Accurate secondary structure prediction and fold recognition for circular dichroism spectroscopy

András Micsonai, Frank Wien, Linda Kernya, Young-Ho Lee, Yuji Goto, Matthieu M. Refregiers, József Kardos

Accurate secondary structure prediction and fold recognition for circular dichroism spectroscopy

András Micsonai[a], Frank Wien[b], Linda Kernya[a], Young-Ho Lee[c], Yuji Goto[c], Matthieu Réfrégiers[b], and József Kardos[a,1]

[a]Department of Biochemistry and MTA-ELTE NAP B Neuroimmunology Research Group, Institute of Biology, Eötvös Loránd University, H-1117 Budapest, Hungary; [b]DISCO Beamline, Synchrotron SOLEIL, 91192 Gif-sur-Yvette, France; and [c]Division of Protein Structural Biology, Institute for Protein Research, Osaka University, Osaka 565-0871, Japan

Edited by Carl Frieden, Washington University School of Medicine, St. Louis, MO, and approved May 6, 2015 (received for review January 15, 2015)

Circular dichroism (CD) spectroscopy is a widely used technique for the study of protein structure. Numerous algorithms have been developed for the estimation of the secondary structure composition from the CD spectra. These methods often fail to provide acceptable results on α/β-mixed or β-structure–rich proteins. The problem arises from the spectral diversity of β-structures, which has hitherto been considered as an intrinsic limitation of the technique. The predictions are less reliable for proteins of unusual β-structures such as membrane proteins, protein aggregates, and amyloid fibrils. Here, we show that the parallel/antiparallel orientation and the twisting of the β-sheets account for the observed spectral diversity. We have developed a method called β-structure selection (BeStSel) for the secondary structure estimation that takes into account the twist of β-structures. This method can reliably distinguish parallel and antiparallel β-sheets and accurately estimates the secondary structure for a broad range of proteins. Moreover, the secondary structure components applied by the method are characteristic to the protein fold, and thus the fold can be predicted to the level of topology in the CATH classification from a single CD spectrum. By constructing a web server, we offer a general tool for a quick and reliable structure analysis using conventional CD or synchrotron radiation CD (SRCD) spectroscopy for the protein science research community. The method is especially useful when X-ray or NMR techniques fail. Using BeStSel on data collected by SRCD spectroscopy, we investigated the structure of amyloid fibrils of various disease-related proteins and peptides.

circular dichroism | secondary structure determination | protein fold | protein aggregation | amyloid

Optically active macromolecules, such as proteins, exhibit differential absorption of circular polarized light. The far-UV circular dichroism (CD) spectroscopy of proteins and peptides (180–250 nm) is predominantly based on the excitation of electronic transitions in amide groups. The peptide backbone forms characteristic secondary structures such as α-helices, β-pleated sheets, turns, and disordered sections with specific Φ, Ψ dihedral angles and H-bond patterns affecting the CD spectrum (1). CD has been exploited for protein folding and stability assays, intermolecular interactions, and ligand binding studies, and has recently been applied in the investigations of protein disorder (2, 3). Synchrotron radiation CD (SRCD) spectroscopy is an emerging technique complementary to small-angle X-ray scattering or infrared spectroscopy, synergistic to biochemical and biophysical assays characterizing the protein folding state. SRCD extends the limits of conventional CD spectroscopy by broadening the spectral range, increasing the signal-to-noise ratio, and accelerating the data acquisition, in the presence of absorbing components (buffers, salts, etc.) (4). Additionally, SRCD has the capability of time-resolved and stopped-flow measurements as well as high-throughput screening (3).

Quantitative analysis of CD spectra allows the prediction of the protein secondary structure content. In the past decades, a multitude of enhanced algorithms, based on variable selection, or singular value decomposition of standardized, scaled, and calibrated reference spectra, have been proposed to predict the secondary structure content, with good overall secondary structure prediction (5, 6). Validated reference spectra are nowadays available and collected in a publicly accessible Protein Circular Dichroism Data Bank (PCDDB) (7). The most populated SP175 reference dataset (8) currently available from PCDDB does not yet fully cover the fold space compared with the X-ray structures presented in Protein Data Bank (PDB) (9) (Fig. 1*C*). As a consequence, the prediction of β-sheet–rich proteins has proven to be difficult and biased due to their spectral variety and lower spectral amplitudes (Fig. 1) (5). This is assumed to be an intrinsic limitation of CD spectroscopy (11). Our goal has been to improve the accuracy and to increase the information content of the secondary structure prediction.

In the absence of high-resolution structures, CD is regarded as the method of choice, providing structural information of proteins in solution. Crystallization failure or the sheer size of macromolecules are the drawbacks for structure determination by X-ray crystallography or solution NMR spectroscopy, respectively. Examples are β-sheet–rich membrane proteins, protein aggregates, and amyloid fibrils. CD spectral results are annually increasingly cited in biophysical and macromolecular structure publications, but are too often limited to qualitative spectral differences and comparisons, lacking quantitative evaluation. SRCD should increase the information content, therefore improving quantitative secondary structure predictions (3,

Significance

Circular dichroism (CD) spectroscopy is widely used for protein secondary structure analysis. However, quantitative estimation for β-sheet–containing proteins is problematic due to the huge morphological and spectral diversity of β-structures. We show that parallel/antiparallel orientation and twisting of β-sheets account for the observed spectral diversity. Taking into account the twist of β-structures, our method accurately estimates the secondary structure for a broad range of protein folds, particularly for β-sheet–rich proteins and amyloid fibrils. Moreover, the method can predict the protein fold down to the topology level following the CATH classification. We provide a general tool for a quick and reliable structure analysis using conventional or synchrotron radiation CD spectroscopy, which is especially useful when X-ray or NMR techniques fail.

Author contributions: A.M., L.K., and J.K. designed research; A.M. and J.K. developed the concept and algorithm; A.M. programmed the algorithm; F.W. and M.R. gave technical support and conceptual advice; Y.G. was a consultant in amyloid work and gave conceptual advice; A.M., L.K., Y.-H.L., and J.K. performed research; Y.G. contributed new reagents/analytic tools; A.M. and J.K. analyzed data; A.M., F.W., M.R., and J.K. wrote the paper; and Y.-H.L. edited the manuscript.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. Email: kardos@elte.hu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1500851112/-/DCSupplemental.

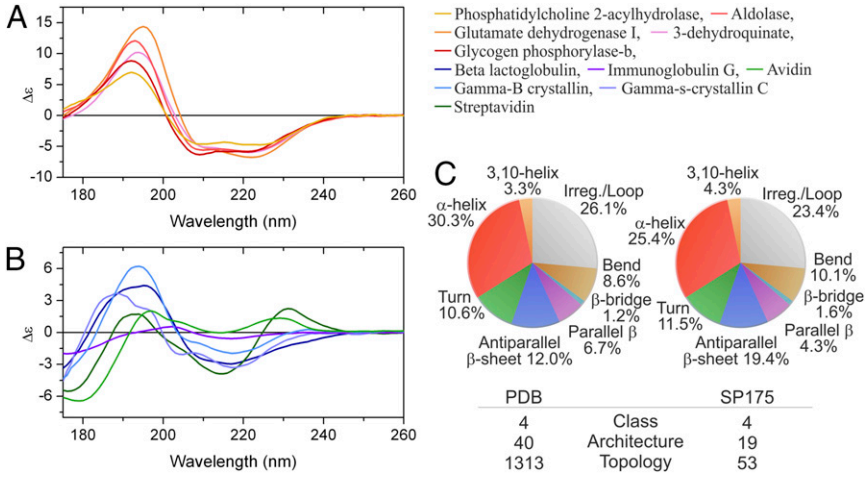

Fig. 1. SRCD spectra of α-helical and β-sheet–rich proteins and the secondary structure and protein fold occurrence in the PDB and in the SP175 CD dataset. The CD spectra of proteins containing ∼50% α-helix (*A*) are similar to one another, whereas proteins having 50% β-sheet with negligible α-helical content (*B*) show spectral properties diverse in amplitude, number, and positions of components. (*C*) The secondary structure composition of proteins in the PDB and in the SP175 reference set. Although the average secondary structure composition is similar in the two, the representation of protein folds is limited in SP175, based on the CATH topology category presented at the bottom (10).

12). For a reliable quantitative analysis of conventional CD and SRCD spectra, a suitable algorithm should correlate the spectral information to the complete fold space. This algorithm accurately predicts the secondary structure content and elucidates the folding pattern.

Protein aggregates play a central role in several degenerative disorders including amyloidosis in the central nervous system, observed for Alzheimer's and Parkinson's diseases. In vitro as in vivo, proteins can form different aggregates of various sizes and morphologies (amorphous aggregates, oligomers, protofibrils, amyloid fibrils), which have distinct physiological effects depending on the environmental conditions (13, 14). Prediction of these β-sheet–rich structures has so far been controversial. The lack of calibrated and standardized CD reference spectra of such proteins, resulted so far in a misestimation of α-helical content due to the influence of strong spectral amplitudes (Table S1 and Fig. S1). Secondary structure information is essential to understand the molecular mechanisms of self-assembly and the pathophysiological effects of these aggregates.

Here, we present a novel algorithm, β-structure selection (BeStSel), which reliably distinguishes parallel from antiparallel β-sheets by CD spectroscopy. We show that the twisting of the β-sheets has a strong influence on the CD spectrum. By taking into account the twisting angles between β-strands, our algorithm improves secondary structure prediction in general, and specifically for β-structure–rich proteins and amyloid fibrils. For the first time (to our knowledge), the increased information content obtained from the CD spectra makes protein fold prediction possible down to the topology level, in terms of the CATH protein structure classification (10).

Results

β-Sheet Twist and Its Effect on CD Spectra. An intrinsic weakness of secondary structure determination by CD arises from the structural and spectral diversity of β-structures (Fig. 1). The structural diversity manifests in variation of length, extent, direction, and degree of twist and distortion of the β-sheets and the orientation of neighboring β-strands, i.e., the antiparallel and parallel nature. These features relate to the diverse CD spectral properties. Careful inspection of the β-structures of some β-sheet–rich proteins revealed that orientation and twist of the β-sheets have a marked influence on the observed spectral features. Previous studies reported poor correlation between β-sheet twist and spectral features (15).

We determined quantitatively the twist angle distribution for the antiparallel and parallel β-sheets on the SP175 reference dataset and selected examples (Fig. 2 *A* and *B*) from their atomic coordinates using the definition of Ho and Curmi (16) for β-sheet twist (Fig. 2 *A*–*C*). For assignment of the β-sheets to the residues in the known 3D structure of a series of test proteins, we used the DSSP (Define Secondary Structure of Proteins) algorithm (17). It is notable that the twist angle distributions of individual proteins are different compared with the overall distribution in the SP175 dataset. We identified several important spectral differences, specifically an inversion of maxima and considerable peak shifts for CD spectra of two proteins containing similar, ∼50% antiparallel β-sheets but substantially different twists (Fig. 2*D*). The two CD spectra are strikingly different. The spectrum of trypsin inhibitor A, e.g., is similar to the spectra of disordered proteins down to 200 nm. These observations demonstrate the strong effect of the β-sheet twist on the CD spectrum. We have to note that the left-hand twisted β-sheet has been considered less stable than the right-hand twisted one in globular proteins (18). In agreement with this, the twist angle distribution on SP175 shows that left-hand twisted β-structure occurs in a low proportion, mainly in antiparallel β-sheets (Fig. 2*A*).

Then we have applied the general assumption that the CD spectrum of a protein is the linear combination of the basis spectra, characteristic of the secondary structural elements present in the protein (19). First, we divided the twisting angle range of antiparallel β-sheets into regions and distinguished them as different secondary structure components. Subsequently, we distinguished regular α-helix, distorted α-helix such as the two residues at each end of a helix, introduced by Sreerama et al. (20), parallel β-sheet, turn, and "others." More details are given in *Methods* and in *Supporting Information*. Basis spectra for the individual structural elements were calculated by the least-squares method applied at every wavelength point for the 71 CD spectra of SP175 reference set. By shifting the positions of the borders between the antiparallel components, we determined the characteristic spectra of the corresponding antiparallel β-sheets. To our knowledge, this is the first presentation of CD basis spectra of β-sheet as a function of the twisting angle (Fig. 2*E*). It becomes obvious that the spectral shape is strongly dependent on the β-sheet twist, and it is useful to divide the antiparallel β-sheets into subgroups. In infrared

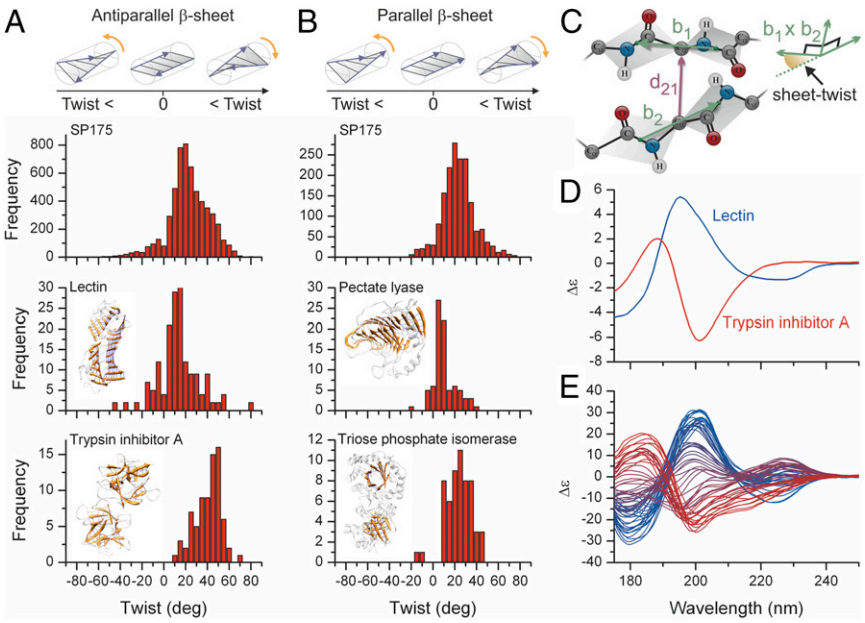

Fig. 2. Distribution of twisting angles. (*A*) Antiparallel and (*B*) parallel β-sheets. Overall distributions in SP175 are shown on the *Top*. The twist of the individual proteins can be largely different, illustrated for lectin, trypsin inhibitor A, pectate lyase, and triose phosphate isomerase. (*C*) The twisting angles were determined from the X-ray structures of the proteins (*Supporting Information*). (*D*) SRCD spectra of lectin and trypsin inhibitor A (left- and right-hand twisted antiparallel β-structures, respectively). (*E*) Calculated basis spectra of antiparallel β-sheets of different twisting angles from −10° (left-twisted, blue) to +60° (right-twisted, red).

spectroscopy, β-sheet twist is reported to correlate with the position of the amide I′ band (21), which is another example when β-sheet twist affects the optical spectrum.

β-Sheet Twist and Secondary Structure Determination. We developed a novel algorithm for secondary structure determination from protein CD spectra, named BeStSel after β-structure selection. Its characteristic features are as follows.

First, we have selected eight secondary structure components. α-Helices were divided into two components, regular and distorted, and the parallel β-sheets were distinguished from antiparallel ones. Antiparallel β-sheets were further divided into three components taking into account the twist of the antiparallel β-sheets: left-hand twisted, relaxed (slightly right-hand twisted), and right-hand twisted by using boundaries +3° and +23° of twisting angles. The last two components are the turn and the "others." Definitions of these eight components, their relation to DSSP, and comparison with secondary structure components used by other CD analysis algorithms are presented in Fig. 3*A*.

Second, a reference set of protein spectra with known structures was used to optimize the basis spectra sets, was based on SP175, and complemented with spectra of proteins with structural compositions that are absent or rare in SP175, such as native β_2-microglobulin, amyloid fibrils of the K3 fragment of β_2-microglobulin, and Alzheimer's amyloid-β (1–42) peptide (referred as SP175+).

Third, for secondary structure determination, we calculated basis-spectra sets on optimized subsets of the reference database by linear least-square approximation. The subsets of reference proteins and wavelength ranges were optimized for each secondary structure separately to provide the best prediction on the entire reference dataset (*Methods* and *Supporting Information*).

Fourth, for the secondary structure determination of an unknown spectrum, each optimized basis set will be used. The final secondary structure content is derived from these. For example, helix1 content will be the helix1 fraction from fitting with a linear combination of the spectra of the basis set that was optimized for helix1. The other seven secondary structure contents will be derived similarly from the other seven fittings. We note that the same, precalculated and fixed basis spectra sets are used for fitting the CD spectrum of any unknown protein.

The performance of the method on the SP175 reference dataset is presented in Table 1 in comparison with most of the presently used algorithms. BeStSel generally performs better than any of the previously published algorithms. Its reliability is the highest even for α-helical structures. However, the real advantage of BeStSel is revealed when analyzing CD spectra of proteins rich in β-structures. Although it has good accuracy for the total β-sheet content, it provides additional information on the antiparallel/parallel β-sheet composition and, for the first time (to our knowledge), the twist of the antiparallel β-sheets. Although some algorithms distinguish antiparallel and parallel β-sheets [e.g., LINCOMB (26), VARSLC (27), CDNN (25)], BeStSel provides unmatched RMSD values and correlation coefficients that are 0.042 and 0.90 for parallel, 0.063 and 0.94 for antiparallel, and 0.057 and 0.94 for the overall β-sheet content, respectively. For comparison, a scatter plot of reference (X-ray) and estimated secondary structure contents for the proteins of the SP175+ dataset is presented for BeStSel in comparison with LINCOMB, VARSLC, and CDNN in Fig. 4. A thorough statistical analysis has been carried out, validating in every respect the performance of BeStSel (Table S2). The algorithm works best with spectra collected down to 175 nm obtained with SRCD and almost as accurately in the conventional, 190- to 250-nm wavelength range promoting it to a general CD analysis tool. An even more dramatic difference was found between BeStSel and the conventional algorithms when they were evaluated on a test set representing CD spectra of proteins rich in β-sheets or having rare structural composition (the list of proteins presented in Table S3). BeStSel performed on this test set similarly as it did on SP175, whereas the other algorithms mostly provided high RMSD values and weak or even negative correlations to the X-ray structures (Table 2).

Protein Fold Prediction by CD Spectroscopy. To date, about 90,000 atomic resolution protein structures have been solved and deposited

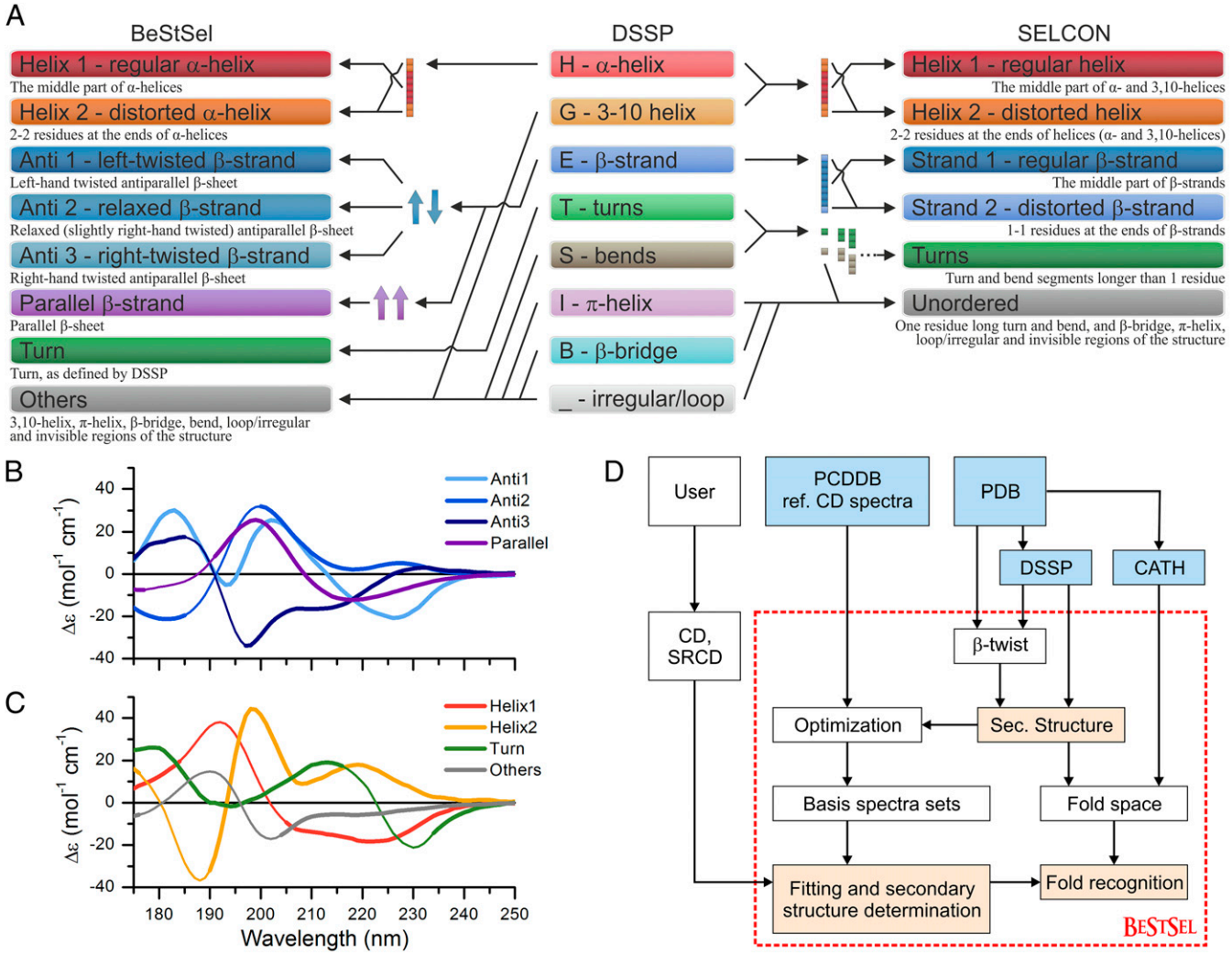

Fig. 3. Secondary structure elements and block diagram of the BeStSel method. (*A*) The secondary structure basis components of BeStSel are derived from DSSP (17). We distinguished parallel and antiparallel β-sheets and divided antiparallel β-sheets into three subgroups: left-hand twisted, relaxed, and right-hand twisted (anti1, anti2, anti3, respectively). The regular part of α-helices (helix1) and the distorted ends (helix2) are separated similarly to SELCON3 (20); however, BeStSel sorts 3_{10} helix to "others." The definition of turn is identical to that in DSSP. For comparison, basis components of SELCON3 [also used for CONTIN (22) and CDSSTR (23)] are also presented. (*B*) Optimized basis spectra of the β-sheet components of BeStSel; each of which was taken out of the respective optimized basis set. (*C*) Basis spectra for helix1, helix2, turn, and others. Wavelength regions left out in the fittings (*SI Methods*) are shown by thin lines. (*D*) Construction of the entire BeStSel package. Blue blocks represent the databases used by the method. Using the reference CD spectra and the corresponding structural information derived from PDB and DSSP, the secondary structure basis components are produced by an optimization process (Fig. S2). This has to be done only once and used for fitting to any spectra collected by CD spectroscopy to determine the secondary structure composition. On the other hand, the eight secondary structure contents provide sufficient information to predict the fold of the unknown protein in terms of the fold classification by the CATH database (10). Pink color shows the blocks of BeStSel, which are available at the bestsel.elte.hu server.

in the PDB. Domains building up these structures can be classified by different methods regarding the secondary structure motifs and their relative organization. We chose CATH (10), a hierarchical protein structure classification method (its first four levels are class, architecture, topology, and homology superfamily). The main question was whether the various protein folds at the level of CATH architectures or topologies differ in secondary structure composition sufficiently to be distinguished. In case of using a 3D secondary structure space, such as α, β and "others" contents, the answer was no, the folds are largely overlapping and different folds may have similar composition. However, using more secondary structure basis components as descriptors of the protein structure, the different folds could be separated. We examined and compared in this respect the basis components of our BeStSel algorithm to that of DSSP and SELCON3 algorithm (CONTIN and CDSSTR algorithms also use the basis components of SELCON3). Any protein structure from the PDB can be represented by a point in the secondary structure space, which has eight dimensions for BeStSel and DSSP and six dimensions for SELCON3.

Taking any structure from the single-domain PDB subset, a search was carried out for the closest structures regarding the Euclidean distance, and then we examined if the closest structures have the same fold. We identified that the eight components of BeStSel are suitable to distinguish protein folds down to the level of topologies and even homology superfamilies (Table 3). At 62%, the closest structure has the correct architecture out of 38 different ones, and at 44%, the closest hit is the right topology, out of 783. Intriguingly, despite its eight components, DSSP provided poorer results, similar to SELCON3. The explanation is that some of its components, the 3_{10} helix, π-helix, and bends are not distinctive enough for the protein fold. In contrast, the parallel β-sheet content and the antiparallel

Table 1. Comparison of the reliability of different methods for secondary structure estimation from the CD spectra: Performance on the original reference set

			Helix		Antiparallel		Parallel		β-sheet		Turn+Others		
Method	Ref. prot.	Range, nm	RMSD*	Corr†	RMSD	Corr	RMSD	Corr	RMSD	Corr	RMSD	Corr	Ref.
BeStSel	73	175–250	0.042	0.98	0.067	0.93	0.039	0.92	0.060	0.93	0.060	0.83	
BeStSel	73	190–250	0.052	0.97	0.068	0.93	0.044	0.89	0.056	0.94	0.058	0.84	
BeStSel	73	200–250	0.044	0.98	0.075	0.91	0.045	0.91	0.071	0.91	0.067	0.79	
SELCON_mod‡	73	175–250	0.082	0.92	0.103	0.82	0.073	0.72	0.088	0.85	0.076	0.69	
SELCON3§	71	175–250	0.063	0.96					0.083	0.86	0.078	0.70	24
CONTIN§	16	178–260	0.075	0.96	0.182	0.23	0.080	0.39	0.218	0.12	0.187	0.50	22
CONTIN	71	175–250	0.066	0.95					0.069	0.90	0.077	0.68	¶
CDSSTR	71	175–250	0.113	0.90					0.090	0.84	0.089	0.65	¶
CDNN§	17	178–260	0.100	0.93	ND	0.91	ND	0.63	0.110	0.73	0.050	0.82	25
LINCOMB§	16	178–260	0.059	0.98	0.097	0.75	0.076	0.42	0.068	0.89	0.094	0.83	26
VARSLC§	16	178–260	0.063	0.97	0.092	0.78	0.055	0.67	0.098	0.76	0.095	0.83	27
CCA§	17	178–260	0.100	0.96					0.180	0.62	0.180	0.39	28
K2D2§,#	71	190–240	0.080	0.92					0.100	0.81	ND	ND	29
K2D3§,#	71	190–240	0.070	0.93					0.100	0.84	ND	ND	30
CAPITO§	107	178–260	0.110	0.96					0.130	0.80	ND	ND	31

Cross-validated statistics. The number of proteins used as reference set for the algorithms are indicated in the second column. For our algorithm, BeStSel, performance on the 190- to 250- and 200- to 250-nm range is also provided.

*Root-mean-square deviation.

†Pearson correlation coefficient.

‡We applied the eight secondary structure elements introduced in the present work for the SELCON [SELMAT (24)] algorithm to compare with the BeStSel method.

§For these methods, it was impossible to obtain or use the original source codes to calculate cross-validated statistics on SP175; thus, parameters from the original publications are provided.

¶Calculated by applying the original algorithms on SP175 reference set.

#K2D2 and K2D3 use 190–240 nm as maximal wavelength range.

β-sheets of different twists provided by BeStSel are better descriptors and suitable to distinguish different folds from each other. Based on these findings, one method to determine the fold of single-domain proteins from CD is to analyze the spectrum by BeStSel and then search the PDB single-domain subset for the closest structures. However, this method does not take into account the possible error of the secondary structure estimation. Thus, the closest structures are not necessarily the ones with the correct fold. Moreover, some folds are overlapping in the secondary structure space. Therefore, it is more informative to examine the neighborhood of the BeStSel result within the expected errors. This defines a "box" in the eight dimensional secondary structure space and allows a survey of the folds and their frequencies (Fig. 5). This method is especially useful when the vicinity of the BeStSel result is crowded. In some cases, such as the glycogen phosphorylase-b (CATH 3.40.50), over 500

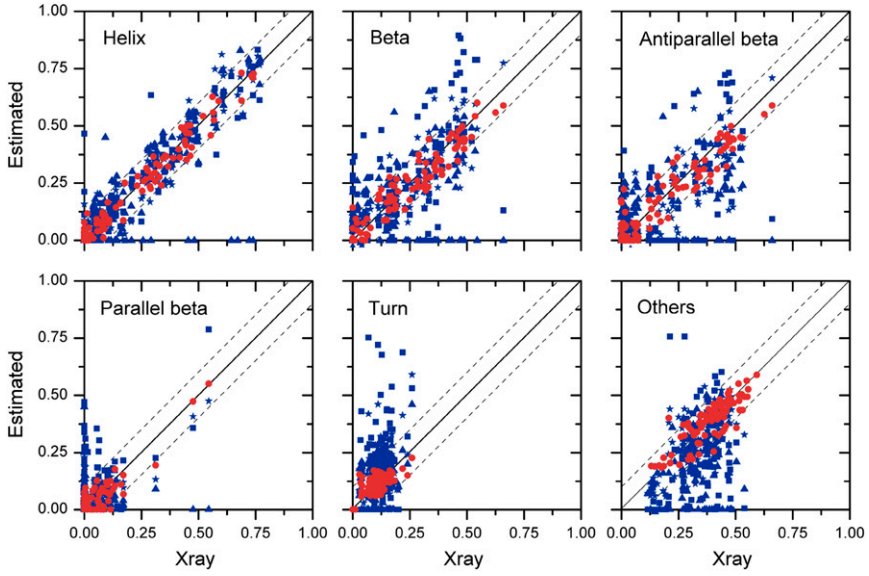

Fig. 4. Scatter plot of reference (X-ray) and estimated content of different secondary structures by BeStSel (red circles), LINCOMB (blue squares), VARSLC (blue triangles), and CDNN (blue stars) algorithms on the proteins of SP175+ dataset. The results are cross-validated for BeStSel. The 45° line represents the perfect prediction when the estimated value equals to that of the X-ray structure. The dashed lines are the ±0.1 borders.

Table 2. Comparison of the reliability of different methods for secondary structure estimation from the CD spectra: Test on β-sheet–rich or rare structures

		Helix		Antiparallel		Parallel		β-sheet		Turn+Others	
Method	Failures*	RMSD	Corr	RMSD	Corr	RMSD	Corr	RMSD	Corr	RMSD	Corr
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 proteins that are either rich in β-sheets or have high α-helical content, or rare structural composition in the SRCD and conventional wavelength ranges. For each spectrum, the widest available wavelength range depending on the protein was used. The list of the proteins is presented in Table S2. Note, that all of the methods except BeStSel show dramatically decreased reliability. The performance of BeStSel is even better than that on SP175 (Table 1), possibly due to the "purer" secondary structure contents.
*In the case of some spectra, the algorithms could not accomplish the procedure and hung up or gave error messages.
†The results are cross-validated except for CDNN, CAPITO, VARSLC, and K2Ds.

structures of various folds can be found in the box. However, the most frequent one is the three-layer sandwich (α–β–α) at the level of architecture and the Rossmann fold at the level of topology, which are characteristic for this protein. Tested on SP175, in 88% the right class is the one in the first place, the architecture is found at 56% and the topology is at 25% in the first place (out of 783). However, if the correct fold is a rare fold overlapping with a highly populated one, it will be necessarily suppressed and appear at the back of the list. Therefore, we cannot expect the first fold in the list to be the most probable. For these reasons, we do not give a probability for the predicted hits. The important point is that the right fold should appear in the list. It was 95% to find the correct architecture within the first 5 most frequent folds and 84% success to predict the correct topology among 10 possible ones from 783 (Table 3). A further hint to find the most probable fold among the hits might be the chain length. In rarely populated areas of the space, finding the closest structures might work better, especially when no structure can be found in the box, i.e., within the expected error of the method (Fig. 5).

Previously, only the tertiary structure class (all-α, all-β, α + β, α/β, and denatured) could be predicted from the CD spectra by CLUSTER algorithm in CDPro package (12, 32). For comparison, we applied that method on the single domain proteins of SP175 by merging the α + β and α/β groups to make four classes, similarly to CATH classes. It predicted the class with 58% success.

In the case of multidomain proteins, one can search for the closest secondary structures in the entire PDB and, for further consideration, examine the domain composition and chain lengths of the structures found.

Our method is complementary to theoretical methods predicting the protein fold by primary sequence similarities. The main advantage is the experimental identification of the protein fold. It is worthwhile to note that in some cases even highly homologous protein sequences can take up different folds, and sometimes sequences with low sequence homology can take up the same one. In the case of recombinant proteins, our method is a fast and inexpensive way to verify the correct fold.

Extending the Limits of Secondary Structure Determination: Selected Examples. The determination of β-sheet content of highly twisted β-structures has been controversial so far. BeStSel accurately estimates the secondary structure of such proteins (*SI Results*, Table S4, and Fig. S3 *A–D*). Moreover, information on the type of β-structure is provided, i.e., highly right-twisted (high Anti3 content) and the likelihood of the successfully predicted fold.

The structure and fold prediction of mixed α/β protein, isopropylmalate dehydrogenase (IPMDH)-containing parallel β-sheet, is presented in *SI Results*, Table S4, and Fig. S3 *E* and *F*.

We analyzed the CD spectra of various forms of β_2-microglobulin, polyglutamine, and Aβ (1–42) that are related to dialysis-related amyloidosis, Huntington's disease, and Alzheimer's disease, respectively. We found that amyloid fibrils show a large spectral and structural variety. Prediction of the β-sheet content and identification of the parallel or antiparallel organization of the β-strands (*SI Results*, Tables S5 and S6, and Fig. S3 *G–M*) correlate with earlier observations regarding the different morphology and behavior of aggregates of these proteins. Such data are important for understanding the aggregation mechanism and the pathological and physiological properties of the aggregates. We also found that the structure of the amyloid fibrils of GNNQQNY peptide fragment of Sup35 yeast prion is markedly different from the crystal structure determined by Eisenberg and coworkers (33) indicating potential differences of crystallized and soluble amyloid fibril proteins (*SI Results*, Table S6, and Fig. S3 *N–P*).

Discussion

A growing community of CD spectroscopists strives for a reliable and accurate spectral analysis. We specifically aimed for the problematic secondary structure determination of proteins with high β-sheet content, encountered in protein aggregates such as amyloids, in β-barrels and in β-helices. We discovered that β-sheet twisting is manifest in the CD spectrum and the spectral differences between β-sheet-containing proteins are clearly related to the different β-sheet twist distributions. Our results demonstrate that inclusion of new β-sheet components, which are distinguished by their twist, significantly improve the secondary structure determination of proteins in general and especially for proteins with high β-sheet content. We extended the secondary structure components to eight ones, including four β-sheet components (antiparallel right, left, and relaxed twisted as well as parallel β-sheets) and redefined the helix, turn, and "other" components.

Table 3. The reliability of fold prediction

Method	PDB*,†					SP175‡,§			
	n	BeStSel	SELCON3	DSSP	K2D3	*n*	BeStSel	SELCON3	K2D3¶
Closest structure method results, %			Class (4)					Class (4)	
	1	91	88	88	86	1	80	63	63
			Architecture (38)					Architecture (38)	
	1	62	50	49	40	1	42	38	20
	5	87	82	81	75	5	75	56	53
			Topology (783)					Topology (783)	
	1	44	29	30	19	1	17	13	5
	5	68	56	54	40	5	32	24	22
	10	75	66	64	51	10	51	34	29
			Homology (1,490)					Homology (1,490)	
	1	35	22	23	11	1	10	7	0
	5	56	43	42	25	5	22	13	12
	10	64	52	51	35	10	46	20	17
	15	68	58	57	42	15	47	23	24
RMSD box method results, %			Class (4)					Class (4)	
	1	92	89	85	89	1	88	79	73
			Architecture (38)					Architecture (38)	
	1	64	54	48	49	1	56	48	41
	5	95	93	88	90	5	95	90	83
			Topology (783)					Topology (783)	
	1	43	31	29	27	1	25	21	17
	5	77	68	58	52	5	67	47	37
	10	88	79	73	66	10	84	59	47
			Homology (1,490)					Homology (1,490)	
	1	32	21	18	14	1	19	10	7
	5	65	49	43	34	5	54	29	25
	10	80	67	60	49	10	72	45	39
	15	85	78	71	59	15	79	64	54

*Closest structure method results: The theoretical reliability of fold prediction based on the closest structures by calculating the Euclidean distance in the eight-dimensional secondary structure space of BeStSel on a single-domain subset of PDB filtered for ≤90% sequential homology; and comparison with the secondary structure space used by SELCON3 algorithm of six components, the eight components of DSSP, or the three components of K2D3. 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 single-domain PDB subset are shown.

†RMSD box method results: Theoretical reliability of protein fold prediction on the above-mentioned PDB subset, by searching for structures within the "RMSD box," i.e., by taking into account the expected error of BeStSel and for comparison, SELCON3, DSSP, and K2D3 algorithms (*Supporting Information*). The percentages represent the ratio of the correct fold within the first "*n*" most frequent folds in the box.

‡Closest structure method results: Reliability of the closest structure method was tested on SP175. The secondary structure compositions were determined from the CD spectra of SP175 in a cross-validated way and then analyzed for the fold.

§RMSD box method results: The real test of the method on the SP175 reference CD spectra by calculating the secondary structures and doing the search on the PDB subset provides similar results.

¶Based on noncross-validated results.

Our observations are fully empirical. There has been a great effort spent to establish the theoretical basis of the origin of protein CD spectra in the last decades. Generally, the appearance of the far-UV CD spectrum is attributed to the $n\pi^*$, $\pi\pi^*$ electronic transitions of the peptide bonds reviewed by Sreerama and Woody (34) and charge transfer transitions between the neighboring peptide bonds (35, 36). Theoretical calculations of the CD spectrum were carried out by ab initio and semiempirical calculations, thoroughly reviewed by Woody (37). Woody pointed out that these methods, despite the respectable correlation coefficients, fall short of achieving satisfactory agreement with experiment. The best results were obtained on α-helical proteins. In case of β-sheets, the calculations gave poor results on β-II proteins, showing a large positive maximum around 200 nm, instead of the experimentally observed negative amplitude. This conflict was explained by Woody and Sreerama by the presence of the PPII structure in such proteins and the inability of the matrix method to treat it (15, 37). Hirst et al. (38) interpreted this result as the effect of the dynamic properties of the protein structure. In our work, we found that β-II proteins have strongly right-hand twisted antiparallel β-sheet structure, and there is a good correlation between the β-sheet twist and the unique spectral features of β-II proteins. The theoretical consideration of the twist of β-sheets is still to be assessed.

Previously, Wallace and coworkers (39, 40) assumed that information gained from SRCD spectroscopy might be the basis for new methods for the identification of the protein folds in the future. Beyond the secondary structure, the orientation and twisting of β-sheets together with the helical and "others" secondary structure components are characteristic also for the protein folding pattern. For the first time (to our knowledge), reliable fold prediction down to the topology level (CATH protein structure classification) is possible. This facilitates the assessment of ubiquitous recombinant proteins in the laboratory before crystallization and X-ray structure determination or NMR spectroscopy.

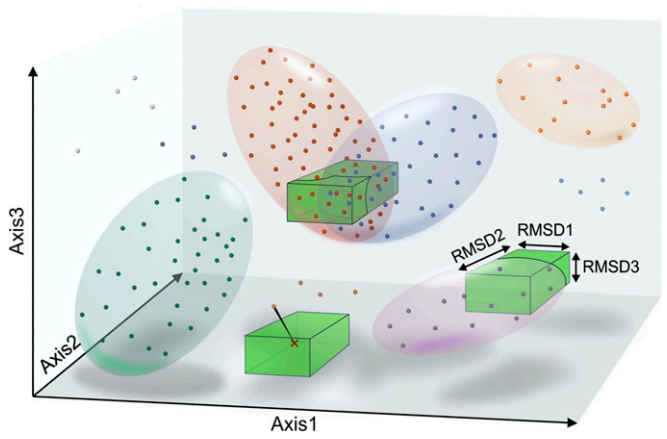

Fig. 5. Fold prediction with BeStSel. Schematic representation of protein structures in the secondary structure space in three dimensions. Every protein structure from the PDB can be represented by a point in this space. Different colors represent different folds. The result of the secondary structure analysis of the CD spectrum and its expected error determines a "box" in this space. Depending on the secondary structure composition, a lot of PDB structures can fall in such a box, sometimes with different, overlapping folds (in the level of architecture or topology). In other cases, a small number of structures can be found in it or sometimes even the closest structure is outside the expected range of the CD analysis result. These cases are represented with green boxes. The fold can be predicted by surveying the structures within the expected error of the CD results, i.e., inside the "box," or by searching for the closest structures by measuring the Euclidean distance.

The findings have been implemented in a freely accessible algorithm, BeStSel (bestsel.elte.hu), which allows structural biologists to solve the secondary structure of their protein samples accurately and to gain more structural information (for short description of the home page, see *Supporting Information*). The versatile algorithm presents a significant improvement compared with existing ones. With fast CD spectral data acquisition and accurate structure determination, our algorithm provides biophysical scientists with a powerful tool. We used a defined, but limited reference set, containing only 73 spectra, which is the limiting factor for the performance of our algorithm. In the future, inclusion of more reference spectra and expansion of the parallel β-sheet components shall further improve the reliability of the algorithm.

Methods

The Twist of the β-Sheets. We adopted the definition of Ho and Curmi (16) for the twist angle of the β-sheet (Fig. 2 *A*–*C*). The twist angle is determined for two adjacent residues of two neighboring β-strands in a β-sheet, the angle between the two neighboring peptide backbones at the location of the residues. This definition provides a number of angle values equal to $(n - 1) \times r$ in an ideal β-sheet of n strands with strand lengths of r residues each. The left and right directions of twist correspond to the twist direction of the two neighboring β-strands around each other (at the *Top* of Fig. 2 *A* and *B*). For assignment of the β-sheets to the residues in the known 3D structure of a series of test proteins, we used the DSSP algorithm (17). We determined quantitatively the twist angle distribution of β-sheets in proteins from their atomic coordinates. We calculated the twist angle distribution for the antiparallel and parallel β-sheets on the SP175 reference dataset and selected examples (Fig. 2 *A* and *B*).

Secondary Structure Determination from CD. A fully detailed description of the BeStSel method is provided in *Supporting Information*. Its main features are as follows.

First, we have selected eight secondary structure components. The assignment of the main secondary structure components to the residues of the known structures of the reference set was based on DSSP algorithm (17). α-Helices were divided into two components, regular (middle part) and distorted (the last two residues at each end). Parallel β-sheets were distinguished from antiparallel ones. Antiparallel β-sheets were further divided into three components taking into account the twist of the antiparallel β-sheets: left-hand twisted, relaxed (slightly right hand-twisted), and right-hand twisted. The last two components are the turn and the "others."

Definitions of these eight components, their relation to DSSP, and comparison with secondary structure components used by other CD analysis algorithms are presented in Fig. 3*A*. We found that the secondary structure estimation gave the best results on SP175 by using boundaries +3° and +23° of twisting angles to separate the three antiparallel groups.

Second, the reference set of protein spectra with known structures, used to optimize the basis spectra sets, was based on SP175 and complemented with spectra of proteins with structural compositions that are absent or rare in SP175, such as native β_2-microglobulin, amyloid fibrils of the K3 fragment of β_2-microglobulin, and Alzheimer's amyloid-β (1–42) peptide.

Third, for secondary structure determination, we calculated basis-spectra sets on optimized subsets of the reference database by linear least-square approximation. For any secondary structure component, a separate subset was generated by subsequently leaving out proteins one by one, in the absence of which the secondary structure prediction on the entire reference set improved most significantly for that particular secondary structure component. Simultaneously, an optimization of the wavelength range was carried out. A wavelength region was left out in the absence of which the prediction on the entire reference set was improved most for the particular secondary structure. In summary, we generated eight sets of eight basis spectra optimized for eight secondary structure components. The subsets of reference proteins and wavelength ranges provided the best prediction on the entire reference dataset.

Fourth, for the secondary structure determination of an unknown spectrum, each optimized basis set will be used, providing eight sets of the eight secondary structure contents. The final secondary structure content is derived from these. For example, helix1 content will be the helix1 fraction from fitting with a linear combination of the spectra of the basis set that was optimized for helix1 (with the constraint that the factors sum up to 1). The remaining seven secondary structure contents will be derived similarly from the other seven fittings. Because they do not necessarily sum up to 1, there is a final normalization. We note that the same, precalculated and fixed basis spectra sets are used for fitting the CD spectrum of any unknown protein.

Fold Prediction from CD Spectra. For prediction of the protein fold from the CD spectrum, we searched for protein molecules with known structures deposited in the PDB that have secondary structure composition similar to the result of the CD spectrum analysis. In the case of single-domain proteins, we expected a reasonable fold prediction by the CATH classification (10) of the similar, known structures. We filtered a single-domain subset from the PDB to make a nonredundant collection of chains containing single CATH domains and filtered for ≤90% sequence homology and resolution better than 3.0 Å. This dataset contains 10,433 polypeptide chains covering four classes, 38 architectures, 783 topologies, and 1,490 superfamilies evenly covering the structural space of proteins.

ACKNOWLEDGMENTS. We thank Ronald Wetzel for the polyQ CD spectrum and helpful advice. We thank Beáta Vértessy, Judit Ovádi, Péter Tompa, Ágnes Tantos, József Dobó, Balázs Major, Péter Gál, Mária Vass, Károly Liliom, Péter Závodszky, Mihály Kovács, Gábor Pál, and László Nyitray for providing protein samples for SRCD measurements, and János Kovács for electron microscopy. J.K. was supported by Bolyai János Scholarship of the Hungarian Academy of Sciences. This work was supported by the Hungarian Scientific Research Fund (K81950) and KTIA_NAP_13-2-2014-0017. SRCD measurements were supported by SOLEIL (Proposals 20140646, 20130475, 20120589, 201110054, and 20110405).

1. Woody RW (1995) Circular dichroism. *Methods Enzymol* 246:34–71.
2. Matsuo K, Sakurada Y, Yonehara R, Kataoka M, Gekko K (2007) Secondary-structure analysis of denatured proteins by vacuum-ultraviolet circular dichroism spectroscopy. *Biophys J* 92(11):4088–4096.
3. Wallace BA (2009) Protein characterization by synchrotron radiation circular dichroism spectroscopy. *Q Rev Biophys* 42(4):317–370.
4. Wallace BA (2000) Synchrotron radiation circular-dichroism spectroscopy as a tool for investigating protein structures. *J Synchrotron Radiat* 7(Pt 5):289–295.
5. Greenfield NJ (2006) Using circular dichroism spectra to estimate protein secondary structure. *Nat Protoc* 1(6):2876–2890.
6. Woollett B, Whitmore L, Janes RW, Wallace BA (2013) ValiDichro: A website for validating and quality control of protein circular dichroism spectra. *Nucleic Acids Res* 41 (Web Server issue):W417–W421.
7. Whitmore L, et al. (2011) PCDDB: The Protein Circular Dichroism Data Bank, a repository for circular dichroism spectral and metadata. *Nucleic Acids Res* 39(Database issue):D480–D486.

8. Lees JG, Miles AJ, Wien F, Wallace BA (2006) A reference database for circular dichroism spectroscopy covering fold and secondary structure space. *Bioinformatics* 22(16):1955–1962.
9. Bernstein FC, et al. (1977) The Protein Data Bank: A computer-based archival file for macromolecular structures. *J Mol Biol* 112(3):535–542.
10. Orengo CA, et al. (1997) CATH—a hierarchic classification of protein domain structures. *Structure* 5(8):1093–1108.
11. Khrapunov S (2009) Circular dichroism spectroscopy has intrinsic limitations for protein secondary structure analysis. *Anal Biochem* 389(2):174–176.
12. Sreerama N, Venyaminov SY, Woody RW (2001) Analysis of protein circular dichroism spectra based on the tertiary structure classification. *Anal Biochem* 299(2):271–274.
13. Orbán G, et al. (2010) Different electrophysiological actions of 24- and 72-hour aggregated amyloid-beta oligomers on hippocampal field population spike in both anesthetized and awake rats. *Brain Res* 1354:227–235.
14. Selkoe DJ (2003) Folding proteins in fatal ways. *Nature* 426(6968):900–904.
15. Sreerama N, Woody RW (2003) Structural composition of betaI- and betaII-proteins. *Protein Sci* 12(2):384–388.
16. Ho BK, Curmi PM (2002) Twist and shear in beta-sheets and beta-ribbons. *J Mol Biol* 317(2):291–308.
17. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22(12):2577–2637.
18. Chothia C (1973) Conformation of twisted beta-pleated sheets in proteins. *J Mol Biol* 75(2):295–302.
19. Adler AJ, Greenfield NJ, Fasman GD (1973) Circular dichroism and optical rotatory dispersion of proteins and polypeptides. *Methods Enzymol* 27:675–735.
20. Sreerama N, Venyaminov SY, Woody RW (1999) Estimation of the number of alpha-helical and beta-strand segments in proteins using circular dichroism spectroscopy. *Protein Sci* 8(2):370–380.
21. Zandomeneghi G, Krebs MR, McCammon MG, Fändrich M (2004) FTIR reveals structural differences between native beta-sheet proteins and amyloid fibrils. *Protein Sci* 13(12):3314–3321.
22. Provencher SW, Glöckner J (1981) Estimation of globular protein secondary structure from circular dichroism. *Biochemistry* 20(1):33–37.
23. Sreerama N, Woody RW (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(2):252–260.
24. Lees JG, Miles AJ, Janes RW, Wallace BA (2006) Novel methods for secondary structure determination using low wavelength (VUV) circular dichroism spectroscopic data. *BMC Bioinformatics* 7:507.
25. Böhm G, Muhr R, Jaenicke R (1992) Quantitative analysis of protein far UV circular dichroism spectra by neural networks. *Protein Eng* 5(3):191–195.
26. Toumadje A, Alcorn SW, Johnson WC, Jr (1992) Extending CD spectra of proteins to 168 nm improves the analysis for secondary structures. *Anal Biochem* 200(2): 321–331.
27. Manavalan P, Johnson WC, Jr (1987) Variable selection method improves the prediction of protein secondary structure from circular dichroism spectra. *Anal Biochem* 167(1):76–85.
28. Perczel A, Park K, Fasman GD (1992) Analysis of the circular dichroism spectrum of proteins using the convex constraint algorithm: A practical guide. *Anal Biochem* 203(1):83–93.
29. Perez-Iratxeta C, Andrade-Navarro MA (2008) K2D2: Estimation of protein secondary structure from circular dichroism spectra. *BMC Struct Biol* 8:25.
30. Louis-Jeune C, Andrade-Navarro MA, Perez-Iratxeta C (2012) Prediction of protein secondary structure from circular dichroism using theoretically derived spectra. *Proteins* 80(2):374–381.
31. Wiedemann C, Bellstedt P, Görlach M (2013) CAPITO—a web server-based analysis and plotting tool for circular dichroism data. *Bioinformatics* 29(14):1750–1757.
32. Venyaminov SYu, Vassilenko KS (1994) Determination of protein tertiary structure class from circular dichroism spectra. *Anal Biochem* 222(1):176–184.
33. Sawaya MR, et al. (2007) Atomic structures of amyloid cross-beta spines reveal varied steric zippers. *Nature* 447(7143):453–457.
34. Sreerama N, Woody RW (2004) Computation and analysis of protein circular dichroism spectra. *Methods Enzymol* 383:318–351.
35. Bulheller BM, Miles AJ, Wallace BA, Hirst JD (2008) Charge-transfer transitions in the vacuum-ultraviolet of protein circular dichroism spectra. *J Phys Chem B* 112(6): 1866–1874.
36. Gilbert ATB, Hirst JD (2004) Charge-transfer transitions in protein circular dichroism spectra. *Theochem* 675(1-3):53–60.
37. Woody RW (2015) The development and current state of protein circular dichroism. *Biomed Spectrosc Imaging* 4(1):5–34.
38. Hirst JD, Bhattacharjee S, Onufriev AV (2003) Theoretical studies of time-resolved spectroscopy of protein folding. *Faraday Discuss* 122:253–267, discussion 269–282.
39. Wallace BA, Janes RW (2001) Synchrotron radiation circular dichroism spectroscopy of proteins: Secondary structure, fold recognition and structural genomics. *Curr Opin Chem Biol* 5(5):567–571.
40. Wallace BA, et al. (2004) Biomedical applications of synchrotron radiation circular dichroism spectroscopy: Identification of mutant proteins associated with disease and development of a reference database for fold motifs. *Faraday Discuss* 126:237–243, discussion 245–254.