“BeStSel: webserver for secondary structure and fold prediction for protein CD spectroscopy” Micsonai et al., Supplementary information

Table S1. Performance indices for the eight structural components of BeStSel on SP175[a]

	Re-optimized BeStSel											
	175-250 nm		180-250 nm		185-250 nm		190-250 nm		195-250 nm		200-250 nm	
	RMSD	Corr	RMSD	Corr	RMSD	Corr	RMSD	Corr	RMSD	Corr	RMSD	Corr
Helix1	0.027	0.98	0.031	0.98	0.031	0.98	0.028	0.98	0.030	0.98	0.030	0.98
Helix2	0.025	0.93	0.024	0.93	0.024	0.93	0.026	0.92	0.029	0.90	0.029	0.90
Anti1	0.016	0.87	0.016	0.87	0.016	0.87	0.018	0.83	0.022	0.74	0.026	0.62
Anti2	0.034	0.92	0.034	0.92	0.034	0.92	0.035	0.92	0.035	0.92	0.035	0.92
Anti3	0.041	0.88	0.040	0.89	0.040	0.89	0.040	0.88	0.043	0.87	0.046	0.85
Parallel	0.041	0.91	0.041	0.91	0.041	0.91	0.041	0.91	0.041	0.91	0.039	0.91
Turn	0.033	0.74	0.033	0.74	0.033	0.73	0.032	0.73	0.033	0.70	0.032	0.72
Others	0.052	0.83	0.054	0.82	0.054	0.82	0.058	0.78	0.059	0.77	0.063	0.74
Helix	0.041	0.98	0.041	0.98	0.041	0.98	0.042	0.98	0.046	0.98	0.045	0.98
Antiparallel	0.064	0.94	0.063	0.94	0.062	0.94	0.064	0.94	0.064	0.94	0.072	0.92
Beta	0.056	0.94	0.055	0.94	0.054	0.95	0.057	0.94	0.061	0.93	0.065	0.92
Turn+Others	0.053	0.88	0.054	0.87	0.054	0.87	0.056	0.84	0.062	0.81	0.061	0.82

	Previous BeStSel											
	175-250 nm		180-250 nm				190-250 nm				200-250 nm	
	RMSD	Corr	RMSD	Corr			RMSD	Corr			RMSD	Corr
Helix1	0.028	0.98	0.037	0.97			0.037	0.97			0.029	0.98
Helix2	0.026	0.92	0.025	0.93			0.027	0.91			0.028	0.91
Anti1	0.017	0.85	0.017	0.85			0.019	0.80			0.023	0.69
Anti2	0.038	0.90	0.035	0.92			0.038	0.91			0.036	0.92
Anti3	0.043	0.87	0.045	0.85			0.038	0.89			0.048	0.84
Parallel	0.039	0.92	0.044	0.90			0.044	0.89			0.045	0.91
Turn	0.036	0.63	0.038	0.60			0.037	0.59			0.034	0.65
Others	0.057	0.81	0.059	0.80			0.058	0.80			0.065	0.75
Helix	0.042	0.98	0.051	0.97			0.052	0.97			0.044	0.98
Antiparallel	0.067	0.93	0.068	0.93			0.068	0.93			0.075	0.91
Beta	0.060	0.93	0.060	0.93			0.056	0.94			0.071	0.91
Turn+Others	0.060	0.83	0.063	0.81			0.058	0.84			0.067	0.79

[a] Root-mean-square-deviation and Pearson correlation for different wavelength ranges are provided. [b]The original performance of the previous BeStSel is from Micsonai et al. (1).

Table S2. Comparison of the reliability of different methods for secondary structure estimation from the CD spectra. Test on β-sheet-rich or rare structures.[a]

		Helix		Antiparallel		Parallel		β-sheet		Turn+Others	
Method	Failures[a]	RMSD[b]	Corr[c]	RMSD	Corr	RMSD	Corr	RMSD	Corr	RMSD	Corr
Re-optimized BeStSel	-	0.034	0.99	0.049	0.97	0.037	0.97	0.035	0.99	0.038	0.91
Previous BeStSel	-	0.038	0.99	0.050	0.98	0.032	0.97	0.039	0.99	0.033	0.95
VARSLC	5	0.089	0.97	0.155	0.62	0.860	-0.08	0.133	0.73	0.130	0.74
LINCOMB	-	0.119	0.91	0.214	0.45	0.198	0.59	0.230	0.51	0.232	0.59
CDNN	-	0.083	0.97	0.122	0.83	0.076	0.91	0.102	0.89	0.115	0.81
SELCON	-	0.147	0.86					0.122	0.82	0.077	0.73
CONTIN	2	0.095	0.95					0.068	0.96	0.074	0.73
CDSSTR	-	0.201	0.76					0.139	0.75	0.099	0.71
K2D	-	0.198	0.84					0.152	0.79	0.153	0.55
K2D2	-	0.222	0.70					0.162	0.71	0.088	0.68
K2D3	-	0.136	0.87					0.184	0.64	0.143	0.65
CAPITO	-	0.260	0.57					0.161	0.85	0.147	0.70

Performance of different algorithms on a set of 25 external CD spectra of proteins that are either rich in β-sheets or have high α-helical content, or rare structural composition. For each spectrum, the widest available wavelength range depending on the protein was used. The performance of the previous BeStSel and other algorithms are from Micsonai et al. (1). The list of the proteins is presented in Table S2 of (1). The results for SELCON, CONTIN, CDSSTR (2), LINCOMB (3) are cross-validated, but for CDNN (4), CAPITO (5), VARSLC (6), and K2Ds(7,8)) are not cross-validated. [a]In the case of some spectra, the algorithms could not accomplish the procedure and hung up or gave error messages. [b]Root-mean-square-deviation, [c]Pearson-correlation coefficient.

Table S3. Performance of different searches for fold recognition on CATH 4.2 and CATH 4.3 single domain datasets.[a]

	CATH 4.2					CATH 4.3			
	n	Closest	Box	WKNN		*n*	Closest	Box	WKNN
Class (4)	1	90	91	93	Class (5)	1	89	91	92
Architecture	1	59	62	73	Architecture	1	59	63	72
(41)	5	86	94	97	(43)	5	86	96	97
Topology	1	38	37	56	Topology	1	39	39	54
(1310)	5	61	64	79	(1467)	5	62	69	79
	10	69	74	85		10	70	80	85
Homology	1	27	23	44	Homology	1	30	28	45
(5398)	5	46	45	66	(6540)	5	50	54	68
	10	54	55	73		10	59	66	76
	15	59	61	77		15	63	73	79

[a]The theoretical reliability of fold prediction on the CATH 4.2 and CATH 4.3 single domain dataset (<95% sequential homology) comparing the 5-fold cross-validated performance of “Closest”, “Box” and WKNN methods. In parentheses, the total numbers of classes, architectures, topologies, and homologies in the CATH 4.2 and 4.3 are shown. Values show the percentage when the correct CATH category is ranked within the top “*n*” upon the prediction. The “Closest” method decides based on the order of Euclidean distance of reference data in the eight-dimensional secondary structure space of BeStSel. Box method takes into account the expected error of BeStSel as an “RMSD box” in which it searches for the most frequent folds. WKNN method predicts the categories based on the sum of the weighted distance (reverse square city block distance) of every reference structures which belong to the same category among the K-nearest neighbors from the query point. Data for CATH 4.2 was presented earlier {Micsonai, 2018 #149}.

SUPPLEMENTARY REFERENCES

1. Micsonai, A., Wien, F., Kernya, L., Lee, Y.H., Goto, Y., Refregiers, M. and Kardos, J. (2015) Accurate secondary structure prediction and fold recognition for circular dichroism spectroscopy. *Proc Natl Acad Sci U S A*, **112**, E3095-3103.

http://www.ncbi.nlm.nih.gov/pubmed/26038575

http://dx.doi.org/10.1073/pnas.1500851112

2. Sreerama, N. and Woody, R.W. (2000) Estimation of protein secondary structure from circular dichroism spectra: comparison of CONTIN, SELCON, and CDSSTR methods with an expanded reference set. *Anal Biochem*, **287**, 252-260.

http://www.ncbi.nlm.nih.gov/pubmed/11112271

3. Toumadje, A., Alcorn, S.W. and Johnson, W.C., Jr. (1992) Extending CD spectra of proteins to 168 nm improves the analysis for secondary structures. *Anal Biochem*, **200**, 321-331.

http://www.ncbi.nlm.nih.gov/pubmed/1632496

4. Provencher, S.W. and Glockner, J. (1981) Estimation of globular protein secondary structure from circular dichroism. *Biochemistry*, **20**, 33-37.

http://www.ncbi.nlm.nih.gov/pubmed/7470476

5. Wiedemann, C., Bellstedt, P. and Gorlach, M. (2013) CAPITO--a web server-based analysis and plotting tool for circular dichroism data. *Bioinformatics*, **29**, 1750-1757.

http://www.ncbi.nlm.nih.gov/pubmed/23681122

6. Manavalan, P. and Johnson, W.C., Jr. (1987) Variable selection method improves the prediction of protein secondary structure from circular dichroism spectra. *Anal Biochem*, **167**, 76-85.

http://www.ncbi.nlm.nih.gov/pubmed/3434802

7. Perez-Iratxeta, C. and Andrade-Navarro, M.A. (2008) K2D2: estimation of protein secondary structure from circular dichroism spectra. *BMC Struct Biol*, **8**, 25.

http://www.ncbi.nlm.nih.gov/pubmed/18477405

8. Louis-Jeune, C., Andrade-Navarro, M.A. and Perez-Iratxeta, C. (2011) Prediction of protein secondary structure from circular dichroism using theoretically derived spectra. *Proteins*, **80**, 374-381.

http://www.ncbi.nlm.nih.gov/pubmed/22095872