Supplementary material of " A max-margin model for predicting residue-base contacts in protein-RNA interactions "

Kengo Sato[1,*] Shunya Kashiwagi[1] Yasubumi Sakakibara[1]

S1 Derivation of scoring functions for max-margin training

The loss function of the prediction $\hat{z}$ against the positive data z (Eq. (14) in the main paper) can be transformed into the following using binary-valued variables:

$$
\begin{aligned}
\Delta(z,\hat{z}) =& \delta^{\text{FN residue}}(\#\text{ of false negative residues}) \\
&+ \delta^{\text{FP residue}}(\#\text{ of false positive residues}) \\
&+ \delta^{\text{FN base}}(\#\text{ of false negative bases}) \\
&+ \delta^{\text{FP base}}(\#\text{ of false positive bases}) \\
&+ \delta^{\text{FN contact}}(\#\text{ of false negative contacts}) \\
&+ \delta^{\text{FP contact}}(\#\text{ of false positive contacts}) \\
=& \delta^{\text{FN residue}} \sum_{i=1}^{N_p} I(x_i = 1) I(\hat{x}_i = 0) + \delta^{\text{FP residue}} \sum_{i=1}^{N_p} I(x_i = 0) I(\hat{x}_i = 1) \\
&+ \delta^{\text{FN base}} \sum_{j=1}^{N_r} I(y_j = 1) I(\hat{y}_j = 0) + \delta^{\text{FP base}} \sum_{j=1}^{N_r} I(y_j = 0) I(\hat{y}_j = 1) \\
&+ \delta^{\text{FN contact}} \sum_{i=1}^{N_p} \sum_{j=1}^{N_r} I(z_{ij} = 1) I(\hat{z}_{ij} = 0) + \delta^{\text{FP contact}} \sum_{i=1}^{N_p} \sum_{j=1}^{N_r} I(z_{ij} = 0) I(\hat{z}_{ij} = 1) \\
=& \sum_{i=1}^{N_p} \left\{ \delta^{\text{FN residue}} x_i (1 - \hat{x}_i) + \delta^{\text{FP residue}} (1 - x_i) \hat{x}_i \right\} \\
&+ \sum_{j=1}^{N_r} \left\{ \delta^{\text{FN base}} y_j (1 - \hat{y}_j) + \delta^{\text{FP base}} (1 - y_j) \hat{y}_j \right\} \\
&+ \sum_{i=1}^{N_p} \sum_{j=1}^{N_r} \left\{ \delta^{\text{FN contact}} z_{ij} (1 - \hat{z}_{ij}) + \delta^{\text{FP contact}} (1 - z_{ij}) \hat{z}_{ij} \right\}.
\end{aligned}
$$

Here, $I(x_i = 1) I(\hat{x}_i = 0) = 1$ if $\hat{x}_i$ is a false negative and 0 otherwise, and $I(x_i = 0) I(\hat{x}_i = 1) = 1$ if $\hat{x}_i$ is a false positive and 0 otherwise. We also use the fact that $I(x_i = 1) = x_i$ and $I(x_i = 0) = 1 - x_i$.

*to whom correspondence should be addressed

[1]Department of Biosciences and Informatics, Keio University, 3–14–1 Hiyoshi, Kohoku-ku, Yokohama 223–8522, Japan

Therefore, the first term of Eq. (13) in the main paper can be simplified into:

$$
\begin{aligned}
f_\lambda(P,R,\hat{z}) + \Delta(z,\hat{z}) &= \sum_{i=1}^{N_p} u_i \hat{x}_i + \sum_{j=1}^{N_r} v_j \hat{y}_j + \sum_{i=1}^{N_p}\sum_{j=1}^{N_r} w_{ij}\hat{z}_{ij} \\
&+ \sum_{i=1}^{N_p} \left\{ \delta^{\text{FN residue}} x_i (1-\hat{x}_i) + \delta^{\text{FP residue}} (1-x_i)\hat{x}_i \right\} \\
&+ \sum_{j=1}^{N_r} \left\{ \delta^{\text{FN base}} y_j (1-\hat{y}_j) + \delta^{\text{FP base}} (1-y_j)\hat{y}_j \right\} \\
&+ \sum_{i=1}^{N_p}\sum_{j=1}^{N_r} \left\{ \delta^{\text{FN contact}} z_{ij} (1-\hat{z}_{ij}) + \delta^{\text{FP contact}} (1-z_{ij})\hat{z}_{ij} \right\} \\
&= \sum_{i=1}^{N_p} \left\{ \left[u_i - \delta^{\text{FN residue}} x_i + \delta^{\text{FP residue}} (1-x_i) \right] \hat{x}_i + \delta^{\text{FN residue}} x_i \right\} \\
&+ \sum_{j=1}^{N_r} \left\{ \left[v_j - \delta^{\text{FN base}} y_j + \delta^{\text{FP base}} (1-y_j) \right] \hat{y}_j + \delta^{\text{FN base}} y_j \right\} \\
&+ \sum_{i=1}^{N_p}\sum_{j=1}^{N_r} \left\{ \left[w_{ij} - \delta^{\text{FN contact}} z_{ij} + \delta^{\text{FP contact}} (1-z_{ij}) \right] \hat{z}_{ij} + \delta^{\text{FN contact}} z_{ij} \right\} \\
&= \sum_{i=1}^{N_p} \bar{u}_i \hat{x}_i + \sum_{j=1}^{N_r} \bar{v}_j \hat{y}_j + \sum_{i=1}^{N_p}\sum_{j=1}^{N_r} \bar{w}_{ij}\hat{z}_{ij} + Const,
\end{aligned}
$$

where

$$
\begin{aligned}
\bar{u}_i &= u_i - \delta^{\text{FN residue}} x_i + \delta^{\text{FP residue}} (1-x_i) \\
&= \begin{cases} u_i - \delta^{\text{FN residue}} & (\text{if } x_i{=}1) \\ u_i + \delta^{\text{FP residue}} & (\text{if } x_i{=}0) \end{cases} \\
\bar{v}_i &= v_j - \delta^{\text{FN base}} y_j + \delta^{\text{FP base}} (1-y_j) \\
&= \begin{cases} v_i - \delta^{\text{FN base}} & (\text{if } y_j{=}1) \\ v_i + \delta^{\text{FP base}} & (\text{if } y_j{=}0) \end{cases} \\
\bar{w}_{ij} &= w_{ij} - \delta^{\text{FN contact}} z_{ij} + \delta^{\text{FP contact}} (1-z_{ij}) \\
&= \begin{cases} w_{ij} - \delta^{\text{FN contact}} & (\text{if } z_{ij}{=}1) \\ w_{ij} + \delta^{\text{FP contact}} & (\text{if } z_{ij}{=}0) \end{cases} \\
Const &= \sum_{i=1}^{N_p} \delta^{\text{FN residue}} x_i + \sum_{j=1}^{N_r} \delta^{\text{FN base}} y_j + \sum_{i=1}^{N_p}\sum_{j=1}^{N_r} \delta^{\text{FN contact}} z_{ij}
\end{aligned}
$$

The last equation indicates that we can maximize the first term of the objective function (13) by replacing scores u_i, v_j and w_{ij} with a constant difference $Const$ that is independent of $\hat{x}_i, \hat{y}_j$ and $\hat{z}_{ij}$.

S2 The maximum number of contacts

To determine the upper bound of the number of contacts for each residue and base (X_i in Eq. (11) and Y_j in Eq. (12) in the main paper), we investigated the distribution of the number of contacts for each residue and base in the dataset assembled in our study. We predicted structural profiles for proteins and RNAs, then counted the contacts for each type of structural profiles. Tables S1 and S2 show the distributions of the number of contacts for each type of structural profiles. We determined the upper bound of the number of contacts for each structural profile such that it can cover the 90 % of contacts observed in our dataset.

Table S1: The number of contacts for each structural element of residues

# of contacts	coil	α helix	β sheet
1	761	335	205
2	704	283	162
3	239	73	38
4	126	25	14
5	53	4	3
6	22	3	0
7	4	1	0

Table S2: The number of contacts for each structural element of bases

# of contacts	bulge	external	hairpin	internal	multibranch	stack
1	9	35	100	57	228	339
2	5	36	66	44	138	229
3	4	21	47	24	74	174
4	3	27	31	23	66	112
5	2	23	18	14	34	67
6	0	17	8	6	13	33
7	0	15	11	1	6	9
8	0	7	4	0	2	2
9	0	8	1	0	2	2
10	0	3	2	0	0	0
11	0	0	1	0	0	0
12	0	0	1	0	0	0
13	0	0	0	0	0	0
14	0	1	0	0	0	0
15	0	1	1	0	0	0

Table S3: The number of residue features with non-zero weights

Type	Context len.	# of features	> 0	< 0
Residues	3	20^3	93	219
	5	20^5	96	224
Simplified alphabets (10 groups)	5	10^5	94	224
	7	10^7	95	221
Simplified alphabets (4 groups)	5	4^5	56	161
	7	4^7	91	218
Secondary structures	3	3^3	1	14
	5	3^5	8	29

Table S4: The number of base features with non-zero weights

Type	Context len.	# of features	> 0	< 0
Bases	3	4^3	0	64
	5	4^5	34	223
Secondary structures	3	6^3	0	34
	5	6^5	4	81

S3 Feature weights

Tables S3–S5 show the number of features with non-zero weights for each feature, indicating that only 10,594 features have non-zero weights (> 0: 2,870 and < 0: 7,724), while the number of potential features is more than 4 billion. This is because ℓ_1 regularization leads to sparse coding.

The 20 largest feature weights among all the features are shown in Table S6. The first two columns indicate the type of features for residues and bases. By contexts of features in the next two columns, each feature can be identified. The contexts of each feature is described by representative character codes as shown in Tables S7, S8 and S9 for structural profiles of residues and bases, simplified alphabets of 10 groups, and those of 4 groups, respectively. Table S6 suggests that the weight for long continuous coil regions in protein sequences have large positive values.

Table S5: The number of residue–base contact features with non-zero weights

Type					
Residue	Base	Context len.	# of features	> 0	< 0
Residues	Bases	3	$20^3 \times 4^3$	251	562
		5	$20^5 \times 4^5$	254	317
Secondary structures	Secondary structures	3	$3^3 \times 6^3$	0	412
		5	$3^5 \times 6^5$	103	552
Simplified alphabets (10 groups)	Bases	3	$10^3 \times 4^3$	224	572
		5	$10^5 \times 4^5$	253	324
Simplified alphabets (10 groups)	Secondary structures	3	$10^3 \times 6^3$	231	640
		5	$10^5 \times 6^5$	253	324
Simplified alphabets (4 groups)	Bases	3	$4^3 \times 4^3$	123	491
		5	$4^5 \times 4^5$	237	442
Simplified alphabets (4 groups)	Secondary structures	3	$4^3 \times 6^3$	107	460
		5	$4^5 \times 6^5$	246	656

Table S6: Top 20 of feature weights

Feature type		Context		
Residue	Base	Residue	Base	Weight
Secondary structures	Secondary structures	CCCCC	SMMMM	5.54003
Secondary structures	Secondary structures	CCCCC	SSSSM	5.43940
Secondary structures	—	CCCCC	—	5.12261
Secondary structures	Secondary structures	CCCCC	SSSSH	4.68329
Secondary structures	Secondary structures	CCCCC	MSSSS	4.34785
Secondary structures	Secondary structures	CCCCC	ISSSS	3.71432
Secondary structures	Secondary structures	CCCCC	HSSSS	2.79177
Secondary structures	Secondary structures	CCCCC	SSSSI	2.56748
Secondary structures	Secondary structures	CCCCC	MMMMS	2.18214
Secondary structures	Secondary structures	CCCCC	SHHHH	2.11508
Secondary structures	Secondary structures	CCCCC	HHHHS	2.10132
Secondary structures	Secondary structures	CCCCC	HHHHH	1.86373
Secondary structures	Secondary structures	ECCCC	SSSSS	0.794036
Simplified alphabets (4 groups)	Secondary structures	AEE	SSS	0.540660
Simplified alphabets (4 groups)	Secondary structures	EEE	SSS	0.533426
Simplified alphabets (4 groups)	Secondary structures	AEA	SSS	0.516197
Secondary structures	—	CCC	—	0.511168
Simplified alphabets (4 groups)	Secondary structures	EEA	SSS	0.506299
Simplified alphabets (4 groups)	Secondary structures	AEF	SSS	0.466757
Secondary structures	Secondary structures	CCCCC	MMSSS	0.448178

Table S7: Structural profiles coding

Residue	α helix	β sheet	Coil			
	H	E	C			
Base	External	Hairpin	Internal	Bulge	Multibranch	Stack
	E	H	I	B	M	S

Table S8: Coding of simplified alphabets (10 groups)

amino acids	LVIM	C	A	G	ST	P	FYW	EDNQ	KR	H
coding	L	C	A	G	S	P	F	E	K	H

Table S9: Coding of simplified alphabets (4 groups)

amino acids	LVIMC	AGSTP	FYW	EDNQKRH
coding	L	A	F	E

S4 Dataset

Table S10 shows PDB IDs and chain IDs used in our dataset.

Table S10: PDB ID and chain IDs used in our dataset

PDBID	Protein chain	RNA chain
1A9N	C	R
1AV6	A	B
1C0A	A	B
1DDL	A	D
1DDL	A	E
1DFU	P	M
1DFU	P	N
1DI2	A	C
1DI2	A	D
1DI2	A	E
1F7U	A	B
1FEU	A	B
1FEU	A	C
1FEU	A	E
1FJG	M	A
1FXL	A	B
1GAX	A	C
1GTF	L	W
1H4S	A	T
1HQ1	A	B
1J1U	A	B
1JBS	A	C
1JID	A	B
1K8W	A	B
1M8W	A	C
1M8W	A	E
1MMS	A	C
1N78	A	C
1OOA	A	C
1Q2R	A	E
1QF6	A	B
1SDS	C	D
1SDS	C	F
1SER	A	T
1VQ8	1	0
1VQ8	3	0
1VQ8	A	0
1VQ8	B	0
1VQ8	C	0
1VQ8	D	0
1VQ8	D	9
1VQ8	E	0
1VQ8	H	0
1VQ8	H	9
1VQ8	J	0
1VQ8	K	0

Table S10: PDB ID and chain IDs used in our dataset (cont.)

PDBID	Protein chain	RNA chain
1VQ8	L	0
1VQ8	M	0
1VQ8	N	0
1VQ8	N	9
1VQ8	O	0
1VQ8	P	0
1VQ8	Q	0
1VQ8	Q	9
1VQ8	R	0
1VQ8	U	0
1VQ8	V	0
1VQ8	W	0
1VQ8	W	9
1VQ8	X	0
1WPU	A	C
1YZ9	A	C
1YZ9	A	D
1YZ9	A	E
1YZ9	A	F
1ZH5	A	C
1ZH5	A	D
1ZHO	A	B
1ZHO	A	H
2A8V	A	D
2ANR	A	B
2ASB	A	B
2BGG	A	P
2BGG	A	Q
2BU1	A	R
2E9T	A	B
2E9T	A	C
2FK6	A	R
2FMT	A	C
2GIC	A	R
2J01	I	A
2J01	R	A
2Q66	A	X
2R8S	L	R
2VQE	B	A
2VQE	C	A
2VQE	D	A
2VQE	F	A
2VQE	G	A
2VQE	H	A
2VQE	I	A
2VQE	J	A
2VQE	K	A
2VQE	N	A
2VQE	P	A

Table S10: PDB ID and chain IDs used in our dataset (cont.)

PDBID	Protein chain	RNA chain
2VQE	R	A
2VQE	S	A
2VQE	T	A
2ZUE	A	B
3GIB	A	H
3IEV	A	D

Table S11: Accuracy under varying $\delta^{\mathrm{FN}\,*}$ with fixing $\delta^{\mathrm{FP}\,*} = 1.0$ and $C = 0.125$.

	Contacts			Binding residues			Binding bases		
$\delta^{\mathrm{FN}\,*}$	PPV	SEN	F	PPV	SEN	F	PPV	SEN	F
2^0	0.4468	0.5836	0.4883	0.5269	0.6351	0.5543	0.5455	0.6405	0.5612
2^1	0.4722	0.5890	0.5076	0.5573	0.6494	0.5765	0.5588	0.6220	0.5618
2^2	0.4797	0.5759	0.5051	0.5766	0.6555	0.5861	0.5653	0.6192	0.5585
2^3	0.4810	0.5494	0.4922	0.5804	0.6368	0.5793	0.5727	0.5925	0.5481
2^4	0.4814	0.4938	0.4592	0.5856	0.6039	0.5608	0.5740	0.5326	0.5063

Table S12: Accuracy under varying C with fixing $\delta^{\mathrm{FP}\,*} = 1.0$ and $\delta^{\mathrm{FN}\,*} = 4.0$.

	Contacts			Binding residues			Binding bases		
C	PPV	SEN	F	PPV	SEN	F	PPV	SEN	F
2^{-4}	0.4862	0.5666	0.5041	0.5806	0.6528	0.5927	0.5693	0.6038	0.5509
2^{-3}	0.4797	0.5759	0.5051	0.5766	0.6555	0.5861	0.5653	0.6192	0.5585
2^{-2}	0.4738	0.5123	0.4766	0.5740	0.5696	0.5475	0.5555	0.5626	0.5406
2^{-1}	0.4395	0.5247	0.4592	0.5388	0.5924	0.5399	0.5345	0.5971	0.5389
2^0	0.4235	0.5204	0.4489	0.5173	0.5951	0.5321	0.5173	0.5921	0.5212
2^1	0.4226	0.4957	0.4381	0.5155	0.5574	0.5113	0.5158	0.5600	0.5116

S5 Hyperparameters

We empirically chose the hyperparameters for the max-margin training: the penalty for positives $\delta^{\mathrm{FN}\,*}$, the penalty for negatives $\delta^{\mathrm{FP}\,*}$, and the weight for ℓ_1 regularization term C. We fixed $\delta^{\mathrm{FP}\,*} = 1.0$ because the balance between $\delta^{\mathrm{FN}\,*}$ and $\delta^{\mathrm{FP}\,*}$ is important. We performed the grid search on $\delta^{\mathrm{FN}\,*} \in \{2^n \mid n = 0, \ldots, 4\}$ and $C \in \{2^n \mid n = -4, \ldots, 1\}$ for the whole dataset. Then, we chose $\delta^{\mathrm{FN}\,*} = 4.0$ and $C = 0.125$ at which good accuracy can be achieved. Tables S11 and S12 show the accuracy under varying $\delta^{\mathrm{FN}\,*}$ and C respectively, indicating that $\delta^{\mathrm{FN}\,*}$ and C are not strongly sensitive to the prediction accuracy.