BeStSel: a webserver for accurate protein secondary structure prediction and fold recognition from the circular dichroism spectra

Micsonai et al., Supplementary information

Table S1. Performance of different searches for fold prediction on CATH 4.2 single domain dataset

	n	Closest	Box	WKNN
Class (4)	1	90	91	93
Architecture (41)	1	59	62	73
	5	86	94	97
Topology (1310)	1	38	37	56
	5	61	64	79
	10	69	74	85
Homology (5398)	1	27	23	44
	5	46	45	66
	10	54	55	73
	15	59	61	77

The theoretical reliability of fold prediction by comparing the "Closest", "Box" and WKNN methods on the CATH 4.2 single domain dataset (<95% sequential homology). The structures of the dataset were taken as queries on fold prediction and a 5-fold cross-validated analyses was carried out. "Closest" structures were calculated by the Euclidean distance in the eight-dimensional secondary structure space of BeStSel. Values show the percentage when the closest structure has the same CATH classification or the correct fold is listed within the closest "*n*" structures in the secondary structure space. In parentheses, the total numbers of classes, architectures, topologies, and homologies in the CATH 4.2 are shown. Box method results: Reliability of protein fold prediction on CATH 4.2 by searching for structures within the "RMSD box," i.e., by taking into account the expected error of BeStSel. The percentages represent the ratio of the correct fold within the first "*n*" most frequent folds in the box. WKNN method: the predicted categories are ordered by the WKNN score, which is defined by the sum of the weighted distance (reverse square city block distance) of every structures (from the query point) among the K-nearest neighbors which belong to the same category. The comparison of the three methods shows that WKNN method exhibits the highest performance.

Table S2. Comparison of the reliability of fold prediction algorithms on SP175 reference set

		Closest			Box			WKNN	
		CATH 3.5	CATH 4.2		CATH 3.5	CATH 4.2		CATH 4.2	
	n	CD	CD	X-ray	CD	CD	X-ray	CD	X-ray
Class (4,4)	1	80	84	96	88	88	95	89	100
Architecture	1	42	42	81	56	54	70	56	82
(38, 41)	5	75	70	95	95	86	96	95	100
Topology	1	17	14	53	25	23	42	21	70
(783, 1310)	5	32	28	82	67	44	86	60	91
	10	51	35	91	84	60	93	77	93
Homology	1	10	5	51	19	16	40	18	72
(1490,5398)	5	22	19	72	54	37	81	46	86
	10	46	25	86	72	47	88	56	88
	15	47	30	88	79	54	93	67	88

The performance of fold prediction algorithms using CATH 3.5 or CATH 4.2 and the WKNN (with CATH 4.2) method was compared on the SP175 reference set. In parentheses, the total numbers of classes, architectures, topologies, and homologies in the CATH 3.5 and CATH 4.2 reference sets are shown, respectively. Performance of the "Closest" and "Box" method on CATH 3.5 was already published in (11). CD spectra of SP175 reference set was analyzed with BeStSel and the different fold prediction procedures were carried out. The values are percentages representing the ratio of the correct fold within the first "*n*" folds in the results of the corresponding searching procedure. Because the results of the CD analyses have their secondary structure estimation error, their reliability strongly affects the performance of the fold prediction. For comparison, we carried out the fold prediction starting with the correct secondary structure contents of the proteins in the SP175 reference set. This prediction is free from the secondary structure estimation errors, and provides a theoretical maximum of the fold prediction method (shown in the columns labelled by "X-ray"). It is clear from this comparison, that the three fold prediction methods have different sensitivity for secondary structure estimation errors. In case of perfect secondary structure estimation (using the X-ray data), the WKNN method outperforms the "Closest" and "Box" methods. However, in the situation starting from the CD spectra even with the quite accurate BeStSel, the performances are decreased and the "Box" and WKNN methods perform similarly. The fold prediction is upgraded to the CATH 4.2 reference database from the original CATH 3.5 version. Although the percentages are sometimes worse, we have to note that CATH 4.2 contains significantly higher numbers of topologies and homologies.

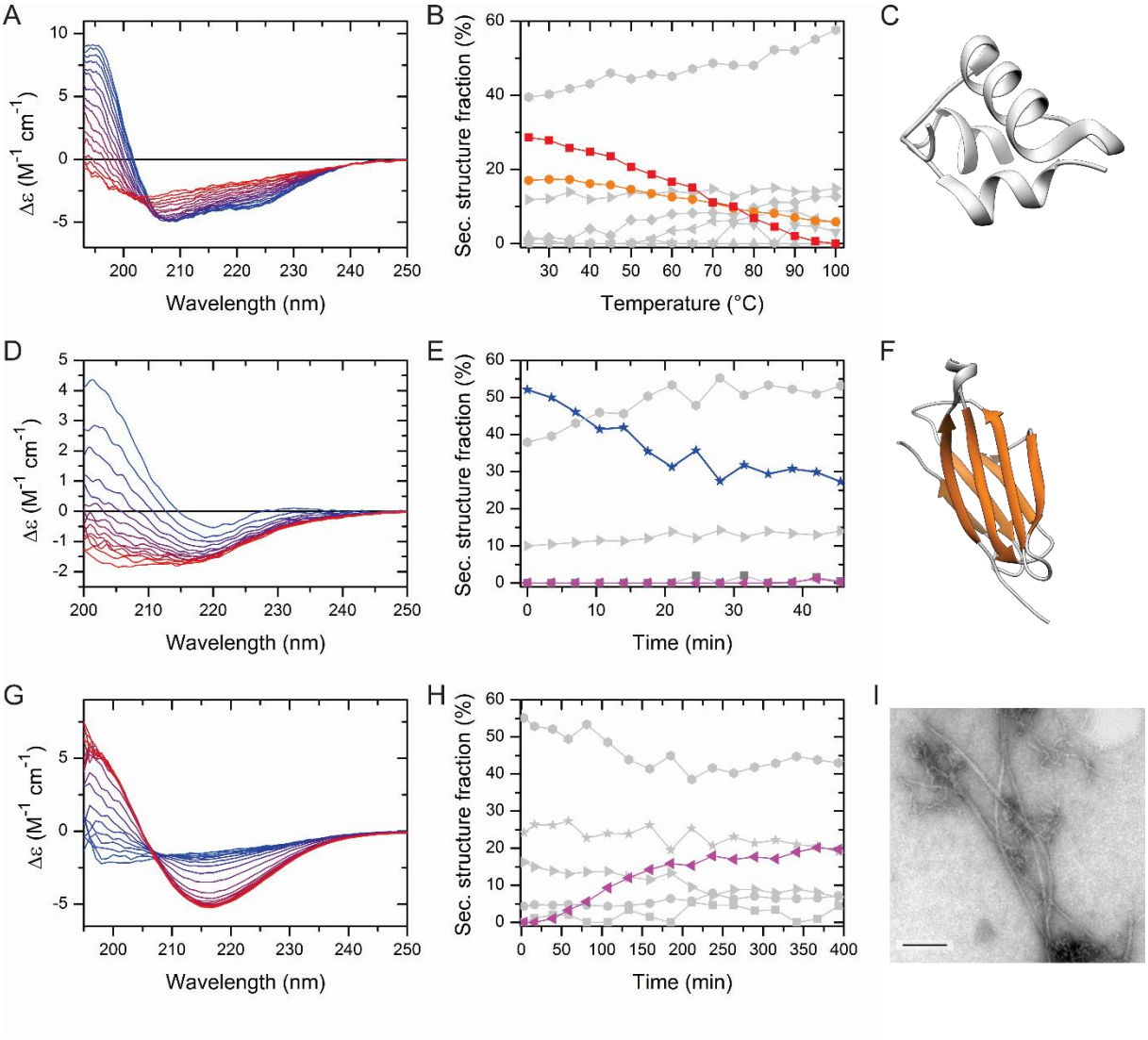

Figure S1. Conformational changes studied by CD spectroscopy and analyzed by the BeStSel web server. (A) Thermal denaturation of human insulin (0.1 mg/ml, 20 mM Na-phosphate, pH 7.4), CD spectra were taken at 5 °C steps from 25 °C to 100 °C on a Jasco J-810 instrument (1 mm pathlength, 20 nm/min scan rate, 4 sec response time, 1 nm bandwidth, accumulation: 3). (B) Secondary structure changes of the α-helical insulin upon thermal denaturation as analyzed by BeStSel. The regular and distorted α-helix components (Helix1 and Helix2) are highlighted in red and orange, respectively. The main conformational change upon denaturation is the disruption of the α-helices and the increase of disordered conformation (grey hexagons). We can conclude that the length and number of α-helices are gradually decreasing in a broad temperature range to an almost total loss at 100 °C. (C) X-ray structure of insulin (PDB: 3U4N, Vinther et al. (2012) *Plos One* 7: e30882). (D) Kinetics of partial unfolding of human β_2-microglobulin Asp38Ala variant as followed by CD spectroscopy in the presence of 0.5 mM SDS at 37 °C in 50 mM Na-phosphate, 50 mM NaCl, pH 7.4. Spectra were repeatedly recorded for 50 minutes, in a 1 mm pathlength cell, at 10 nm/min scanning rate, using 8 sec response time, 1 nm bandwidth. (E) The main conformational change is the significant decrease of the overall antiparallel β-

sheet content (blue stars). There is no parallel β-sheet detectable. (F) Antiparallel β-sheet structure of native β2m (PDB: 2YXF, Iwata et al. (2007) J. Biochem.(Tokyo) 142: 413). (G) Aggregation of amyloid-β (1-42) peptide at 0.2 mg/ml concentration in a buffer of 50 mM Na-phosphate, 30 mM NaCl, pH 7.4, as followed by CD spectroscopy at 37 °C. (H) After a lagtime, we can see the formation of parallel β-sheet structure (highlighted with purple left triangles) in parallel with the formation of the amyloid fibrils as verified by electron microscopy (I; negative staining, scale bar 100 nm). However, at neutral pH, we can observe the formation of oligomers and non-fibrillar aggregates as well, which explains the moderate parallel β-sheet content (~20 %).