SUPPLEMENTARY MATERIALS

Bioinformatic analysis of protein families for identification of variable amino acid residues responsible for functional diversity

Dmitry Suplatov[1], Daria Shalaeva[2], Evgeny Kirilin[2], Vladimir Arzhanik[2], Vytas Švedas[1,2]

Lomonosov Moscow State University [1] Belozersky Institute of Physicochemical Biology and [2] Faculty of Bioengineering and Bioinformatics, Vorobjev hills, 1-73, 119991 Moscow, Russia

Table S1. Subfamily-specific positions in β/α-barrel fold basic amino acid decarboxylases

Subfamilies are shown for ornithine decarboxylases (ODC), arginine decarboxylases (ADC), diaminopimelate decarboxylases (DAPDC) and carboxynorspermidine decarboxylases (CANSDCs). Hits are ranked in declined statistical significance. Positions are numbered as in ornithine decarboxylase from *Trypanosoma brucei* (TbODC, 1F3T). Most frequently occurring residues are shown for every subfamily. * - catalytic site residues.

Rank	S_i	Z-score	P-value	Position	Subfamily 1: DAPDC (1884 sequences)	Subfamily 2: ADC (682 sequences)	Subfamily 3: CANSDC (421 sequences)	Subfamily 4: ODC (297 sequences)	Subfamily 5: ODC (217 sequences)
1	0.59	5.09	1.9e-05	325S	L (61.8%) N (29.4%)	S (97.5%)	H (93.6%)	S (56.2%) N (24.9%) T (8.1%) A (3.7%)	G (100.0%)
2	0.49	4.45	9.8e-08	323Y*	N (51.3%) S (19.9%) T (12.9%) A (5.0%)	F (99.4%)	E (40.9%) T (28.0%) A (17.1%) S (6.2%)	Y (93.6%)	F (97.2%)
4	0.71	4.23	7.9e-14	393G*	M (92.0%)	L (85.3%) M (9.5%)	K (99.8%)	A (40.7%) S (20.2%) I (6.1%) G (5.7%) L (5.7%) T (5.4%)	Y (99.1%)
5	0.51	4.04	1.0e-15	363L*	G (75.8%) S (11.1%) N (6.7%)	D (99.4%)	G (100.0%)	L (47.1%) I (24.2%) V (5.1%) M (5.1%)	A (71.9%) V (14.3%) M (6.5%)
7	0.50	3.83	8.6e-20	329I*	A (71.7%) S (13.2%) T (6.1%)	W (99.4%)	L (80.3%) I (11.6%)	I (54.9%) V (16.5%) L (8.1%) K (7.1%)	T (74.2%) M (13.8%)
10	0.60	3.64	4.2e-26	361D*	E (99.7%)	D (100.0%)	L (99.5%)	D (96.0%)	D (100.0%)
15	0.61	3.34	6.9e-34	327N*	R (99.9%)	D (99.3%)	D (93.1%)	N (58.2%) S (21.9%) A (5.1%) M (3.7%)	A (64.5%) E (19.4%) I (9.7%)
16	0.40	3.14	1.1e-31	328C*	P (90.3%)	A (44.0%) S (15.5%) F (11.7%) T (8.8%)	V (40.6%) T (21.4%) L (18.8%) C (16.4%)	C (49.5%) N (17.5%) S (16.2%) G (8.1%)	E (80.2%) T (18.9%)
18	0.44	3.02	2.0e-33	333H*	A (76.0%) S (23.4%)	Q (98.4%)	Y (59.9%) E (27.3%) F (5.7%)	H (65.7%) N (9.1%) K (5.4%) D (4.4%)	A (50.2%) I (22.1%) S (16.6%) F (3.2%)
19	0.28	3.01	3.1e-35	392V*	S (58.0%) V (17.6%) T (10.6%) A (9.8%)	I (67.6%) V (11.9%) A (8.8%) T (5.0%)	V (97.1%)	A (37.0%) C (21.9%) V (10.4%) S (7.7%) P (7.7%) G (4.4%)	T (62.2%) S (28.1%)
23	0.48	2.66	1.2e-33	390T*	G (77.2%) C (9.2%) N (5.3%)	Q (90.5%)	T (76.5%) S (23.3%)	T (83.2%) S (9.4%)	T (92.6%)
26	0.27	2.60	1.2e-37	358P*	P (52.5%) K (28.6%) S (7.1%) G (2.5%)	I (68.8%) L (27.0%)	N (28.0%) P (20.2%) K (19.5%) C (15.0%) A (5.9%) L (3.6%)	P (95.3%)	P (96.8%)
29	0.47	2.47	1.2e-38	331Y*	Y (99.7%)	I (89.0%) L (7.9%)	M (57.0%) Y (38.2%)	F (51.9%) Y (30.0%) S (4.4%) M (4.4%)	D (54.4%) E (21.2%) G (12.4%) N (4.6%)
30	0.23	2.43	3.9e-39	332D*	G (32.6%) D (22.0%) Q (21.6%) E (10.7%)	D (69.9%) G (13.8%) E (5.3%) N (4.4%)	P (60.1%) R (19.0%) K (7.6%) Q (6.2%)	D (79.1%) E (11.1%)	E (70.0%) A (10.6%) S (7.8%) G (4.6%)
31	0.35	2.40	8.8e-40	322V*	M (78.0%) F (15.2%)	L (59.5%) V (28.6%) I (7.3%)	A (39.9%) I (16.9%) F (15.0%) T (14.7%)	V (66.3%) L (17.8%) I (10.8%)	K (72.4%) R (12.0%) V (10.1%)
32	0.26	2.40	1.9e-41	364D*	D (99.7%)	G (100.0%)	D (100.0%)	D (100.0%)	D (100.0%)

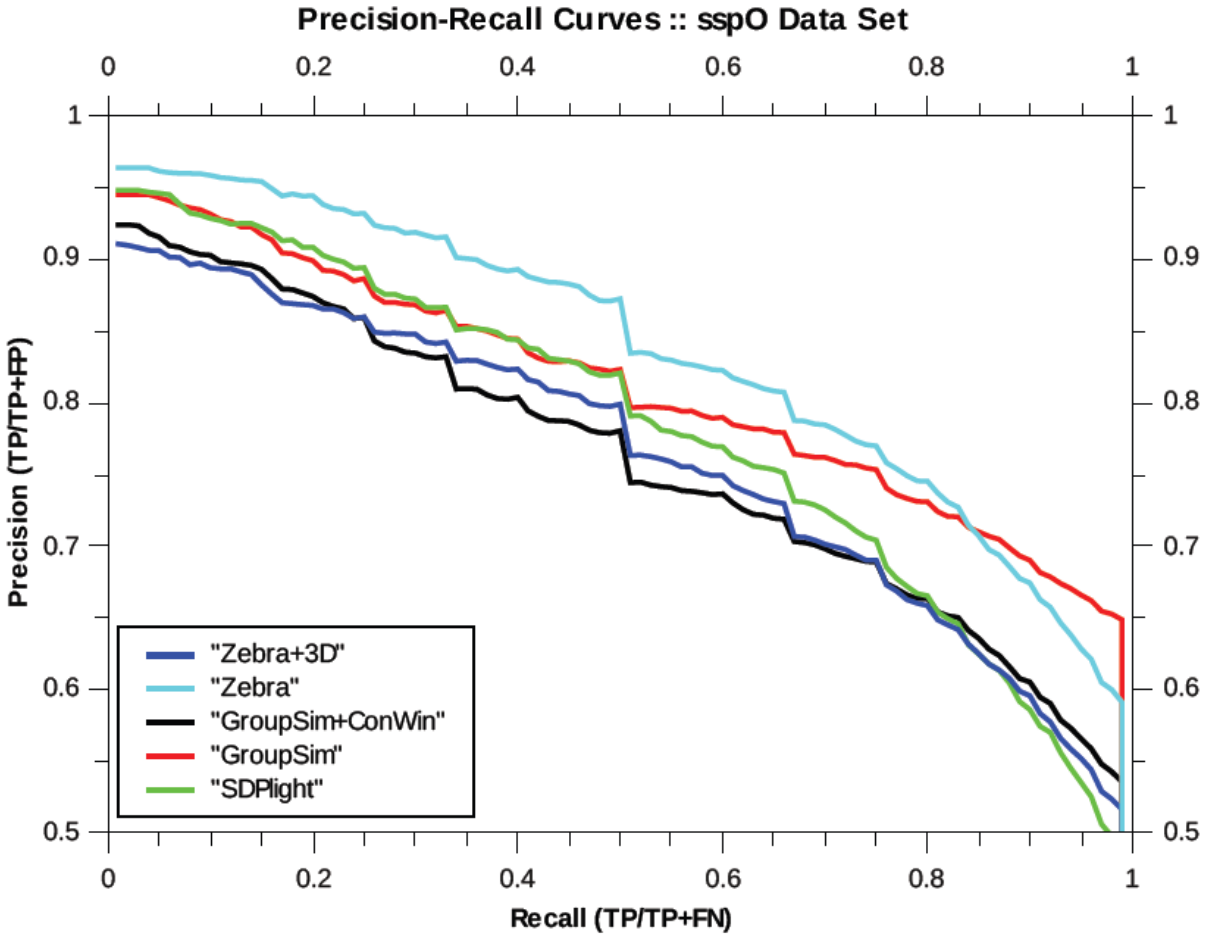

Fig S1. Precision-Recall curves for representative SSP prediction methods on sspO dataset.

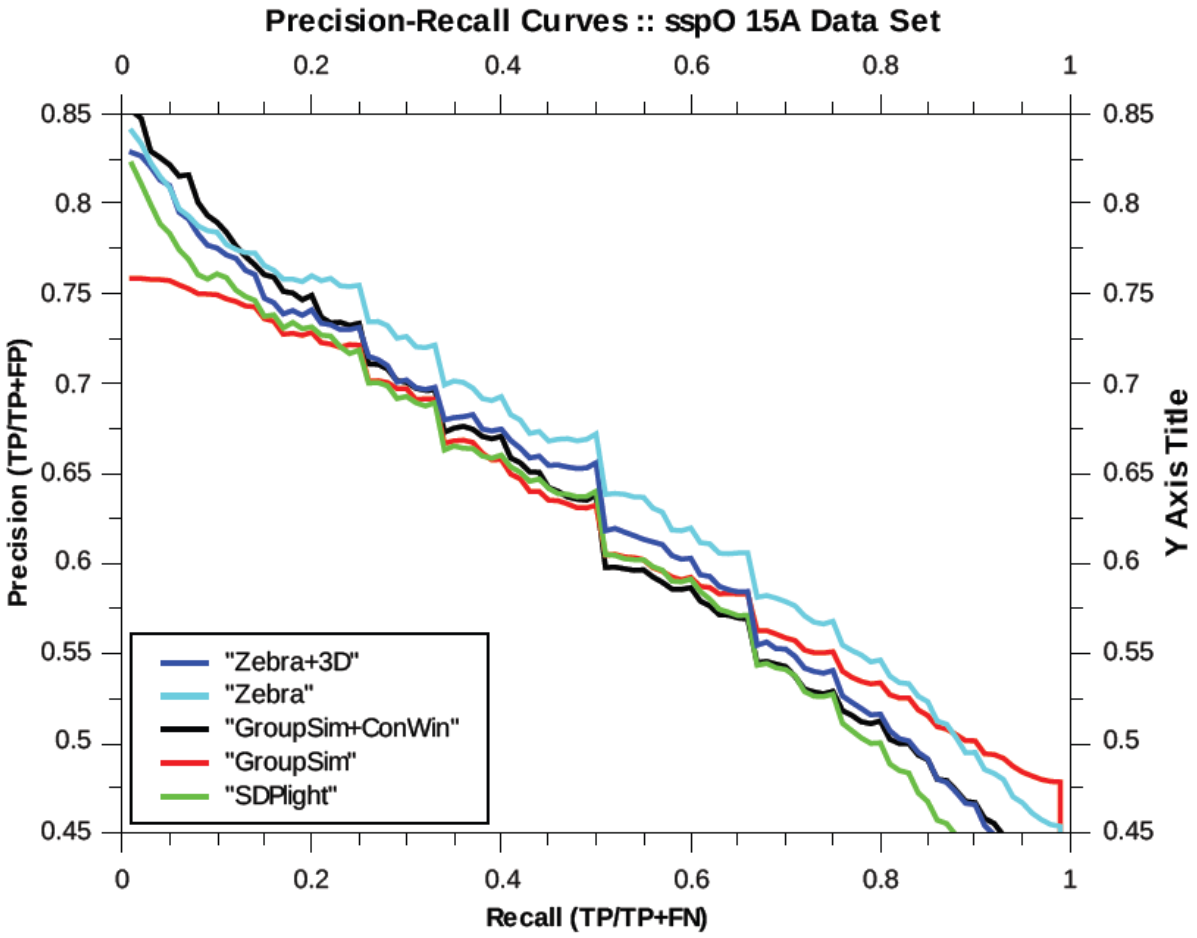

Fig S2. Precision-Recall curves for representative SSP prediction methods on sspO_15 dataset.

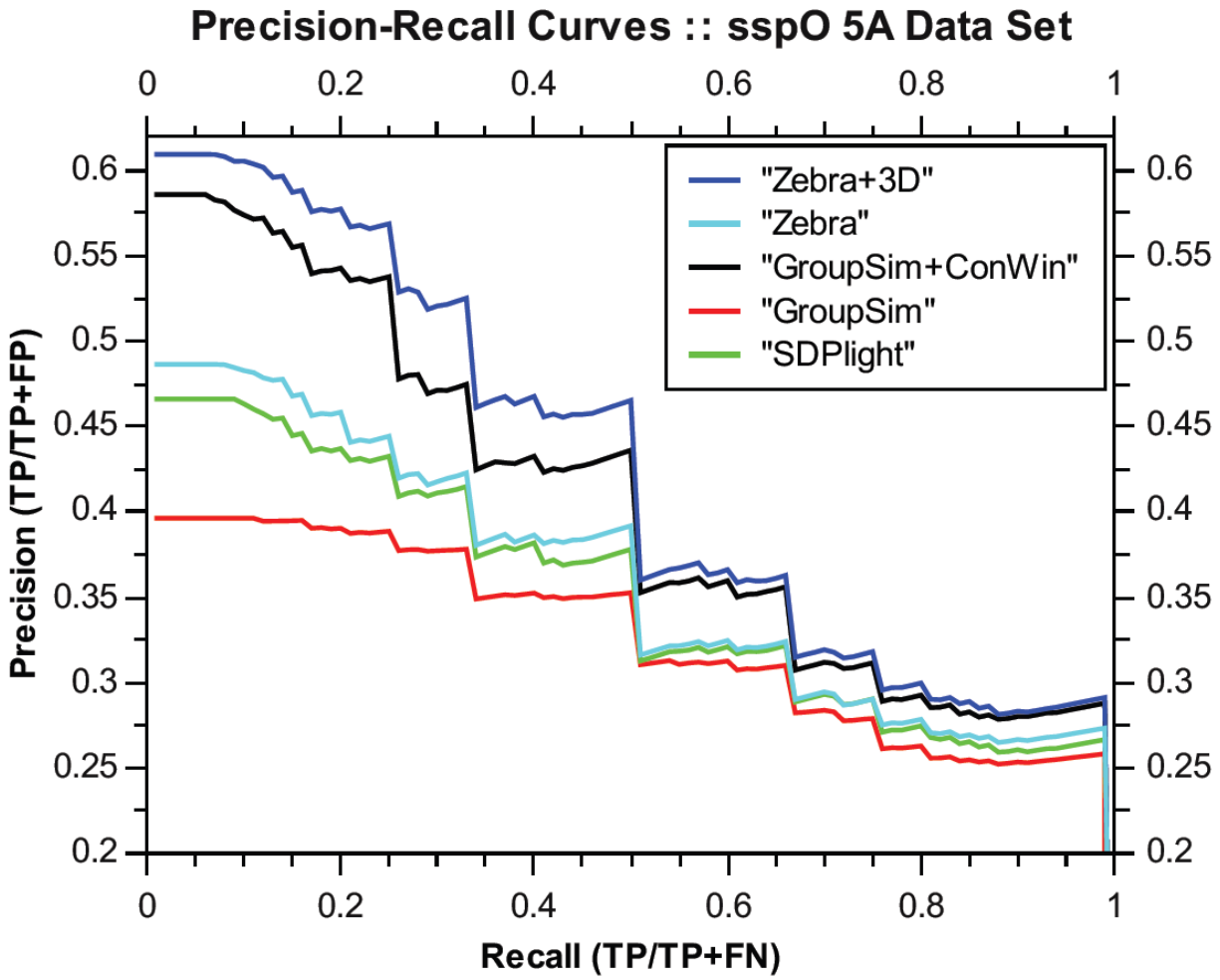

Fig S3. Precision-Recall curves for representative SSP prediction methods on sspO_5 dataset.

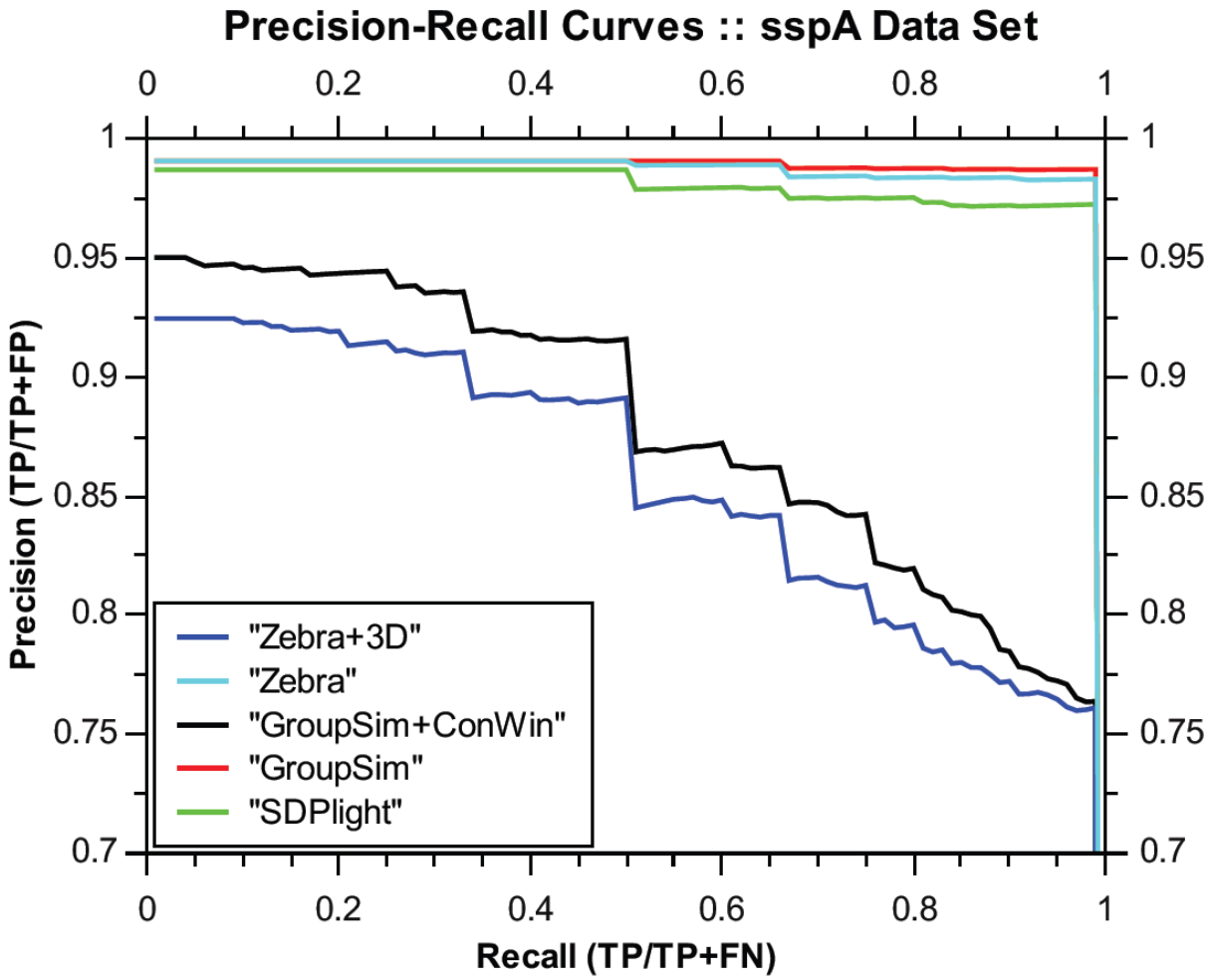

Fig S4. Precision-Recall curves for representative SSP prediction methods on sspA dataset.

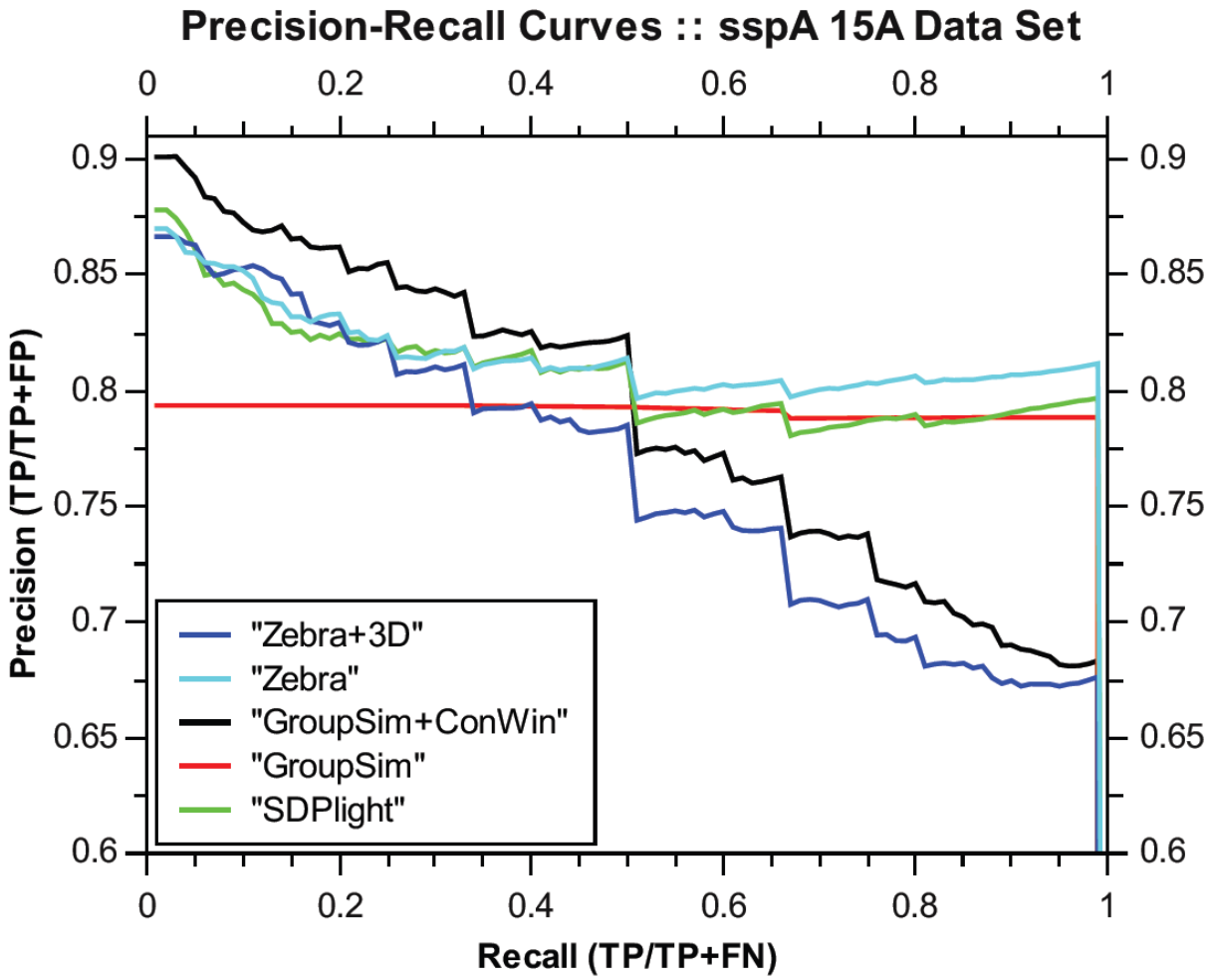

Fig S5. Precision-Recall curves for representative SSP prediction methods on sspA_15 dataset.

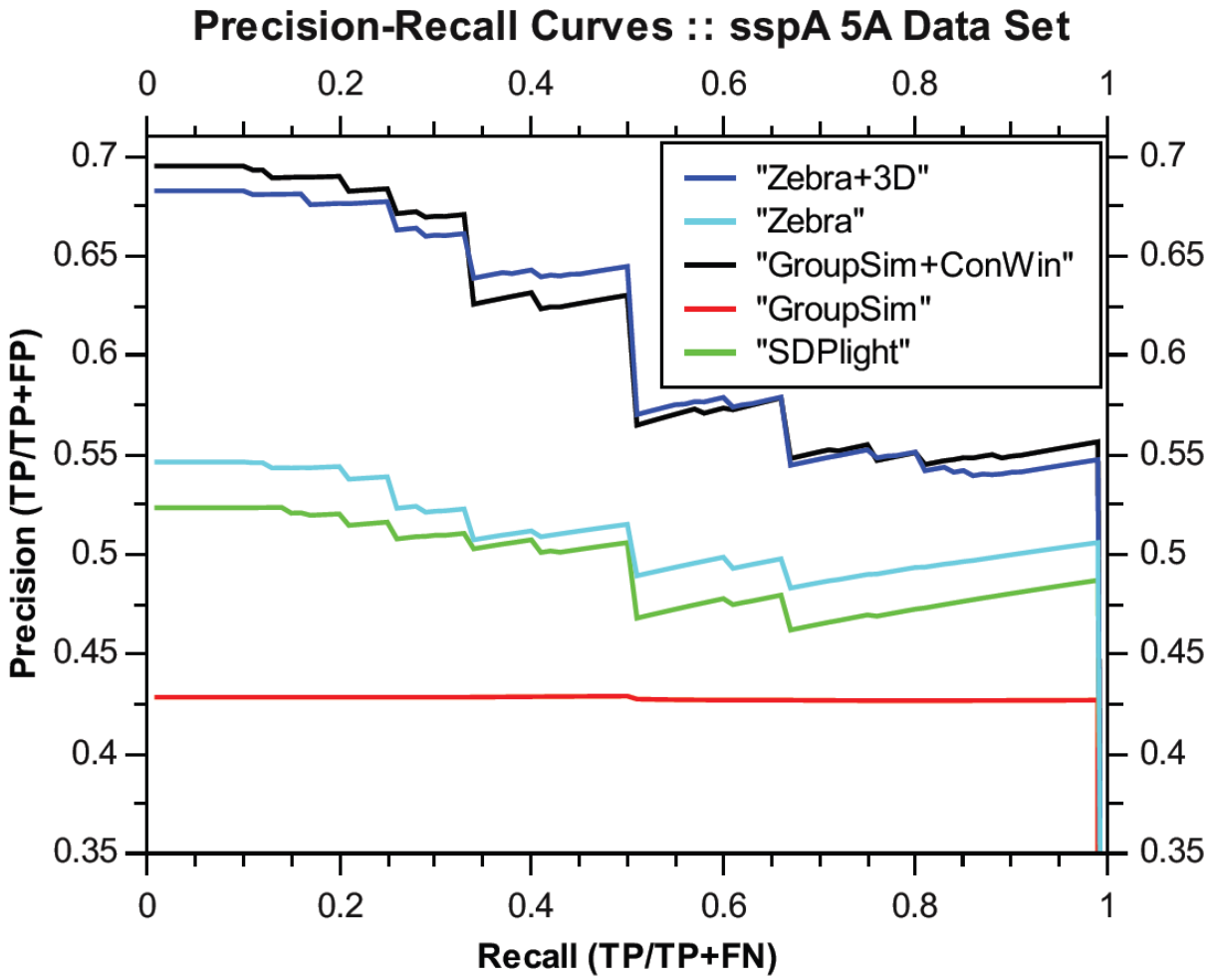

Fig S6. Precision-Recall curves for representative SSP prediction methods on sspA_5 dataset.

Table S2. Distribution of different types of SSPs and conserved positions in protein structures and secondary structure elements

	All	**Subfamily-specific**		**Conserved**
		sspO	sspA	
∞				
Ratio to group	100	100	100	100
Ratio to all within radius	100	**8,38**	**2,84**	**13.17**
β-strand	23,69	24,42	34,88	26.79
α-helix	36,66	33,01	26,12	24.85
Loop	39,65	42,58	38,99	48.35
15Å				
Ratio to group	48,97	52,55	61,04	72.89
Ratio to all within radius	100	**8,99**	**3,54**	**19.6**
β-strand	27,23	25,93	37,56	28.05
α-helix	33,89	31,95	23,98	24.16
Loop	38,88	42,12	38,45	47.79
5Å				
Ratio to group	8,17	7,77	13,15	28.22
Ratio to all within radius	100	**7,98**	**4,57**	**45.5**
β-strand	20,63	21,78	25,84	22.21
α-helix	24,70	24,89	21,96	23.3
Loop	54,67	53,33	52,20	54.49

Positions are assigned to the most strict pattern filter they pass. Secondary structures were taken from DSSP (Kabasch and Sander, 1983). α-Helix corresponds to H, G or I states in DSSP, β-sheet to E or B and loop to S, T or C. Specific positions are more frequent in loop regions. However, distribution of specific positions in secondary structure elements is in general similar to the background distribution.

REFERENCES

Capra,J.A. and Singh,M. (2008) Characterization and prediction of residues determining protein functional specificity. *Bioinformatics*, 24(13), 1473-1480

Kabasch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22, 2577–2637