Supporting Information

ATPbind: accurate protein-ATP binding site prediction by combining sequence-profiling and structure-based comparisons

Jun Hu[†,‡], Yang Li[†,‡], Yang Zhang[‡,*] and Dong-Jun Yu[†,*]

[†]School of Computer Science and Engineering, Nanjing University of Science and Technology, Xiaolingwei 200, Nanjing, China, 210094, [‡]Department of Computational Medicine and Bioinformatics, University of Michigan, 100 Washtenaw, Ann Arbor, MI 48109-2218, USA

*Email: zhng@umich.edu or njyudj@njust.edu.cn

Supporting Texts

Text S1: Comparison between S-SITEatp with S-SITE

Table S2 compares the results from S-SITEatp and S-SITE on PATP-TEST. From Table S2, we can find that S-SITEatp has a comparable performance with the original S-SITE although having a slightly lower MCC value than S-SITE.

In order to further compare S-SITEatp and S-SITE, one protein (4xbrA), which consists of 17 ATP-binding residues and 298 non-ATP-binding residues, is used for case study. In this case study, although the true positive numbers of S-SITEatp and S-SITE are both 17, the false positive number of S-SITEatp, which is 34, is less than that of S-SITE, i.e., 76. The running time of S-SITEatp is 105.55s, while the running time of S-SITE is 24,705.89s in the same computational condition. S-SITE searches 2,133 template protein pockets with different ligand types (not only ATP) to locate the binding sites together, so that it cannot tell us the predicted binding sites whether bind ATP or not. In a word, S-SITEatp, which is an ATP-specific S-SITE, is more suitable than a general-purpose predictor, S-SITE, for protein-ATP binding sites prediction.

Text S2: Comparison between TM-SITEatp with TM-SITE

Table S3 compares the discriminative performance between TM-SITEatp and TM-SITE on PATP-TEST. It is easily found that TM-SITEatp has a slightly higher MCC value than TM-SITE.

To further compare TM-SITEatp and TM-SITE, the protein 4xbrA is also employed for case study. In this case study, although the true positive numbers of TM-SITEatp and TM-SITE are both 17, the false positive number of TM-SITEatp, which is 14, is less than that of TM-SITE, i.e., 36. The running time of TM-SITEatp is 253.57s, while the running time of TM-SITE is 13,784.63s in the same computational condition. Since TM-SITE searches 392 template protein pockets with different ligand types (not only ATP) to identify the binding sites together, it does not tell us the predicted binding sites whether bind ATP or not.

Text S3: Cross-validation test

In the K-fold cross-validation test, the training proteins are first randomly partitioned into K disjoint subsets; K-1 subsets are used for training model and remaining one subset is employed for testing; this practice continued until all the K subsets of the training dataset are traversed over; the final performance is obtained by evaluating the combination of the predicted results of all the testing subsets.

Supporting Figures

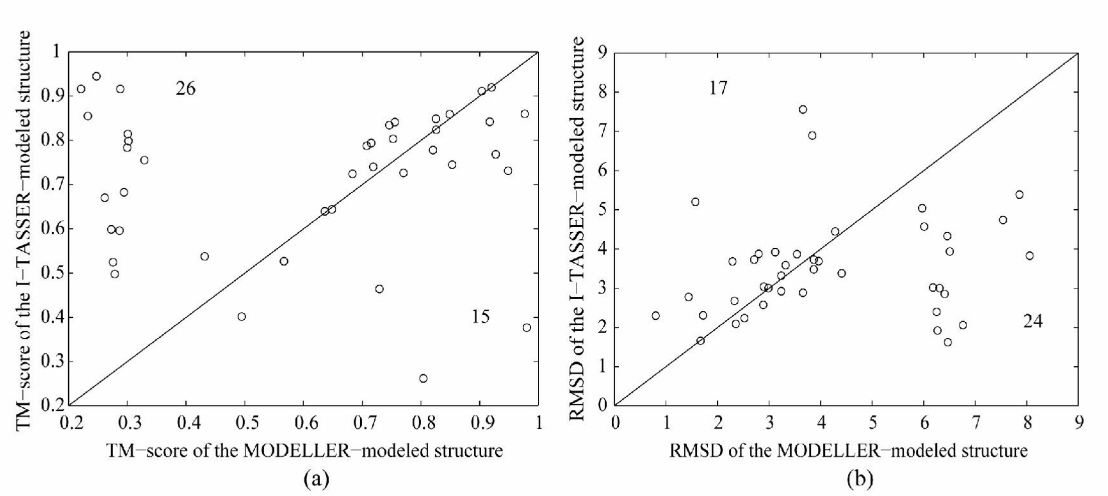

Figure S1. Head-to-head comparisons between I-TASSER and MODELLER on PATP-TEST: (a) TM-score-based comparison; (b) RMSD-based comparison. The numbers in each panel represent the number of points in the upper and lower triangle, respectively.

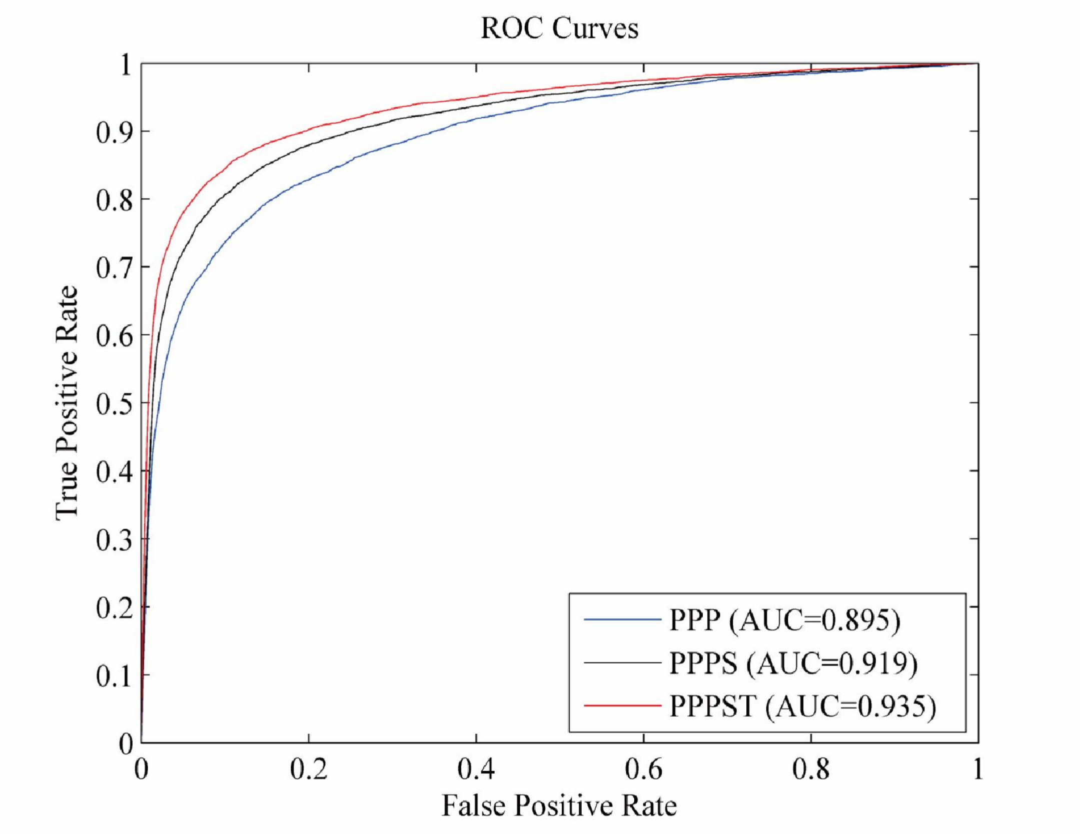

Figure S2. ROC curves of the PPP, PPPS, and PPPST features over five-fold cross-validation tests with the single SVM classifier.

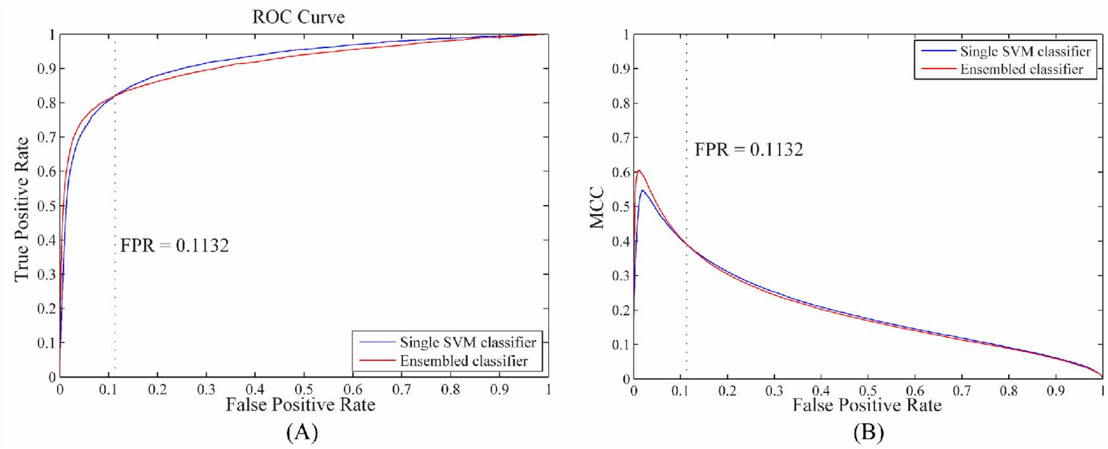

Figure S3. (A) ROC curves of the ensembled classified and single SVM classifier over five-fold cross-validation tests with the PPPS feature. (B) The variation curves of MCC versus false positive rate of the ensembled classified and single SVM classifier over five-fold cross-validation tests with the PPPS features. FPR and TPR represent false positive rate and true positive rate, respectively.

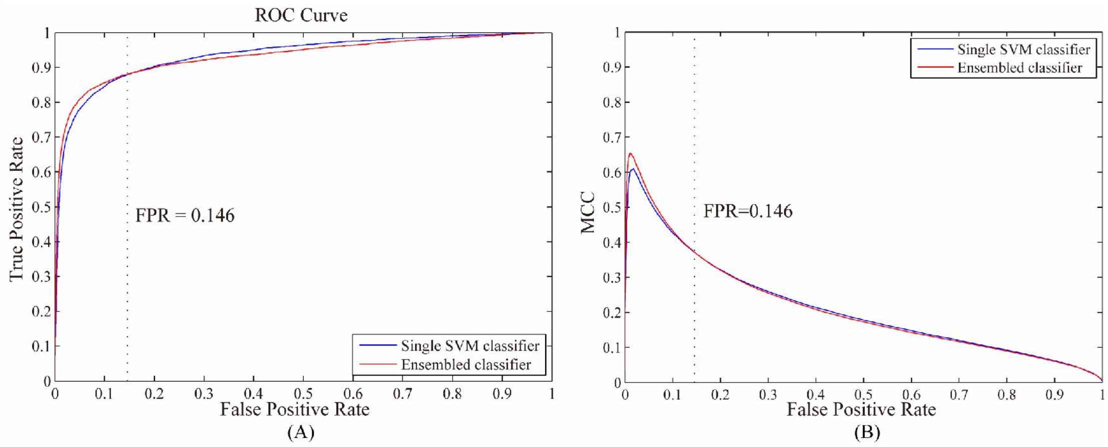

Figure S4. (A) ROC curves of the ensembled classified and single SVM classifier over five-fold cross-validation tests with the PPPST feature. (B) The variation curves of MCC versus false positive rate of the ensembled classified and single SVM classifier over five-fold cross-validation tests with the PPPST features. FPR and TPR represent false positive rate and true positive rate, respectively.

Supporting Tables

Table S1. The maximum sequence identity of each protein in PATP-TEST against all proteins in PATP-388

Protein in PATP-TEST	Protein in PATP-388 [a]	Maximum Sequence Identity (%)	Protein in PATP-TEST	Protein in PATP-388 [a]	Maximum Sequence Identity (%)
5DN6A	3EPSA	17.31	3J94A	1SVMC	19.18
3J8YK	3GNIB	21.52	4WVYA	4O5FA	14.75
4XVUH	4AZWA	16.71	5A98A	4WH3A	13.87
5D15B	3J1FM	8.51	4YB7A	1ZAOA	18.24
4RX6B	4OZNA	33.62	5ECKA	1N48A	14.41
4XBRA	4K6RA	17.06	5DN3A	1II0B	13.92
5EWGA	1OS1A	15.83	4RV7C	3UQDD	13.59
5DB4A	1KO5A	17.88	5EOUB	1B0UA	17.05
4S1KA	4JYZA	12.55	5COUA	2ZT7A	13.40
4YDSA	2CBZA	21.30	5APBA	2HIXA	19.93
4X2DA	3I7VA	13.22	4D79C	4QQXA	8.75
5BSMA	2KMXA	9.81	5D9HB	2Y27B	16.31
5CYRB	4WAEA	17.47	4XHOA	3REUA	17.89
5E84F	1J7KA	13.20	5F1CB	1W7AA	11.13
4ZGNA	1WKLB	11.83	5A99A	3WHLB	10.12
5D6JA	1ESQC	12.06	4XRUA	4BJRA	13.80
4XJXB	2Q66A	15.06	5BURA	5HKKO	15.37
5E3IB	3HNEB	14.09	4XRUB	3TY5A	23.10
5ETLD	1QHXA	17.98	4ZS4B	1EE1A	19.19
4WZYA	2E5YA	10.86	5D1OA	3LL3A	17.43
4RQVA	4FO0A	14.54			

[a] Protein in PATP-388 which has the maximum sequence identity with the corresponding protein in PATP-TEST.

Table S2. Comparison between S-SITEatp and S-SITE on PATP-TEST

Predictor	*Sen* (%)	*Spe* (%)	*Acc* (%)	*Pre* (%)	*MCC*
S-SITEatp	67.51	92.65	91.51	30.41	0.416
S-SITE	64.69	93.50	92.19	32.13	0.420

Table S3. Comparison between TM-SITEatp and TM-SITE on PATP-TEST

Predictor	*Sen* (%)	*Spe* (%)	*Acc* (%)	*Pre* (%)	*MCC*
TM-SITEatp	78.78	96.27	95.48	50.14	0.607
TM-SITE	57.12	98.69	96.80	67.43	0.604

Table S4. Comparison of the quality of the I-TASSER-modeled structure with that of the MODELLER-modeled structure on PATP-TEST

Protein	I-TASSER		MODELLER		Protein	I-TASSER		MODELLER	
	TM-score	RMSD [a]	TM-score	RMSD		TM-score	RMSD	TM-score	RMSD
5DN6A	0.78743	3.48	0.70776	3.87	3J94A	0.53735	6.90	0.43203	3.84
3J8YK	0.68244	4.33	0.29436	6.46	4WVYA	0.46395	3.73	0.72938	2.71
4XVUH	0.72636	3.88	0.77028	2.80	5A98A	0.94468	1.62	0.24736	6.47
5D15B	0.91912	1.66	0.92047	1.67	4YB7A	0.63931	3.59	0.6362	3.32
4RX6B	0.82375	2.09	0.82601	2.36	5ECKA	0.66999	4.74	0.26173	7.54
4XBRA	0.79851	2.86	0.30174	6.41	5DN3A	0.81421	3.00	0.301	6.31
5EWGA	0.83404	3.73	0.74621	3.87	4RV7C	0.52651	2.68	0.5669	2.33
5DB4A	0.84174	2.31	0.91739	1.72	5EOUB	0.84885	3.00	0.82551	2.99
4S1KA	0.85492	2.40	0.2327	6.25	5COUA	0.52448	5.04	0.2758	5.97
4YDSA	0.8034	3.04	0.75254	2.90	5APBA	0.77817	3.32	0.8206	3.24
4X2DA	0.8588	2.58	0.84862	2.89	4D79C	0.74033	2.93	0.71889	3.24
5BSMA	0.76876	3.68	0.92713	2.29	5D9HB	0.78348	3.02	0.29984	6.18
5CYRB	0.72463	3.87	0.68364	3.54	4XHOA	0.75503	3.94	0.32911	6.51
5E84F	0.59583	3.83	0.28664	8.06	5F1CB	0.26159	7.56	0.80403	3.66
4ZGNA	0.49753	4.57	0.27882	6.01	5A99A	0.91625	1.92	0.22132	6.27
5D6JA	0.79357	4.45	0.71556	4.28	4XRUA	0.40156	2.89	0.49453	3.66
4XJXB	0.37701	5.20	0.98009	1.57	5BURA	0.74539	3.92	0.85328	3.12
5E3IB	0.84095	3.38	0.75542	4.41	4XRUB	0.64416	3.69	0.64831	3.96
5ETLD	0.85984	2.30	0.97654	0.80	4ZS4B	0.73152	2.78	0.94825	1.44
4WZYA	0.59873	5.39	0.27251	7.86	5D1OA	0.91156	2.24	0.90375	2.52
4RQVA	0.91564	2.06	0.2881	6.76	Average	0.721	3.502	0.605	4.197

[a] The measurement unit is Å.

Table S5. The prediction performance of ATPbind (EXP [*]) on each protein in PATP-TEST.

Protein name	Sen (%)	Spe (%)	Acc (%)	Pre (%)	MCC	Protein name	Sen (%)	Spe (%)	Acc (%)	Pre (%)	MCC
5DN6A	55.56	99.40	98.61	62.50	0.582	3J94A	34.48	98.26	94.49	55.56	0.411
3J8YK	46.67	99.68	97.27	87.50	0.628	4WVYA	61.90	99.75	97.89	92.86	0.749
4XVUH	59.09	99.31	96.46	86.67	0.699	5A98A	28.57	96.93	94.89	22.22	0.226
5D15B	40.91	100.00	92.07	100.00	0.612	4YB7A	0.00	99.64	94.59	0.00	-0.013
4RX6B	60.00	98.95	92.17	92.31	0.706	5ECKA	26.32	98.55	96.13	38.46	0.299
4XBRA	76.47	97.32	96.19	61.90	0.668	5DN3A	90.00	98.77	98.11	85.71	0.868
5EWGA	92.31	99.51	99.29	85.71	0.886	4RV7C	0.00	99.31	89.44	0.00	-0.026
5DB4A	80.95	98.10	96.09	85.00	0.808	5EOUB	100.00	99.11	99.15	84.21	0.914
4S1KA	0.00	99.12	94.92	0.00	-0.019	5COUA	47.06	98.14	95.58	57.14	0.496
4YDSA	84.62	99.06	98.23	84.62	0.837	5APBA	94.12	98.34	98.20	66.67	0.784
4X2DA	93.33	97.64	97.36	73.68	0.816	4D79C	94.44	98.71	98.41	85.00	0.888
5BSMA	95.24	100.00	99.81	100.00	0.975	5D9HB	76.19	97.79	96.25	72.73	0.724
5CYRB	30.77	98.77	97.44	33.33	0.307	4XHOA	70.00	99.14	97.56	82.35	0.747
5E84F	85.71	99.66	99.17	90.00	0.874	5F1CB	0.00	99.71	96.06	0.00	-0.010
4ZGNA	61.11	98.85	96.42	78.57	0.675	5A99A	0.00	96.75	96.36	0.00	-0.012
5D6JA	94.12	99.35	99.21	80.00	0.864	4XRUA	90.00	100.00	99.65	100.00	0.947
4XJXB	44.44	99.88	98.73	88.89	0.624	5BURA	100.00	99.56	99.58	90.00	0.947
5E3IB	92.31	99.48	99.25	85.71	0.886	4XRUB	72.22	99.47	98.22	86.67	0.782
5ETLD	12.50	94.44	86.25	20.00	0.086	4ZS4B	47.06	99.58	96.11	88.89	0.631
4WZYA	63.64	98.86	98.00	58.33	0.599	5D1OA	70.59	99.72	98.40	92.31	0.800
4RQVA	100.00	99.26	99.31	90.48	0.948						

[*] EXP means that the experimental 3D structure information is used by ATPbind

Table S6. The prediction performance of ATPbind (ITA [*]) on each protein in PATP-TEST.

Protein name	Sen (%)	Spe (%)	Acc (%)	Pre (%)	MCC	Protein name	Sen (%)	Spe (%)	Acc (%)	Pre (%)	MCC
5DN6A	55.56	99.40	98.61	62.50	0.582	3J94A	31.03	98.26	94.29	52.94	0.378
3J8YK	20.00	98.73	95.15	42.86	0.271	4WVYA	52.38	98.52	96.25	64.71	0.563
4XVUH	68.18	99.31	97.11	88.24	0.761	5A98A	42.86	97.37	95.74	33.33	0.356
5D15B	40.91	100.00	92.07	100.00	0.612	4YB7A	13.33	98.93	94.59	40.00	0.209
4RX6B	80.00	98.95	95.65	94.12	0.843	5ECKA	21.05	97.82	95.25	25.00	0.205
4XBRA	76.47	97.32	96.19	61.90	0.668	5DN3A	90.00	98.77	98.11	85.71	0.868
5EWGA	92.31	99.51	99.29	85.71	0.886	4RV7C	6.25	95.17	86.34	12.50	0.020
5DB4A	90.48	98.73	97.77	90.48	0.892	5EOUB	100.00	99.11	99.15	84.21	0.914
4S1KA	20.00	99.12	95.76	50.00	0.298	5COUA	70.59	98.45	97.05	70.59	0.690
4YDSA	84.62	99.06	98.23	84.62	0.837	5APBA	94.12	98.76	98.60	72.73	0.821
4X2DA	93.33	98.11	97.80	77.78	0.841	4D79C	88.89	98.71	98.01	84.21	0.854
5BSMA	90.48	99.80	99.43	95.00	0.924	5D9HB	71.43	97.79	95.90	71.43	0.692
5CYRB	30.77	99.08	97.74	40.00	0.340	4XHOA	35.00	99.14	95.66	70.00	0.476
5E84F	85.71	99.66	99.17	90.00	0.874	5F1CB	23.08	96.49	93.80	20.00	0.183
4ZGNA	44.44	98.47	94.98	66.67	0.520	5A99A	0.00	98.37	97.98	0.00	-0.008
5D6JA	94.12	99.02	98.89	72.73	0.822	4XRUA	90.00	100.00	99.65	100.00	0.947
4XJXB	44.44	100.00	98.84	100.00	0.663	5BURA	100.00	99.56	99.58	90.00	0.947
5E3IB	84.62	99.48	98.99	84.62	0.841	4XRUB	61.11	99.73	97.97	91.67	0.739
5ETLD	6.25	93.75	85.00	10.00	0.000	4ZS4B	47.06	99.58	96.11	88.89	0.631
4WZYA	72.73	98.18	97.56	50.00	0.591	5D1OA	58.82	99.44	97.60	83.33	0.689
4RQVA	100.00	99.26	99.31	90.48	0.948						

[*] ITA means that the I-TASSER-modeled 3D structure information is used by ATPbind

Table S7. The prediction performance of ATPseq on each protein in PATP-TEST.

Protein name	Sen (%)	Spe (%)	Acc (%)	Pre (%)	MCC	Protein name	Sen (%)	Spe (%)	Acc (%)	Pre (%)	MCC
5DN6A	55.56	98.59	97.82	41.67	0.470	3J94A	31.03	98.26	94.29	52.94	0.378
3J8YK	46.67	99.68	97.27	87.50	0.628	4WVYA	47.62	99.75	97.19	90.91	0.647
4XVUH	63.64	99.31	96.78	87.50	0.731	5A98A	0.00	99.12	96.17	0.00	-0.016
5D15B	54.55	100.00	93.90	100.00	0.714	4YB7A	13.33	98.93	94.59	40.00	0.209
4RX6B	80.00	96.84	93.91	84.21	0.784	5ECKA	0.00	99.82	96.49	0.00	-0.008
4XBRA	76.47	97.32	96.19	61.90	0.668	5DN3A	90.00	98.77	98.11	85.71	0.868
5EWGA	92.31	99.51	99.29	85.71	0.886	4RV7C	6.25	97.93	88.82	25.00	0.080
5DB4A	80.95	97.47	95.53	80.95	0.784	5EOUB	100.00	99.40	99.43	88.89	0.940
4S1KA	0.00	98.67	94.49	0.00	-0.024	5COUA	5.88	99.38	94.69	33.33	0.123
4YDSA	92.31	100.00	99.56	100.00	0.959	5APBA	94.12	98.34	98.20	66.67	0.784
4X2DA	86.67	98.11	97.36	76.47	0.800	4D79C	88.89	98.28	97.61	80.00	0.831
5BSMA	95.24	100.00	99.81	100.00	0.975	5D9HB	71.43	98.16	96.25	75.00	0.712
5CYRB	0.00	99.85	97.89	0.00	-0.005	4XHOA	25.00	99.43	95.39	71.43	0.405
5E84F	90.48	99.83	99.50	95.00	0.925	5F1CB	0.00	99.71	96.06	0.00	-0.010
4ZGNA	44.44	99.23	95.70	80.00	0.577	5A99A	0.00	99.19	98.79	0.00	-0.006
5D6JA	94.12	99.35	99.21	80.00	0.864	4XRUA	80.00	100.00	99.29	100.00	0.891
4XJXB	44.44	100.00	98.84	100.00	0.663	5BURA	100.00	99.56	99.58	90.00	0.947
5E3IB	69.23	99.48	98.49	81.82	0.745	4XRUB	16.67	99.47	95.69	60.00	0.301
5ETLD	6.25	98.61	89.38	33.33	0.108	4ZS4B	11.76	98.75	93.00	40.00	0.189
4WZYA	0.00	99.77	97.34	0.00	-0.007	5D1OA	35.29	100.00	97.07	100.00	0.585
4RQVA	100.00	99.26	99.31	90.48	0.948						