Recognizing metal and acid radical ion binding sites by integrating *ab initio* modeling with template-based transferals

Xiuzhen Hu, Qiwen Dong, Jianyi Yang, Yang Zhang

Supplementary Information

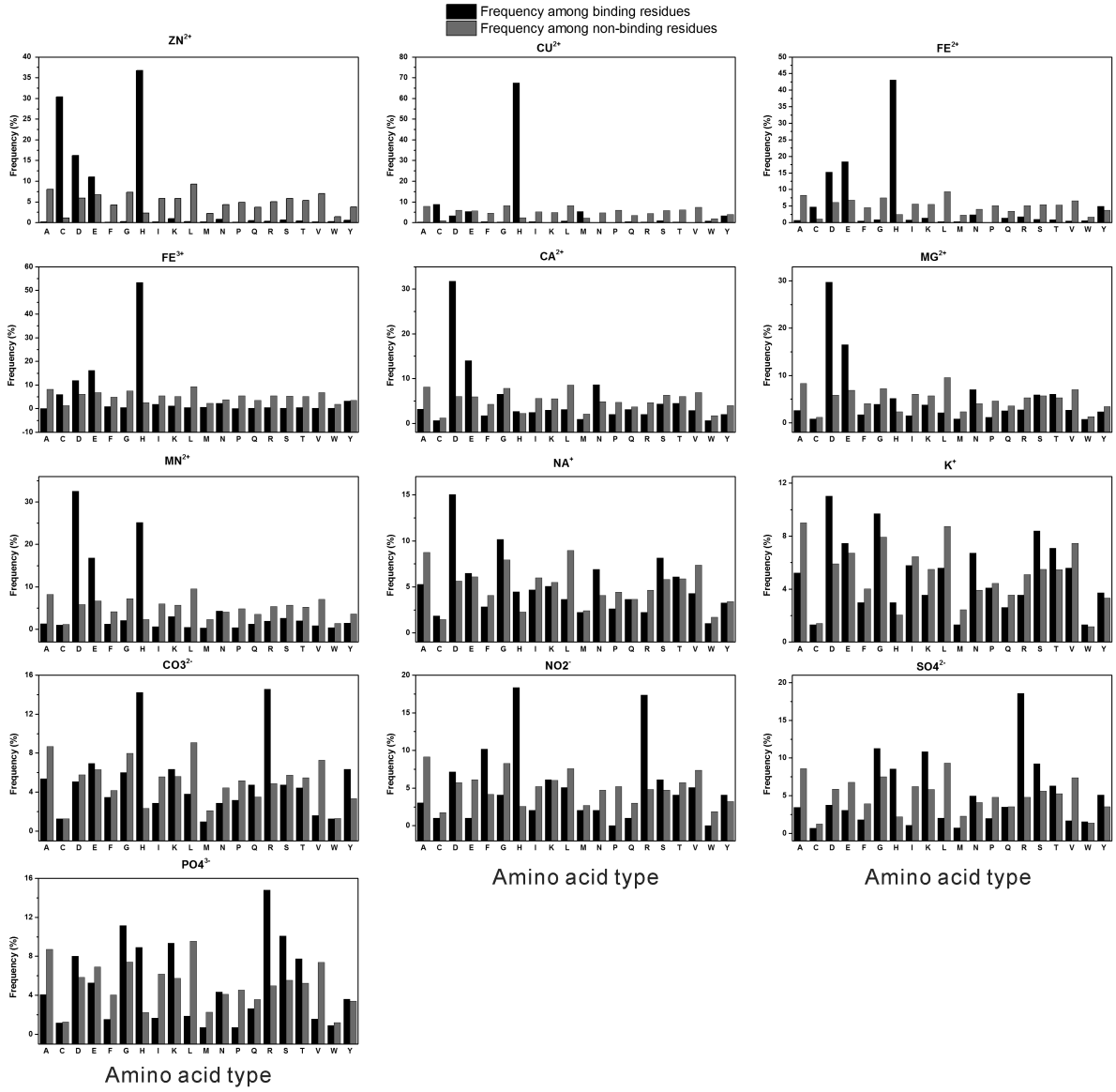

Figure S1. Relative frequency of 20 amino acids appearing on the binding sites (dark) and the non-binding (gray) sites of the 13 ion ligands involved in this study. Data are collected from the BioLiP database.

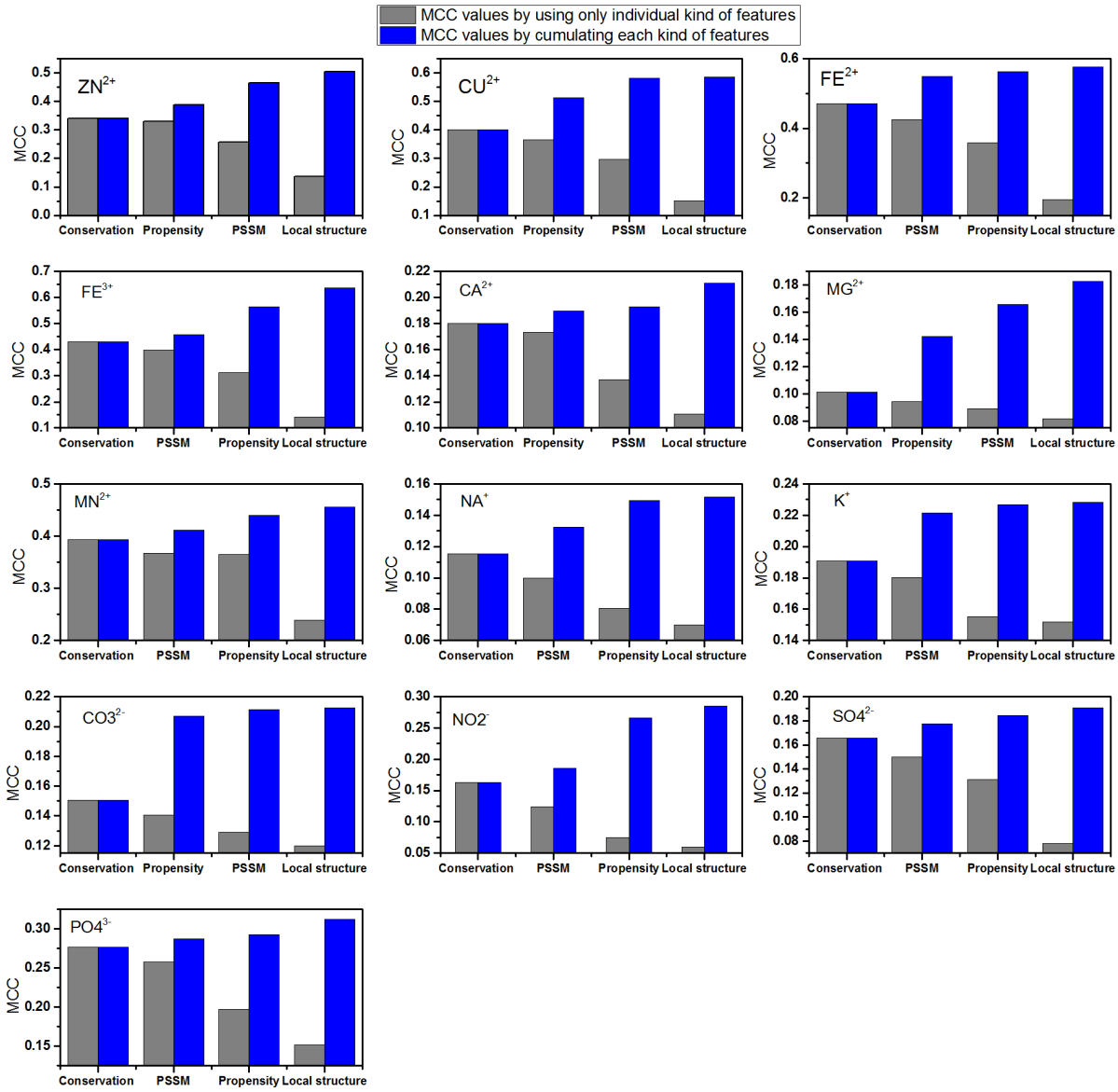

Fig. S2. Performance of IonSeq when trained on different groups of features. The gray bars indicate the results that use individual feature types, where the features are sorted based on the MCC values in each box. The blue bars represent the results obtained by IonSeq trained on multiple features from all entries on the left of the bars.

Table S1. A list of all features used for training IonSeq.

Feature type	Feature	Description	No. of features
Position specific scoring matrix (PSSM)	$PSSM_C$	Column C in position specific scoring matrix	20**L*, *L* is the window width
	$PSSM_M$	Column M in position specific scoring matrix	
	$PSSM_F$	Column F in position specific scoring matrix	
	$PSSM_I$	Column I in position specific scoring matrix	
	$PSSM_L$	Column L in position specific scoring matrix	
	$PSSM_V$	Column V in position specific scoring matrix	
	$PSSM_W$	Column W in position specific scoring matrix	
	$PSSM_Y$	Column Y in position specific scoring matrix	
	$PSSM_A$	Column A in position specific scoring matrix	
	$PSSM_G$	Column G in position specific scoring matrix	
	$PSSM_T$	Column T in position specific scoring matrix	
	$PSSM_S$	Column S in position specific scoring matrix	
	$PSSM_Q$	Column Q in position specific scoring matrix	
	$PSSM_N$	Column N in position specific scoring matrix	
	$PSSM_E$	Column E in position specific scoring matrix	
	$PSSM_D$	Column D in position specific scoring matrix	
	$PSSM_H$	Column H in position specific scoring matrix	
	$PSSM_R$	Column R in position specific scoring matrix	
	$PSSM_K$	Column K in position specific scoring matrix	
	$PSSM_P$	Column P in position specific scoring matrix	
Local structure properties	H	Alpha-helix	6**L*
	E	Beta-strand	
	C	Coil	
	RSA	Relative solvent accessibility	
	PHI	The φ dihedral angle	
	PSI	The ψ dihedral angle	
Conservation scores	RE	Relative entropy conservation score	*L*
	JSD	Jensen-Shannon divergence score	*L*
	SEG	Segment conservation score	*1*
Ion binding propensity	PA	Ion binding propensity for different amino acids	*L*
Summary			*29*L+1*

Table S2. Results of IonSeq on HEM binding site prediction using different feature sets.

Features	Accuracy	Sensitivity	Specificity	MCC
Using the four types of features	0.9227	0.6225	0.9480	0.5179
Using additional physicochemical features	0.8937	0.7839	0.9093	0.5143

Table S3. Summary of the training parameters that are used in IonSeq and IonCom.

Ligand[a]	Method[b]	L[c]	SVM-C[d]	SVM-gamma[e]
Zn^{2+}	IonCom	11	1	0.0009765625
	IonSeq	13	16	0.00195315
Cu^{2+}	IonCom	15	1	0.00097625
	IonSeq	17	2	0.125
Fe^{2+}	IonCom	9	1	0.00097625
	IonSeq	11	1	0.25
Fe^{3+}	IonCom	9	1	0.000390625
	IonSeq	9	2	0.25
Ca^{2+}	IonCom	13	1	0.00097625
	IonSeq	13	4	0.25
Mg^{2+}	IonCom	15	1	0.001953125
	IonSeq	15	1	0.5
Mn^{2+}	IonCom	11	1	0.0009765625
	IonSeq	11	1	0.25
Na^{+}	IonCom	11	8	0.00097625
	IonSeq	13	4	0.0325
K^{+}	IonCom	11	16	0.5
	IonSeq	11	2	0.5
$CO3^{2-}$	IonCom	13	2	0.125
	IonSeq	13	1	0.25
$NO2^{-}$	IonCom	11	1	0.03125
	IonSeq	11	1	0.25
$SO4^{2-}$	IonCom	13	1	0.0001953125
	IonSeq	11	1	0.25
$PO4^{3-}$	IonCom	9	1	0.001953125
	IonSeq	11	1	0.25

[a]The ligand types

[b]The methods proposed in this study

[c]The optimal window length

[d]The cost parameter in SVM

[e]The gamma parameter of RBF kernel in SVM

Table S4. Performance of six involved predictors on the binding-site prediction of 13 metal and acid radical ion ligands.

Ligand	Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	MCC
Zn^{2+}	IonCom	99.48	48.86	99.86	0.5896
	IonSeq	99.21	43.56	99.75	0.5043
	COACH	98.65	57.38	99.14	0.4952
	S-site	99.12	35.49	99.86	0.5133
	COFACTOR	98.83	36.41	99.56	0.4190
	TM-site	97.71	56.43	98.20	0.3794
Cu^{2+}	IonCom	99.26	53.08	99.90	0.6799
	IonSeq	99.01	50.65	99.69	0.5868
	COACH	98.86	61.12	99.39	0.5901
	S-site	99.04	40.19	99.85	0.5607
	COFACTOR	99.04	47.10	99.76	0.5830
	TM-site	97.98	60.37	98.50	0.4564
Fe^{2+}	IonCom	98.73	59.64	99.32	0.5762
	IonSeq	98.84	54.08	99.51	0.5772
	COACH	97.95	66.82	98.42	0.5009
	S-site	98.54	36.68	99.47	0.4265
	COFACTOR	98.41	59.37	99.00	0.5214
	TM-site	96.93	59.55	97.49	0.3835
Fe^{3+}	IonCom	99.32	59.77	99.83	0.6959
	IonSeq	99.21	52.27	99.81	0.6370
	COACH	99.20	62.41	99.67	0.6607
	S-site	98.90	18.22	99.94	0.3749
	COFACTOR	99.19	56.26	99.74	0.6402
	TM-site	98.28	42.14	99.00	0.3760
Ca^{2+}	IonCom	98.87	17.72	99.80	0.2963
	IonSeq	98.18	22.72	99.04	0.2111
	COACH	96.53	31.59	97.47	0.2048
	S-site	98.54	14.96	99.77	0.2632
	COFACTOR	98.06	12.03	99.32	0.1480
	TM-site	96.62	30.28	97.59	0.2010
Mg^{2+}	IonCom	99.47	25.32	99.86	0.3425
	IonSeq	99.49	5.57	99.98	0.1825
	COACH	97.96	44.52	98.40	0.2817
	S-site	99.08	28.80	99.67	0.3439
	COFACTOR	98.58	29.27	99.16	0.2504
	TM-site	96.88	42.41	97.33	0.2117
Mn^{2+}	IonCom	98.95	48.65	99.55	0.5193
	IonSeq	99.01	31.07	99.82	0.4553
	COACH	98.54	54.44	99.07	0.4656
	S-site	98.95	23.51	99.85	0.3882

	COFACTOR	98.77	44.71	99.42	0.4568
	TM-site	98.01	47.36	98.62	0.3619
Na^{+}	IonCom	92.03	43.27	92.90	0.1777
	IonSeq	74.09	77.14	74.04	0.1516
	COACH	96.91	14.52	98.38	0.1259
	S-site	97.83	2.25	99.54	0.0334
	COFACTOR	98.12	5.93	99.76	0.1289
	TM-site	97.91	7.98	99.52	0.1260
K^{+}	IonCom	94.37	20.93	96.49	0.1460
	IonSeq	97.32	8.52	99.88	0.2283
	COACH	93.95	12.69	96.27	0.0752
	S-site	96.96	0.93	99.70	0.0186
	COFACTOR	96.62	4.66	99.25	0.0696
	TM-site	96.72	3.92	99.37	0.0639
$CO3^{2-}$	IonCom	98.47	12.81	99.67	0.2068
	IonSeq	98.58	10.62	99.82	0.2127
	COACH	98.39	8.86	99.63	0.1420
	S-site	98.55	6.33	99.83	0.1430
	COFACTOR	98.57	6.33	99.85	0.1486
	TM-site	98.24	6.01	99.52	0.0866
$NO2^{-}$	IonCom	98.92	17.00	99.93	0.3534
	IonSeq	98.79	18.00	99.78	0.2847
	COACH	98.86	21.43	99.79	0.3395
	S-site	98.85	6.12	99.96	0.1997
	COFACTOR	98.86	15.31	99.86	0.2932
	TM-site	98.50	4.08	99.63	0.0628
$SO4^{2-}$	IonCom	97.73	15.15	99.49	0.2338
	IonSeq	97.53	13.65	99.32	0.1906
	COACH	97.21	19.15	98.87	0.2114
	S-site	97.62	10.54	99.48	0.1680
	COFACTOR	97.73	11.67	99.56	0.1963
	TM-site	96.98	14.40	98.73	0.1525
$PO4^{3-}$	IonCom	98.00	31.75	99.28	0.3728
	IonSeq	97.95	24.15	99.38	0.3121
	COACH	97.52	35.33	98.72	0.3381
	S-site	97.86	21.45	99.33	0.2764
	COFACTOR	97.98	24.08	99.41	0.3163
	TM-site	97.29	27.86	98.63	0.2667

Table S5. Performance of six methods for the binding-site prediction of 13 metal and radical acid ions on the "hard" target proteins that have no close homology in the PDB.

Ligand[a]	N[b]	Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	MCC
Zn^{2+}	2	IonCom	97.83	20	99.63	21.91
		COACH	97.47	20	99.26	17.61
		S-SITE	97.1	20	98.89	15.02
		IonSeq	87	62.5	87.66	21.27
		COFACTOR	97.83	0	100	0
		TM-SITE	97.83	0	100	0
Cu^{2+}	3	IonCom	98.33	0	100	0
		COACH	98.76	16.67	100	23.42
		S-SITE	98.76	33.33	99.56	27.04
		IonSeq	98.33	0	100	0
		COFACTOR	98.33	0	100	0
		TM-SITE	98.33	0	100	0
Mn^{2+}	11	IonCom	98.46	31.97	99.54	31.71
		COACH	97.56	36.06	98.54	28.95
		S-SITE	97.95	42.05	98.84	33.61
		IonSeq	98.45	0	99.9	-0.08
		COFACTOR	98.48	9.09	99.87	7.33
		TM-SITE	98.28	6.82	99.67	4.34
Na^{+}	3	IonCom	95.48	11.11	96.95	3.32
		COACH	98.35	0	100	0
		S-SITE	98.35	0	100	0
		IonSeq	78.76	55.55	79.26	9.62
		COFACTOR	98.35	0	100	0
		TM-SITE	98.35	0	100	0
$SO4^{2-}$	5	IonCom	97.3	6.67	99.8	9.02
		COACH	97.02	6.67	99.51	7.08
		S-SITE	97.11	0	100	0
		IonSeq	96.83	0	99.71	-0.61
		COFACTOR	97.39	6.67	99.9	10
		TM-SITE	97.02	6.67	99.51	7.08
$PO4^{3-}$	8	IonCom	97.65	2.5	99.56	2.2
		COACH	97.54	5	99.43	2.9
		S-SITE	97.81	0	99.76	-0.36
		IonSeq	97.36	15.48	98.89	10.59
		COFACTOR	97.91	0	99.86	-0.17
		TM-SITE	98.05	0	100	0

[a]The ligands with "hard" target proteins less than 2 are ignored.

[b]The number of "hard" target proteins in each type of ligands.

Table S6. Comparison of general-purpose and ligand-specific model of IonSeq training on a sub-dataset of BioLiP by five-fold cross validation.

Ligand	Model Type	Acc (%)	Sen (%)	Spe (%)	MCC
Zn^{2+}	General	99.16	10.74	99.70	0.1352
	Specific	92.72	45.43	92.78	0.5023
Cu^{2+}	General	98.70	34.65	99.47	0.3829
	Specific	99.12	49.70	99.23	0.5876
Fe^{2+}	General	98.64	26.55	99.77	0.4062
	Specific	98.83	53.54	99.52	0.5732
Fe^{3+}	General	99.09	25.61	99.99	0.4920
	Specific	99.20	51.36	99.81	0.6299
Ca^{2+}	General	98.65	15.73	99.54	0.1999
	Specific	98.17	22.43	99.03	0.2078
Mg^{2+}	General	99.50	21.54	99.84	0.2793
	Specific	99.54	5.42	99.92	0.1835
Mn^{2+}	General	98.85	21.91	99.83	0.3629
	Specific	99.12	31.71	98.32	0.4545
Na^{+}	General	97.51	6.36	99.25	0.0829
	Specific	73.98	77.14	73.93	0.1512
K^{+}	General	97.12	14.58	99.72	0.2908
	Specific	97.13	8.46	99.73	0.2238
$CO3^{2-}$	General	98.96	14.29	99.93	0.3132
	Specific	98.67	9.06	99.93	0.2312
$NO2^{-}$	General	98.99	23.53	99.81	0.3626
	Specific	98.47	17.60	99.35	0.2844
$SO4^{2-}$	General	97.66	2.98	99.67	0.0608
	Specific	97.79	10.07	99.66	0.1879
$PO4^{3-}$	General	98.16	11.45	99.75	0.2225
	Specific	97.57	24.52	99.27	0.3124

Table S7. Comparison of the ion binding residues defined by BioLip and LPC.

Ligand	Only in BioLip[a]	Only in LPC[b]	Common[c]
Zn^{2+}	65	449	632
Cu^{2+}	69	299	466
Fe^{2+}	203	522	912
Fe^{3+}	55	376	384
Ca^{2+}	179	552	729
Mg^{2+}	103	97	701
Mn^{2+}	298	1437	1480
Na^{+}	249	228	240
K^{+}	303	167	233
$CO3^{2-}$	54	243	262
$NO2^{-}$	11	40	87
$SO4^{2-}$	289	2380	1836
$PO4^{3-}$	407	1777	1761

[a]The number of binding residues that are defined by BioLip but not by LPC

[b]The number of binding residues that are defined by LPC but not by BioLiP

[c]The number of binding residues that are defined by both BioLip and LPC