SAMPDI-3D: predicting the effects of protein and DNA mutations on protein–DNA interactions

Gen Li [1], Shailesh Kumar Panday[1], Yunhui Peng[1] and Emil Alexov [1,*]

1 Department of Physics and Astronomy, Clemson University, Clemson, SC 29634, USA; genl@clemson.edu (G.L.); spanday@clemson.edu (S. P.); yunhuip@g.clemson.edu (Y. P.); ealexov@clemson.edu (E.A.)

* Correspondence: ealexov@clemson.edu;

Blind dataset preparation for protein mutation (T227)

A database of the binding affinity changes upon single base pair substitution in the protein-DNA interactions have been constructed using the recent experimental data (1,2). To construct the database, we took the processed M-word binding scores of the transcriptional factors (TFs) binding to DNA and these binding scores are calculated based on observed experimental enrichment counts from the HT-SELEX experiments (1). The flitted high-quality HT-SELEX experiments data initially comprises of 219 transcriptional factors (TFs) from 29 families. Since the structural information is crucial for our database, we firstly filtered out the TFs without available protein-DNA complex structures in Protein Data Bank. Next, we removed the TFs for which the DNA sequences in the corresponding 3D structure does not match the sequence of the DNA used in the experiment. After filtering, for each remaining TFs, we collected the M-word binding scores (ΔM) of the DNA sequences under single base pair substitution in respect to the sequence in PDB structures. In total, we collected binding score for 227 DNA single base pair substitution from 18 TFs. We use the $\Delta\Delta M=\ln(\Delta M_w/\Delta M_m)$ to reflect the change M-word binding scores (ΔΔM) of single base pair substitution. In this way, the larger ΔΔM means binding affinity decrease, smaller ΔΔM means binding affinity increase.

Blind dataset preparation for disruptive and non-disruptive protein mutations (D101)

First, we downloaded a data set containing 283 mutation effect descriptions from the dbAMEPNI database. Then, removed the structure containing the following content: hybrid DNA/RNA, confusing description of mutation effects, without DNA, modified DNA, mutation site interact with small molecules and unreasonable structure. After filtering, our final blind dataset includes 101 alanine mutations in 28 proteins.

Table S1. Performance of SAMPDI-3D and other methods in predicting disruptive and non-disruptive protein mutations.

Method	Accuracy	Precision	Recall	MCC	AUC
SAMPDI-3D	0.94	0.88	0.88	0.84	0.96
SAMPDI	0.77	0.50	0.63	0.41	0.67
PremPDI	0.86	1.00	0.38	0.56	0.69
mCSM-NA	0.89	0.83	0.63	0.66	0.82

Table S2. Number of disruptive and non-disruptive mutations in training and blind test datasets.

Dataset	Disruptive	Non-disruptive
S419	147	272
S200	53	147
D463	149	314
D101	50	51

We classify the disruptive mutations as $|\Delta\Delta G|>1$ kcal/mol and non-disruptive as $|\Delta\Delta G|<1$ kcal/mol

Table S3. Number of features in each category for the model of predicting protein mutations or DNA mutations.

Feature groups	Numbers	
	Predicting mutations in protein	Predicting mutations in DNA
Physicochemical properties	9	None
Protein secondary structure element	6	
Amino acid properties	4	None
Protein-DNA interactions	4	
Experimental condition	1	None
Knowledge-based	None	3
Structural feature of mutation site	None	18

Table S4. Performance for Interfacial and Non-interfacial protein mutations on S200 dataset

Method	Interfacial mutations		Non-interfacial mutations	
	PCC	MSE	PCC	MSE
SAMPDI-3D	0.39	1.08	0.43	0.79
SAMPDI	-0.01	1.58	0.21	0.96
PremPDI	0.17	1.78	0.37	1.13
mCSM-NA	0.17	2.71	0.37	1.77
FoldX	0.01	4.44	0.09	8.71

Table S5 Performance for Interfacial and Non-interfacial DNA mutations on T227 dataset

Method	Interfacial mutations	No-interfacial mutations
	PCC	PCC
SAMPDI-3D	0.28	0.44
FoldX	-0.15	0.21

1. Yang, L., Orenstein, Y., Jolma, A., Yin, Y., Taipale, J., Shamir, R. and Rohs, R. (2017) Transcription factor family-specific DNA shape readout revealed by quantitative specificity models. Molecular systems biology, 13, 910.
2. Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K.R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G. *et al.* (2013) DNA-binding specificities of human transcription factors. Cell, 152, 327-339.