Supporting Information

SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network

Yang Shen and Ad Bax

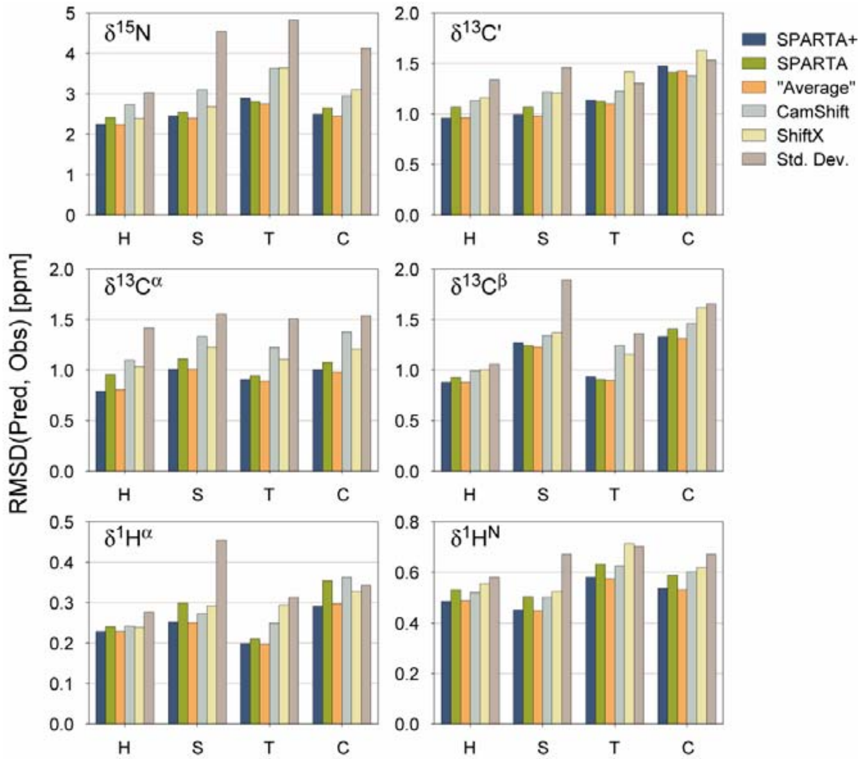

Figure S1. Chemical shift prediction performance of multiple methods for residues with different secondary structure. Chemical shifts were predicted using different methods (see text) for eleven proteins which are not present in the neural network database. The prediction performance (vertical axis) of the ^{15}N, $^{13}C'$, $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{1}H^{\alpha}$ and $^{1}H^{N}$ chemical shifts is defined by the rms difference between the experimental and the predicted chemical shifts, and plotted for residues with different secondary structure (horizontal axis), including α-helix ("H": DSSP classifications of "G", "H" and "I"), β-strand ("S": "B", "E" or "S"), turn ("T": "T") and coil ("C": blank), as defined by the program DSSP (Kabsch and Sander 1983). For comparison, the standard deviation (Std. Dev.) relative to the average of the chemical shifts for each type of secondary structure for the chemical shifts in the training database is also shown.

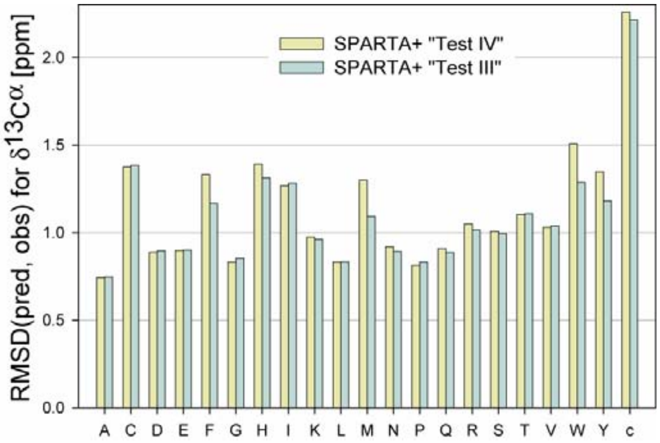

Figure S2. Impact of including the χ_2 torsion angle information on the accuracy of SPARTA+ $^{13}C^{\alpha}$ chemical shift prediction, for each amino acid type (c: oxidized Cys). The root mean square deviations between the observed $^{13}C^{\alpha}$ chemical shifts and the SPARTA+ $^{13}C^{\alpha}$ chemical shifts, as predicted by ANN networks Test III and Test IV, are calculated for each amino acid type in the validation data set. Results indicate that the addition of χ_2 torsion angle information (Test III vs Test IV) improves the prediction performance of $\delta^{13}C^{\alpha}$ for aromatic amino acids (Phe, His, Tyr and Trp) and Met.

Table S1. Chemical shift prediction performances for eleven test proteins which are not included in the SPARTA+ training database.

BMRB/PDB	RMSD (δ^{pred}, δ^{obs}) [ppm]					
	δ^{15}N	δ^{1}H$^{\alpha}$	δ^{13}C'	δ^{13}C$^{\alpha}$	δ^{13}C$^{\beta}$	δ^{1}H^{N}
			SPARTA			
16146/1enfA	2.907			1.132	1.622	0.561
16321/1wzvA	2.297	0.308	0.967	0.810	0.903	0.466
16362/1gwyA	2.352	0.297		0.902	1.196	0.522
16447/1phpA	2.750		1.242	1.099	1.050	0.653
16537/2etlA	2.738	0.297	1.223	1.416	1.369	0.585
16572/3hn9A	2.343	0.300		0.775	1.098	0.421
16656/3ipfA	2.428	0.257	0.941	0.900	0.881	0.425
16661/3gzmA	2.210			1.026	0.761	0.415
16684/3l48C	2.130	0.252		0.885	0.991	0.455
dinI/1ghh_	1.897	0.255	0.833	0.849	0.613	0.325
5589/1nxi_	2.679	0.335	1.148	0.840	1.199	0.638
Average	**2.557**	**0.290**	**1.141**	**1.041**	**1.155**	**0.544**
			SPARTA+			
16146/1enfA	2.482			1.087	1.576	0.497
16321/1wzvA	2.147	0.279	0.998	0.756	0.892	0.505
16362/1gwyA	2.394	0.255		0.806	1.206	0.454
16447/1phpA	2.537		1.165	0.893	1.049	0.551
16537/2etlA	3.023	0.254	1.182	1.333	1.290	0.58
16572/3hn9A	2.478	0.214		0.755	1.045	0.366
16656/3ipfA	2.103	0.241	0.969	0.639	0.882	0.394
16661/3gzmA	1.802			0.882	0.789	0.363
16684/3l48C	1.947	0.208		0.878	1.003	0.375
dinI/1ghh_	1.603	0.216	0.853	0.673	0.767	0.286
5589/1nxi_	2.733	0.326	1.081	0.863	1.204	0.630
Average	**2.450**	**0.251**	**1.093**	**0.935**	**1.139**	**0.493**
			Average(SPARTA, SPARTA+)			
16146/1enfA	2.524			1.071	1.564	0.500
16321/1wzvA	2.093	0.281	0.946	0.711	0.847	0.471
16362/1gwyA	2.286	0.253		0.792	1.157	0.456
16447/1phpA	2.535		1.158	0.913	1.024	0.568
16537/2etlA	2.857	0.258	1.162	1.336	1.289	0.562
16572/3hn9A	2.346	0.224		0.717	1.017	0.358
16656/3ipfA	2.108	0.228	0.911	0.666	0.851	0.376
16661/3gzmA	1.823			0.876	0.745	0.360
16684/3l48C	1.839	0.205		0.843	0.968	0.380
dinI/1ghh_	1.563	0.212	0.798	0.688	0.637	0.269
5589/1nxi_	2.620	0.316	1.065	0.819	1.162	0.608
Average	**2.399**	**0.250**	**1.075**	**0.932**	**1.113**	**0.491**
			CamShift			
16146/1enfA	3.140			1.228	1.599	0.508
16321/1wzvA	2.774	0.306	1.080	1.073	0.989	0.494
16362/1gwyA	3.066	0.276		1.056	1.341	0.524
16447/1phpA	3.255		1.278	1.418	1.237	0.630
16537/2etlA	3.401	0.256	1.315	1.572	1.384	0.611
16572/3hn9A	3.039	0.292		1.057	1.174	0.419
16656/3ipfA	2.840	0.294	1.093	1.172	0.996	0.445
16661/3gzmA	2.393			1.107	0.938	0.409
16684/3l48C	2.456	0.224		1.259	1.227	0.424
dinI/1ghh_	1.790	0.234	0.866	1.055	0.799	0.376
5589/1nxi_	2.979	0.356	1.261	1.112	1.245	0.645
Average	**3.029**	**0.282**	**1.214**	**1.261**	**1.253**	**0.540**
			ShiftX			
16146/1enfA	2.661			1.181	1.662	0.567
16321/1wzvA	2.562	0.286	1.170	0.958	1.044	0.543
16362/1gwyA	2.610	0.287		1.026	1.353	0.531
16447/1phpA	2.862		1.323	1.208	1.216	0.630
16537/2etlA	3.190	0.290	1.479	1.509	1.547	0.676
16572/3hn9A	2.833	0.280		0.990	1.229	0.436
16656/3ipfA	2.341	0.294	1.186	0.902	0.999	0.584

16661/3gzmA	2.222			1.087	0.916	0.391
16684/3l48C	2.434	0.268		1.102	1.279	0.488
dinI/1ghh_	1.876	0.261	0.928	0.918	0.706	0.396
5589/1nxi_	3.612	0.344	1.338	1.103	1.376	0.726
Average	**2.793**	**0.287**	**1.298**	**1.155**	**1.296**	**0.575**
			ProShift [a]			
16146/1enfA	3.231			1.313	1.752	0.635
16321/1wzvA	3.052	0.264	1.246	1.084	1.293	0.607
16362/1gwyA	3.205	0.424		1.157	1.579	0.610
16447/1phpA	3.339		1.424	1.344	1.479	0.705
16537/2etlA	3.299	0.324	1.411	1.478	1.400	0.694
16572/3hn9A	3.196	0.445		1.174	1.446	0.535
16656/3ipfA	2.761	0.345	1.319	1.160	1.415	0.505
16661/3gzmA	2.698			1.126	2.441	0.478
16684/3l48C	2.827	0.295		1.114	1.873	0.433
dinI/1ghh_	2.555	0.243	1.177	0.894	1.019	0.372
5589/1nxi_	2.612	0.348	1.334	1.035	1.339	0.541
Average	**3.128**	**0.361**	**1.364**	**1.233**	**1.552**	**0.611**

[a] All ProShift predicted ^{13}C chemical shifts show an offset of ~-2.0ppm. The listed RMS deviations for ProShift predicted ^{13}C chemical shifts are calculated after adding a 2.0 ppm offset.

Table S2. SPARTA+ chemical shift prediction performance of the three individual neural networks for eleven test proteins, and impact of averaging the predicted chemical shifts.

	RMSD [a] [ppm]	RMSD_1 [b] [ppm]	RMSD_2 [b] [ppm]	RMSD_3 [b] [ppm]	Average [c] [ppm]	Decrease [d] [%]
Network "Full"						
δ^{15}N	2.450	2.558	2.569	2.632	2.586	5.3
δ^{1}H$^{\alpha}$	0.251	0.260	0.257	0.256	0.258	2.6
δ^{13}C'	1.093	1.155	1.128	1.107	1.130	3.3
δ^{13}C$^{\alpha}$	0.935	1.010	0.987	0.991	0.996	6.1
δ^{13}C$^{\beta}$	1.139	1.224	1.186	1.169	1.193	4.5
δ^{1}H^{N}	0.493	0.503	0.508	0.513	0.508	3.0
Network "Test IV"						
δ^{15}N	2.504	2.622	2.582	2.593	2.599	3.7
δ^{1}H$^{\alpha}$	0.291	0.293	0.301	0.298	0.297	2.1
δ^{13}C'	1.129	1.144	1.160	1.156	1.153	2.1
δ^{13}C$^{\alpha}$	1.023	1.050	1.068	1.053	1.057	3.2
δ^{13}C$^{\beta}$	1.131	1.164	1.159	1.167	1.163	2.8
δ^{1}H^{N}	0.580	0.585	0.590	0.592	0.589	1.5

[a] RMS deviations between the predicted chemical shifts, obtained by averaging over the outputs from three separately trained neural networks, and the experimental chemical shifts.

[b] RMS deviations between the predicted chemical shifts from each of the three separately trained neural networks and the experimental chemical shifts.

[c] Average RMS deviations between the predicted chemical shifts from each of the three separately trained neural networks and the experimental chemical shifts.

[d] Decrease of the RMS deviation between the experimental and the predicted chemical shifts, obtained by averaging over the outputs from the three separately trained neural networks, versus the average RMS deviation when using the three separate neural networks.

Table S3. Impact of improved SPARTA+ chemical shift prediction on CS-Rosetta structure selection.

	DinI		Vc0424	
	SPARTA+ rescored	SPARTA rescored	SPARTA+ rescored	SPARTA rescored
Rmsd to native [Å] [a]				
C$^{\alpha}$ atoms	1.455 +/- 0.154	1.782 +/- 0.507	1.717+/- 0.101	1.763+/- 0.102
All heavy atoms	2.194 +/- 0.087	2.455 +/- 0.397	2.451+/- 0.156	2.503+/- 0.151

[a] For each protein, 10 structures with lowest energy, after scoring using the SPARTA or SPARTA+ predicted chemical shifts, are selected; the RMS coordinate deviations between these selected CS-Rosetta structures and the experimental structure are listed.

Table S4. Impact of the quality of the reference structure on the SPARTA+ chemical shift prediction performance.

BMRB#/PDB# /Resolution [a]	*RMS*($\delta^{15}N$)	*RMS*($\delta^{1}H^{\alpha}$)	*RMS*($\delta^{13}C'$)	*RMS*($\delta^{13}C^{\alpha}$)	*RMS*($\delta^{13}C^{\beta}$)	*RMS*($\delta^{1}H^{N}$)
5618/1PJX/0.85	2.57	0.24	1.14	1.00	1.17	0.56
6321/3PYP/0.85	2.04	0.26		0.78	1.26	0.45
4562/1IEE/0.94		0.25		1.14	1.61	0.38
5601/1BRF/0.95	2.65	0.22	1.29	0.92	1.44	0.66
5967/2NMZ/0.97	2.36			0.81	0.88	0.35
7132/1TQG/0.98	2.06			0.98	0.84	0.40
5223/2IIM/1.00	2.03	0.21		0.82	0.88	0.30
5623/1MN8/1.00	2.39	0.21	1.59[b]	0.90	0.88	0.45
5631/2H3L/1.00	2.13	0.20	1.06	0.96	0.97	0.44
5794/1LKK/1.00	2.00	0.20	1.18	0.73	1.18	0.44
Average	**2.30**	**0.23**	**1.22**	**0.94**	**1.15**	**0.46**
4339/1NMU/2.31	2.89	0.28	1.34	1.54	1.14	0.50
5789/1A6J/2.35	2.60	0.21	1.08	0.87	1.24	0.47
15531/2GMF/2.40	2.15	0.24	1.21	1.47	1.50	0.59
4035/1A7G/2.40	2.40	0.26		1.29	0.87	0.45
4188/1RLW/2.40	2.25	0.27	1.08	0.88	1.00	0.45
4735/1JOB/2.40	2.28	0.29	1.38	1.02	1.19	0.39
5821/1JMC/2.40	2.51	0.27	1.28	1.25	1.45	0.60
5940/1N26/2.40	2.49	0.36	1.36	0.89	1.48	0.43
6611/3DZD/2.40	2.23	0.20		0.92	0.82	0.46
interleukin5/1HUL/2.40	2.56	0.22	0.94	0.99	0.94	0.57
Average	**2.45**	**0.26**	**1.21**	**1.11**	**1.19**	**0.49**
7178/2ES7/2.80	3.13	0.25		1.30	1.20	
15796/1FBV/2.90	3.02	0.27	0.96	1.41	1.20	0.50
4417/1FSK/2.90	2.36	0.21	0.90	0.81	0.97	0.40
4437/1OB1/2.90	2.60	0.28	0.97	1.23	1.67	0.51
5155/1S3S/2.90	2.87	0.39		1.17	1.29	0.49
7301/3BEG/2.90	3.13	0.33	1.02	0.87	1.26	0.57
6442/1SYR/2.95	2.31	0.26	0.96	0.91		0.43
15578/3BA0/3.00	3.54	0.36	1.26	1.19	1.61	0.64
4843/3BPO/3.00	2.50		0.93	1.03	1.20	0.47
5022/1AY9/3.00	2.39	0.33		0.93	1.28	0.57
Average	**2.90**	**0.30**	**1.06**	**1.09**	**1.35**	**0.53**

[a] Ten proteins, with a high-resolution (≤1.0 Å) X-ray structure and with a modest-resolution (2.3-2.4 Å), respectively, are selected from the training database. Their chemical shifts data are taken from the original chemical shifts, i.e., before applying the corrections for the possible referencing offset or deuteration, and removal of the shift outliers and the shifts for the residues in the dynamic regions. Another ten testing proteins with lower resolution (~3.0 Å) X-ray structure are also prepared. The SPARTA+ predicted chemical shifts are calculated for these proteins, and the accuracy are listed in the table.

[b] The error for this protein appears to be dominated by a relatively large, ca 1ppm, $\delta^{13}C'$ reference offset for the chemical shift entries of this protein, which has not been corrected for in the value reported here.

Reference

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