein
100	$VDWC_{w}$	96.59	94.64	112.81	118.36	0.0055	Energy penalization due to van der Waals' clashes for wild-type protein
101	$EntSC_{w}$	77.59	76.71	79.62	84.13	0.31	Entropy cost of fixing the side chain for wild-type protein
102	$EntMC_{w}$	235.62	232.66	224.66	237.25	0.27	Entropy cost of fixing the main chain for wild-type protein
103	TC_{w}	26.92	26.92	30.44	31.67	0.11	Van der Waals' torsional clashes for wild-type protein
104	$BBVDW_{w}$	83.40	83.40	91.46	99.66	0.0061	Backbone-backbone wan der Waals for wild-type protein
105	$EleHelix_{w}$	-1.82	-1.82	-1.64	-1.83	0.036	Electrostatic contribution of the helix dipole for wild-type protein
106	Ion_{w}	0.16	0.16	0.30	0.29	0.033	Ionization for wild-type protein
107	$FTotal_{m}$	205.12	205.32	238.91	257.71	0.47	Overall stability for mutant protein
108	$BBHbond_{m}$	-66.57	-66.64	-67.17	-73.49	0.0099	Contribution of backbone Hbonds for mutant protein
109	$SCHbond_{m}$	-10.34	-10.19	-10.81	-11.55	0.0063	Contribution of sidechain-sidechain and sidechain-backbone Honds for mutant protein
110	F_VDW_{m}	-133.88	-135.08	-134.19	-145.13	0.027	Contribution of the van der Waals for mutant protein
111	F_ELE_{m}	-3.36	-2.91	-3.20	-3.15	0.00013	Electrostatic interactions for mutant protein
112	SP_{m}	200.30	201.60	201.41	217.58	0.055	Penalization for burying polar groups for mutant protein
113	SH_{m}	-173.61	-176.02	-172.51	-186.61	0.028	Contribution of hydrophobic groups for mutant protein
114	$VDWC_{m}$	83.26	106.94	125.29	142.48	0.0087	Energy penalization due to van der Waals' clashes for mutant protein

115	EntSC$_m$	64.00	64.38	62.62	66.95	0.18	Entropy cost of fixing the side chain for mutant protein
116	EntMC$_m$	232.98	228.70	222.22	234.94	0.29	Entropy cost of fixing the main chain for mutant protein
117	TC$_m$	15.85	15.08	15.79	16.32	0.16	Van der Waals' torsional clashes for mutant protein
118	BBVDW$_m$	79.59	79.85	89.25	97.76	0.0022	Backbone-backbone wan der Waals for mutant protein
119	EleHelix$_m$	-0.39	-0.70	-0.94	-1.14	0.009	Electrostatic contribution of the helix dipole for mutant protein
120	Ion$_m$	0.19	0.16	0.25	0.30	0.38	Ionization for mutant protein

DM[a]: Destablization mutations with $\Delta\Delta G<0$.

SM[b]: Stablization mutations with $\Delta\Delta G>0$.

M-W test[c]: Mann-Whitney test to determine whether two datasets are drawn from the same distribution. If the p-value is lower than 0.05, the hypothesis that the distributions of the two datasets are the same can be rejected.

Table S2. Summary of mutation stability predictions on the p53 protein by different methods.

Methods	**γ**	**σ (Kcal/mol)**
I-Mutent3.0	0.57	1.48
INPS	0.71	1.49
mCSM	0.67	1.40
PoPMuSiC	0.56	1.58
STRUM	0.69	1.34

Table S3. Summary of performance trained by different groups of features. The data are generated by protein-level 5-fold cross validation on the S2648 dataset.

Features groups	**N**[a]	**Rank**[b]	**Top *n***[c]	**γ**[d]	**σ**[e]
Physicochemical + MSA + local structure prediction	37	3, 5, 7, 9, 2, 37, 16, 14, 15, 31, 30, 12, 36, 21, 28, 20, 18, 24, 26, 13, 27, 23, 19, 25, 17, 29, 22, 11, 32, 34, 33, 35, 8, 6, 4, 10, 1	Top 5	0.36	1.44
			Top 10	0.42	1.40
			Top 20	0.46	1.35
			Top 37	0.47	1.34
Physicochemical + Threading-based features	15	42, 41, 40, 38, 39, 3, 9, 7, 5, 2, 1, 6, 10, 4, 8	Top 5	0.14	1.64
			Top 10	0.38	1.43
			Top 15	0.41	1.41
Physicochemical + I-TASSER-based features	88	64, 65, 68, 66, 63, 67, 58, 61, 81, 59, 57, 60, 118, 62, 56, 117, 111, 90, 84, 55, 82, 89, 3, 112, 114, 83, 54, 115, 107, 87, 116, 52, 113, 85, 91, 5, 92, 53, 86, 108, 88, 110, 51, 50, 7, 9, 119, 109, 2, 1, 120, 4, 8, 6, 10, 105, 97, 95, 103, 79, 106, 76, 49, 104, 100, 70, 93, 47, 78, 75, 77, 72, 80, 69, 94, 48, 102, 71, 45, 73, 46, 101, 99, 74, 43, 44, 96, 98	Top 5	0.19	1.57
			Top 10	0.31	1.50
			Top 20	0.37	1.42
			Top 50	0.47	1.32
			Top 88	0.49	1.31

[a] N: Number of features
[b] Rank: Rank of features in Table S1 by the importance as calculated by the Scikit-learn program
[c] Top *n*: The first *n* features are used to train the predictor
[d] γ: Pearson correlation coefficient between predicted and experimental $\Delta\Delta G$
[e] σ: root mean square error of $\Delta\Delta G$ prediction in Kcal/mol

Table S4. Summary of performance trained using different number of top features selected from different feature groups. The data are generated by protein-level 5-fold cross validation on the S2648 dataset.

Top *n*[a]	**Nt**[b]	γ[c]	σ[d]
Top 5	15	0.44	1.37
Top 10	25	0.48	1.32
Top 20	50	0.50	1.29
Top 50	86	0.51	1.27

[a] Top *n*: The first *n*-ranking features in each feature group are selected and merged into a set of training features.
[b] Nt: Total number of features in the final training feature set
[c] γ: Pearson correlation coefficient between predicted and experiment $\Delta\Delta G$
[d] σ: root mean square error of $\Delta\Delta G$ prediction in Kcal/mol

Table S5. Dependence of STRUM performance on structure accuracy. Data are generated from the 'mutation-level' 5-fold cross validation on S2648.

Protein structure	Total		TM-score ≥ 0.5 (109 proteins)		TM-score < 0.5 (23 proteins)	
	γ[a]	σ[b]	γ[a]	σ[b]	γ[a]	σ[b]
I-TASSER model	0.77	0.94	0.78	0.94	0.67	0.98
Native structure	0.78	0.92	0.78	0.94	0.72	0.92

[a]γ: PCC between predicted and experiment $\Delta\Delta G$

[b]σ: RMSE of $\Delta\Delta G$ prediction

Table S6 Performance of STRUM when using different quality of I-TASSER models. Data are from protein-level 5-fold cross validation on S2648.

Quality of structure models	**#proteins**	**#mutations**	$\mathbf{P_{good}}$[b]	$\mathbf{P_{bad}}$[a]
TM-score<0.5	23	423	0.07	0.09
TM-score≥0.5	109	2225	0.10	0.06

[a]P_{good}: Portion of good predictions that are defined as those with the predicted and experimental $\Delta\Delta G$ having the same sign (i.e. both >0 or <0) and the difference below 0.01.

[b]P_{bad}: Portion of bad predictions that are defined as those with the predicted and experimental $\Delta\Delta G$ having the opposite sign and the difference above 1.0.

Table S7. Pearson correlation coefficient (PCC) of the NMA feature values and the experimental $\Delta\Delta G$ when different templates are used to calculate the NMA features.

Features	**TM-score**[a]	**PCC**[b]			
		Fw[c]	**Fm**[d]	**Fm-Fw**	**Rmsip**[e]
First LOMETS template (T1)	0.651	0.047	0.030	0.019	-0.049
10th LOMETS template (T10)	0.627	0.023	0.020	0.013	-0.077
The last LOMETS template (TN)	0.599	0.029	0.020	0.006	0.050
The first 10 LOMETS templates (TF10)	0.640	0.032	0.030	0.020	-0.050
The last 10 LOMETS templates (TL10)	0.611	0.023	0.020	0.009	-0.037

[a] TM-score: Average TM-score of the Modeller models built on LOMET templates

[b] PCC: Pearson correlation coefficient between NMA features and the experimental $\Delta\Delta G$.

[c] Fw: Wild-type template conformational fluctuation score defined by Eq. 8 (Feature #40 in Table S1)

[d] Fm: Mutant template conformational fluctuation score defined by Eq. 8 (Feature #41 in Table S1)

[e] Rmsip: Root mean square inner product between wild-type and mutation templates defined by Eq. 8 (Feature #42 in Table S1)

Table S8. Performance of STRUM when replacing the full-set NMA features by the NMA features from the five template selections. The results are from protein-level 5-fold cross-validation on S2648 set.

Features	**PCC**	**RMSE**
STRUM with T1	0.53	1.27
STRUM with T10	0.53	1.25
STRUM with TN	0.53	1.26
STRUM with TF10	0.53	1.26
STRUM with TL10	0.53	1.25