Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment

Jianyi Yang[1], Ambrish Roy[1] and Yang Zhang[1,2*]

[1]Department of Computational Medicine and Bioinformatics, [2]Department of Biological Chemistry, University of Michigan, 100 Washtenaw Avenue, Ann Arbor, MI 48109-2218, USA

*Email: zhng@umich.edu

Supplementary Materials

Normalized BLOSUM62 scoring matrix

To have the binding and conservation terms comparable to other alignment terms in Eqs. (1) and (3), we normalize the BLOSUM62 matrix with the element value in [0, 1], i.e.

$$B(i, j) = \frac{b(i, j) - b_{\min}}{b_{\max} - b_{\min}}, \qquad i, j \in \{1,2,...20\} \tag{S1}$$

where $b(i, j)$ is the original BLOSUM62 martix. $b_{\min} = -4$ and $b_{\max} = 11$ are the minimum and maximum values of the elements of the BLOSUM62 matrix (Henikoff and Henikoff, 1992), respectively.

Jensen–Shannon divergence score

The Jensen-Shannon divergence (JSD) (Capra and Singh, 2007) score is an index to measure the extent of the evolutionary conservation/variation of each residue position in a protein chain. To compute JSD, we first construct the multiple sequence alignment (MSA) of the query protein by running PSI-BLAST (Altschul, et al., 1997) against the NCBI non-redundant sequence database (with parameters "-j 3 -h 0.001"). After obtaining the MSA of the query protein, Henikoff and Henikoff scheme (Henikoff and Henikoff, 1994) is used to weight the aligned sequences in MSA for subsequent computation of the weighted occurring frequency $\boldsymbol{p_i}$ of amino acids in the ith column of the MSA (corresponds to one position/residue in the query protein). Denote the background frequency by $\boldsymbol{q}$, which can be estimated using a large set of random sequences. A new frequency vector $\boldsymbol{c_i}$ is computed by combining $\boldsymbol{p_i}$ and $\boldsymbol{q}$ by $\boldsymbol{c_i} = \lambda\, \boldsymbol{p_i} + (1- \lambda)\, \boldsymbol{q}$, where λ is a parameter and it is set to be 0.5 as in (Capra and Singh, 2007). The JSD score for the ith column is calculated by:

$$JSD_i = \lambda \sum_{\alpha \in AA} p_i(\alpha) \log \frac{p_i(\alpha)}{c_i(\alpha)} + (1-\lambda) \sum_{\alpha \in AA} q(\alpha) \log \frac{q(\alpha)}{c_i(\alpha)} \tag{S2}$$

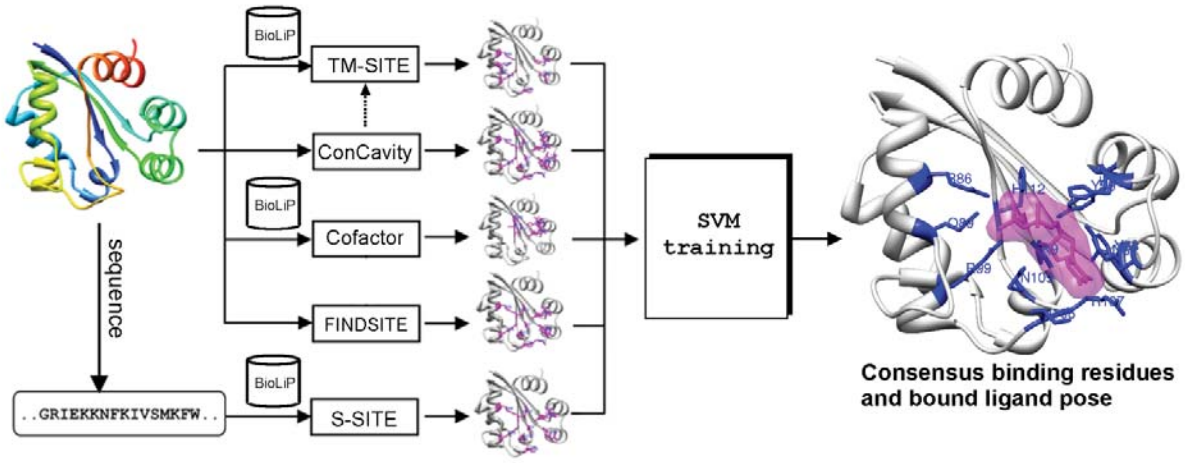


Figure S1. Flowchart of the COACH algorithm. COACH combines complementary predictions from TM-SITE, S-SITE, COFACTOR, FINDSITE and ConCavity.

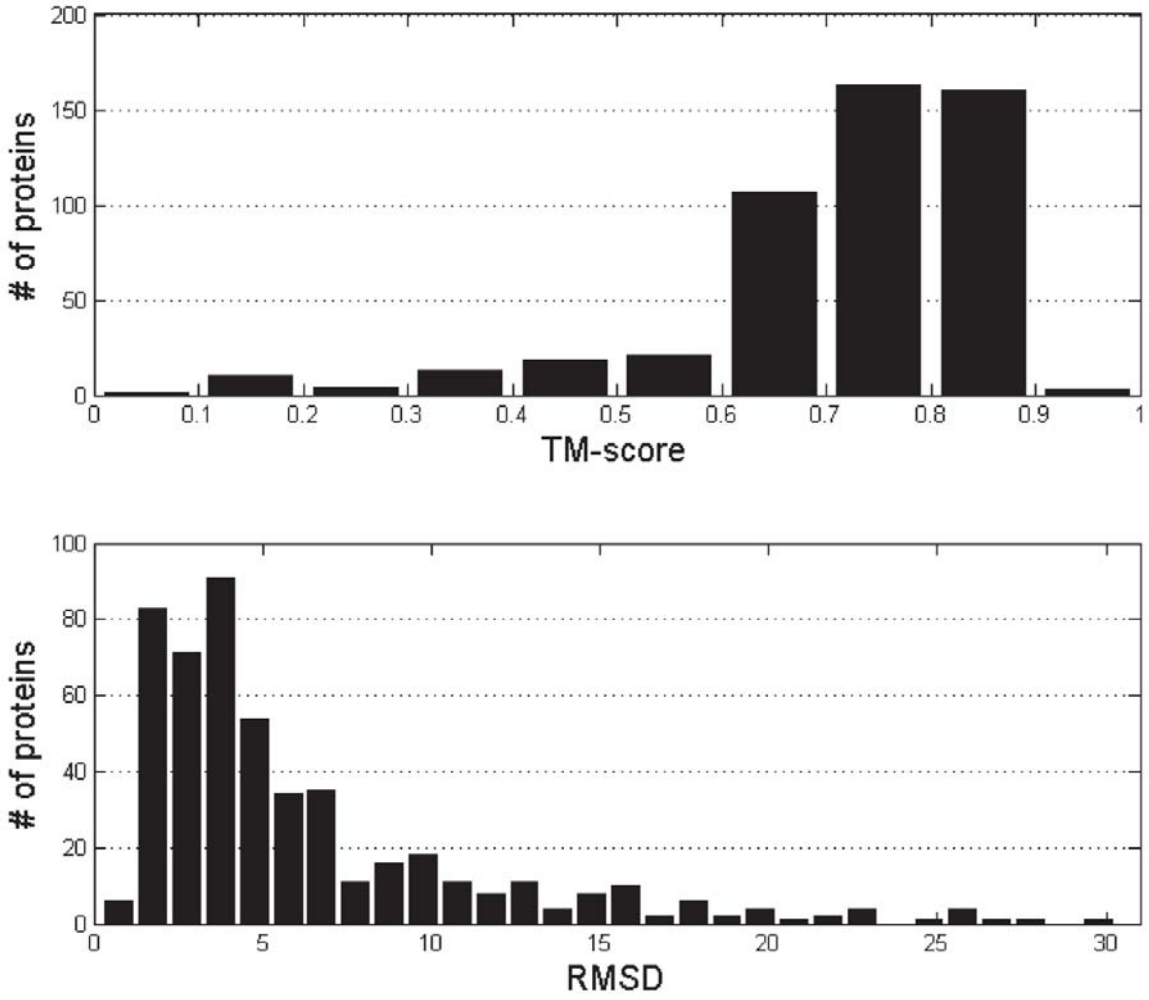


Figure S2. TM-score and RMSD distributions of the I-TASSER models for the 500 test proteins.

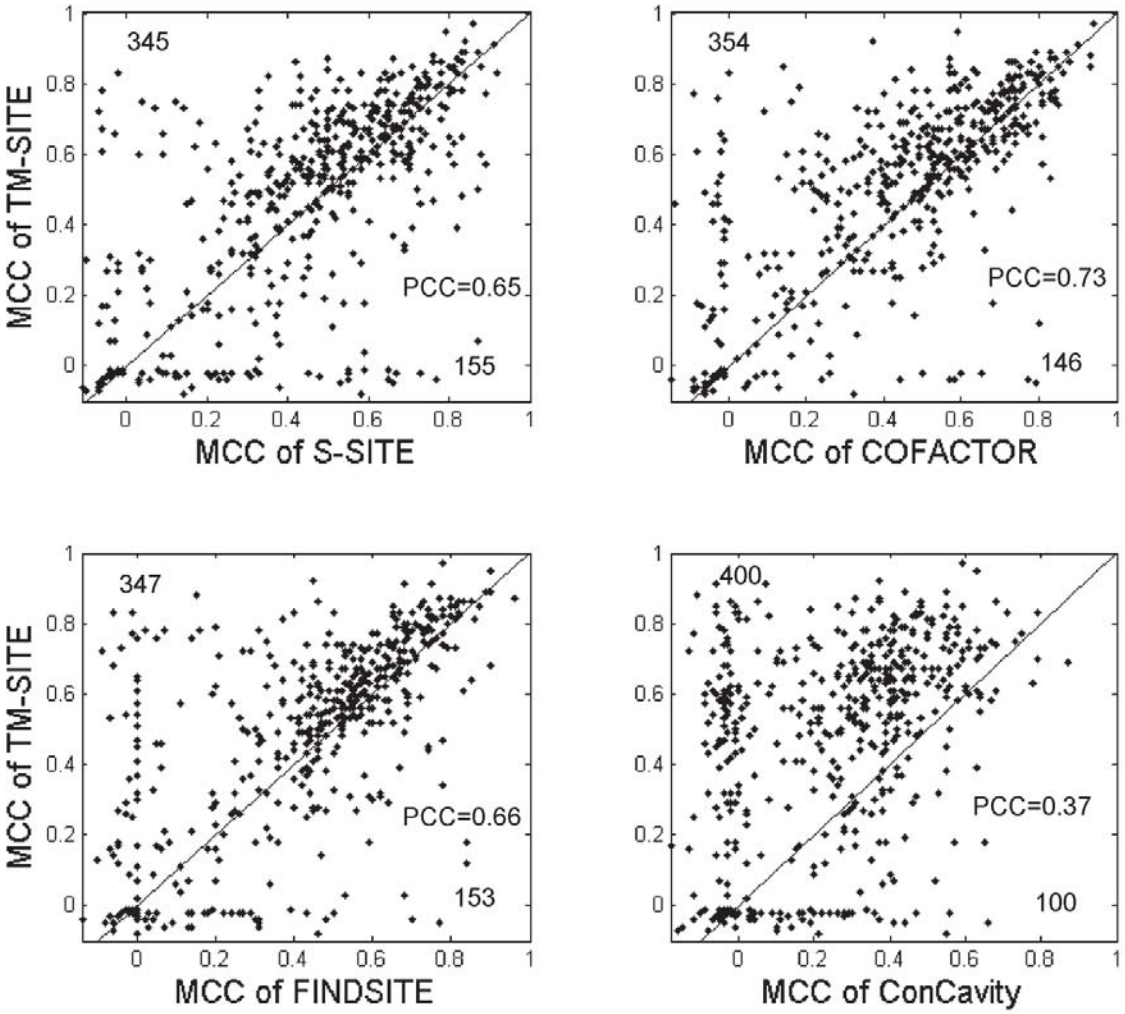


Figure S3. Head-to-head comparisons of MCC scores between TM-SITE and other LBS prediction methods on the 500 test proteins. The numbers in each panel represent the number of points in the upper and lower triangles, respectively. PCC is the Pearson's correlation coefficient between the MCCs of the two compared methods.

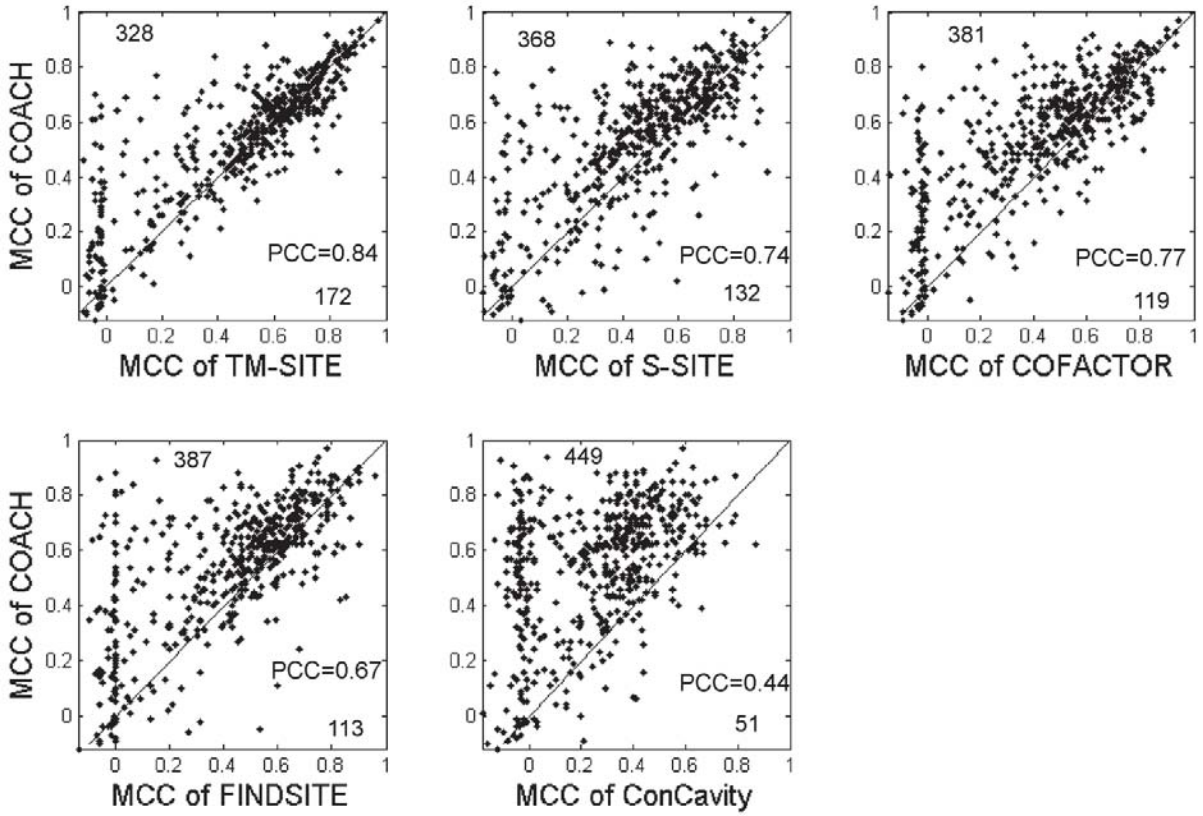
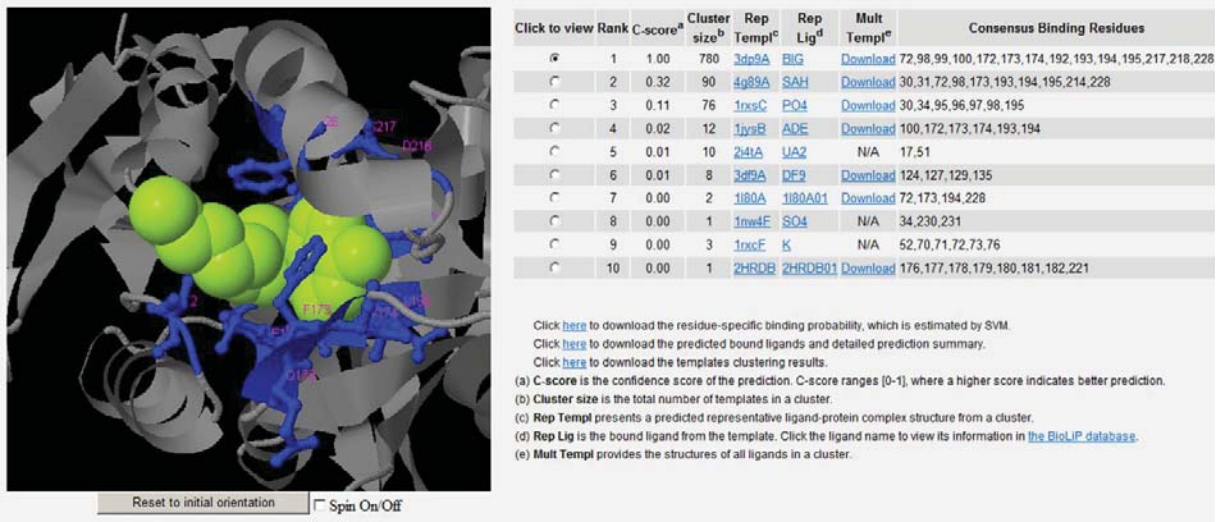


Figure S4. Head-to-head comparisons between COACH and individual component methods on the 500 test proteins. The numbers in each panel represent the number of points in the upper and lower triangle, respectively. PCC is the Pearson's correlation coefficient between the MCCs of the two compared methods.



Click to view	Rank	C-score[a]	Cluster size[b]	Rep Templ[c]	Rep Lig[d]	Mult Templ[e]	Consensus Binding Residues
◉	1	1.00	780	3dp9A	BIG	Download	72,98,99,100,172,173,174,192,193,194,195,217,218,228
○	2	0.32	90	4g89A	SAH	Download	30,31,72,98,173,193,194,195,214,228
○	3	0.11	76	1rxsC	PO4	Download	30,34,95,96,97,98,195
○	4	0.02	12	1jysB	ADE	Download	100,172,173,174,193,194
○	5	0.01	10	2i4tA	UA2	N/A	17,51
○	6	0.01	8	3df9A	DF9	Download	124,127,129,135
○	7	0.00	2	1l80A	1l80A01	Download	72,173,194,228
○	8	0.00	1	1nw4F	SO4	N/A	34,230,231
○	9	0.00	3	1rxcF	K	N/A	52,70,71,72,73,76
○	10	0.00	1	2HRDB	2HRDB01	Download	176,177,178,179,180,181,182,221

Click here to download the residue-specific binding probability, which is estimated by SVM.
Click here to download the predicted bound ligands and detailed prediction summary.
Click here to download the templates clustering results.

(a) **C-score** is the confidence score of the prediction. C-score ranges [0-1], where a higher score indicates better prediction.
(b) **Cluster size** is the total number of templates in a cluster.
(c) **Rep Templ** presents a predicted representative ligand-protein complex structure from a cluster.
(d) **Rep Lig** is the bound ligand from the template. Click the ligand name to view its information in the BioLiP database.
(e) **Mult Templ** provides the structures of all ligands in a cluster.

Figure S5. An illustration of the LBS prediction results on the COACH server. The top 10 predictions by COACH are presented in a table. Individual predictions can also visualized by the Jmol Applet on the left panel. The picutre is a snapshot taken from the COACH example page at http://zhanglab.ccmb.med.umich.edu/COACH/CH000001/.

Table S1. The p-values in student t-test for the difference in MCC score between predictors on the 500 test proteins where receptor models are generated by I-TASSER.

Predictors	COACH	TM-SITE	S-SITE	COFACTOR	FINDSITE	ConCavity
COACH		1.6e-14	6.2e-25	1.9e-40	1.8e-31	4.6e-92
TM-SITE			1.1e-03	2.6e-11	1.7e-09	1.2e-53
S-SITE				3.6e-03	0.01	6.7e-41

Table S2. A detail list of ligand-binding site predictions in CAMEO (data from 2012/12/07 to 2013/05/03, taken from http://www.cameo3d.org/ligand_binding/weekly_summary.html).

Date	Server Name	# Targets released	# Targets modelled	Average Accuracy (all targets)	Average Accuracy (modeled targets)
2013.05.03	Random	71	0	0.5	nan
	naive_homology	71	15	0.56	0.79
	naive_pocket	71	68	0.67	0.68
	naive_conservation	71	71	0.67	0.67
	INTFOLD-FN	71	39	0.59	0.67
	server5	71	61	0.65	0.68
	HHfunc	71	33	0.59	0.7
	COACH	71	71	0.86	0.86
2013.04.26	Random	89	89	0.51	0.51
	naive_homology	89	35	0.61	0.77
	naive_pocket	89	89	0.72	0.72
	naive_conservation	89	89	0.68	0.68
	INTFOLD-FN	89	64	0.6	0.65
	server5	89	67	0.62	0.66
	HHfunc	89	27	0.58	0.75
	COACH	89	89	0.88	0.88
2013.04.19	Random	64	64	0.5	0.5
	naive_homology	64	33	0.64	0.77
	naive_pocket	64	63	0.76	0.76
	naive_conservation	64	64	0.68	0.68
	INTFOLD-FN	64	40	0.63	0.7
	server5	64	61	0.66	0.66
	HHfunc	64	22	0.63	0.87
	COACH	64	64	0.92	0.92
2013.04.12	Random	55	55	0.53	0.53
	naive_homology	55	18	0.61	0.84
	naive_pocket	55	52	0.67	0.68
	naive_conservation	55	54	0.67	0.67
	INTFOLD-FN	55	24	0.57	0.65

	server5	55	39	0.64	0.7
	HHfunc	55	37	0.64	0.71
	COACH	55	55	0.88	0.88
2013.04.05	Random	38	38	0.52	0.52
	naive_homology	38	10	0.59	0.85
	naive_pocket	38	37	0.75	0.76
	naive_conservation	38	37	0.71	0.72
	INTFOLD-FN	38	27	0.63	0.68
	server5	38	25	0.63	0.71
	HHfunc	38	25	0.73	0.85
	COACH	38	38	0.88	0.88
2013.03.29	Random	54	54	0.47	0.47
	naive_homology	54	8	0.54	0.8
	naive_pocket	54	51	0.68	0.69
	naive_conservation	54	53	0.66	0.66
	INTFOLD-FN	54	28	0.59	0.67
	server5	54	34	0.6	0.67
	HHfunc	54	35	0.62	0.68
	COACH	54	54	0.86	0.86
2013.03.22	Random	68	68	0.52	0.52
	naive_homology	68	28	0.64	0.84
	naive_pocket	68	68	0.7	0.7
	naive_conservation	68	68	0.64	0.64
	INTFOLD-FN	68	46	0.65	0.72
	server5	68	59	0.7	0.72
	HHfunc	68	41	0.7	0.82
	COACH	68	68	0.9	0.9
2013.03.15	Random	70	70	0.5	0.5
	naive_homology	70	44	0.71	0.83
	naive_pocket	70	67	0.81	0.83
	naive_conservation	70	69	0.72	0.72
	INTFOLD-FN	70	48	0.66	0.74
	server5	70	63	0.7	0.72
	HHfunc	70	29	0.64	0.83
	COACH	70	70	0.93	0.93
2013.03.08	Random	26	26	0.49	0.49
	naive_homology	26	4	0.55	0.81
	naive_pocket	26	26	0.65	0.65
	naive_conservation	26	26	0.67	0.67
	INTFOLD-FN	26	6	0.56	0.78
	server5	26	20	0.64	0.68

	HHfunc	26	0	0.5	nan
	COACH	26	26	0.82	0.82
2013.03.01	Random	90	90	0.53	0.53
	naive_homology	90	17	0.58	0.91
	naive_pocket	90	87	0.58	0.59
	naive_conservation	90	90	0.61	0.61
	INTFOLD-FN	90	39	0.55	0.61
	server5	90	46	0.6	0.69
	HHfunc	90	0	0.5	nan
	COACH	90	90	0.77	0.77
2013.02.22	Random	26	26	0.54	0.54
	naive_homology	26	7	0.57	0.76
	naive_pocket	26	24	0.63	0.65
	naive_conservation	26	25	0.66	0.66
	INTFOLD-FN	26	16	0.57	0.62
	server5	26	17	0.57	0.61
	HHfunc	26	15	0.61	0.7
	COACH	26	26	0.79	0.79
2013.02.15	Random	48	48	0.51	0.51
	naive_homology	48	27	0.7	0.86
	naive_pocket	48	48	0.79	0.79
	naive_conservation	48	48	0.65	0.65
	INTFOLD-FN	48	42	0.69	0.71
	server5	48	40	0.66	0.69
	HHfunc	48	22	0.69	0.9
	COACH	48	48	0.93	0.93
2013.02.08	Random	65	65	0.5	0.5
	naive_homology	65	29	0.69	0.93
	naive_pocket	65	64	0.61	0.61
	naive_conservation	65	0	0.5	nan
	INTFOLD-FN	65	50	0.69	0.74
	server5	65	58	0.73	0.75
	HHfunc	65	24	0.58	0.72
	COACH	65	65	0.89	0.89
2013.02.01	Random	78	78	0.5	0.5
	naive_homology	78	29	0.58	0.72
	naive_pocket	78	70	0.66	0.68
	naive_conservation	78	0	0.5	nan
	INTFOLD-FN	78	57	0.61	0.65
	server5	78	48	0.62	0.7
	HHfunc	78	44	0.63	0.73

	COACH	78	27	0.63	0.87
2013.01.25	Random	42	42	0.52	0.52
	naive_homology	42	17	0.63	0.82
	naive_pocket	42	38	0.73	0.76
	naive_conservation	42	0	0.5	nan
	INTFOLD-FN	42	31	0.63	0.68
	server5	42	33	0.61	0.65
	HHfunc	42	33	0.76	0.83
	COACH	42	42	0.88	0.88
2013.01.18	Random	61	61	0.53	0.53
	naive_homology	61	19	0.59	0.78
	naive_pocket	61	59	0.76	0.77
	naive_conservation	61	0	0.5	nan
	INTFOLD-FN	61	48	0.64	0.68
	server5	61	50	0.68	0.72
	HHfunc	61	37	0.72	0.86
	COACH	61	61	0.87	0.87
2013.01.11	Random	56	56	0.51	0.51
	naive_homology	56	22	0.58	0.7
	naive_pocket	56	56	0.74	0.74
	naive_conservation	56	56	0.68	0.68
	INTFOLD-FN	56	41	0.66	0.72
	server5	56	37	0.65	0.73
	HHfunc	56	56	0.7	0.7
	COACH	56	56	0.87	0.87
2013.01.04	Random	16	16	0.53	0.53
	naive_homology	16	4	0.57	0.79
	naive_pocket	16	16	0.85	0.85
	naive_conservation	16	15	0.67	0.68
	INTFOLD-FN	16	15	0.75	0.77
	server5	16	14	0.74	0.77
	HHfunc	16	16	0.72	0.72
	COACH	16	16	0.91	0.91
2012.12.28	Random	18	18	0.5	0.5
	naive_homology	18	2	0.52	0.66
	naive_pocket	18	18	0.76	0.76
	naive_conservation	18	18	0.52	0.52
	INTFOLD-FN	18	13	0.57	0.6
	server5	18	11	0.57	0.61
	HHfunc	18	18	0.62	0.62
	COACH	18	18	0.82	0.82

2012.12.21	Random	56	56	0.51	0.51
	naive_homology	56	23	0.64	0.84
	naive_pocket	56	55	0.75	0.75
	naive_conservation	56	55	0.68	0.69
	INTFOLD-FN	56	41	0.68	0.74
	server5	56	46	0.66	0.7
	HHfunc	56	56	0.74	0.74
	COACH	56	56	0.86	0.86
2012.12.14	Random	63	63	0.5	0.5
	naive_homology	63	17	0.6	0.88
	naive_pocket	63	62	0.69	0.7
	naive_conservation	63	62	0.57	0.57
	INTFOLD-FN	63	53	0.62	0.65
	server5	63	55	0.63	0.64
	HHfunc	63	0	0.5	nan
	COACH	63	63	0.84	0.84
2012.12.07	Random	49	49	0.49	0.49
	naive_homology	49	23	0.67	0.86
	naive_pocket	49	45	0.67	0.69
	naive_conservation	49	46	0.66	0.67
	INTFOLD-FN	49	35	0.65	0.7
	server5	49	39	0.65	0.69
	HHfunc	49	49	0.74	0.74
	COACH	49	49	0.88	0.88

REFERENCES

Altschul, S.F., et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res*, **25**, 3389-3402.

Capra, J.A. and Singh, M. (2007) Predicting functionally important residues from sequence conservation, *Bioinformatics*, **23**, 1875-1882.

Henikoff, S. and Henikoff, J.G. (1992) Amino acid substitution matrices from protein blocks, *Proc Natl Acad Sci U S A*, **89**, 10915-10919.

Henikoff, S. and Henikoff, J.G. (1994) Position-based sequence weights, *J Mol Biol*, **243**, 574-578.