BIOINFORMATICS ORIGINAL PAPER
Vol. 29 no. 14 2013, pages 1750–1757
doi:10.1093/bioinformatics/btt278

Structural bioinformatics
Advance Access publication May 15, 2013

CAPITO—a web server-based analysis and plotting tool for circular dichroism data

Christoph Wiedemann*, Peter Bellstedt and Matthias Görlach*

Biomolecular NMR Spectroscopy, Leibniz Institute for Age Research—Fritz Lipmann Institute, Beutenbergstr. 11, 07745 Jena, Germany

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Circular dichroism (CD) spectroscopy is one of the most versatile tools to study protein folding and to validate the proper fold of purified proteins. Here, we aim to provide a readily accessible, user-friendly and platform-independent tool capable of analysing multiple CD datasets of virtually any format and returning results as high-quality graphical output to the user.

Results: CAPITO (CD Anaylsis and Plotting Tool) is a novel web server-based tool for analysing and plotting CD data. It allows reliable estimation of secondary structure content utilizing different approaches. CAPITO accepts multiple CD datasets and, hence, is well suited for a wide application range such as the analysis of temperature or pH-dependent (un)folding and the comparison of mutants.

Availability: http://capito.nmr.fli-leibniz.de.

Contact: cwiede@fli-leibniz.de or mago@fli-leibniz.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on December 20, 2012; revised on May 10, 2013; accepted on May 13, 2013

1 INTRODUCTION

The past 20 years have witnessed a dramatic growth of the number of high-resolution protein structures deposited in the protein data bank (PDB: Berman *et al*., 2000). The progress in structural biology has been driven by developments in recombinant protein expression technology, as well as by advances in methodology, data analysis and bioinformatics. *Escherichia coli* is, so far, the most widely used host for structural studies, which in turn require significant amounts of recombinant protein. Before resource-intensive detailed structural and functional studies, it is extremely helpful, if not essential, to validate the proper fold of purified recombinant proteins and one of the most versatile tools to study protein fold(ing) constitutes circular dichroism (CD) spectroscopy. Compared with X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy, the structural information obtained from CD is limited. However, CD spectroscopy carries a number of advantages: it is a well-established label-free technique requiring comparably small amounts of material, and a short time is necessary for assessing structural parameters of proteins like secondary structure, conformational changes, (un)folding and interactions (Whitmore *et al*., 2010).

A broad range of mathematical methods have been devised to extract structural information from CD spectra to provide for an estimate of the secondary structure composition of proteins via multilinear regression (Greenfield and Fasman, 1969), singular value decomposition (Hennessey and Johnson, 1981), ridge regression (Provencher and Glöckner, 1981), principal component factor analysis (Pribiċ, 1994), convex constraint analysis (Perczel *et al*., 1991), neural network-based analysis (Andrade *et al*., 1993; Böhm *et al*., 1992) and the self consistent method (Sreerama and Woody, 1993), respectively. All these methods are based on the assumption that the CD spectrum of a given protein represents a linear combination of basis spectra. Different secondary structural elements give rise to bands characteristic in wavelength and intensity (Raussens *et al*., 2003).

$$[\Theta]_\lambda = \sum f_n S_{\lambda n} + noise \qquad (1)$$

The CD spectrum of a given protein can be represented by the molar ellipticity $[\Theta]_\lambda$ as a function of wavelength λ, where f_n is the fraction of each secondary structure n, and $S_{\lambda n}$ is the ellipticity at each wavelength of each *nth* secondary structural element (Greenfield, 2006). The sum of all fractional weights $\sum f_n$ is equal to 1 in constrained fits.

The quality of the output of the aforementioned methods relies on the availability of a reference database of CD spectra of proteins whose 3D structure is known (Lees *et al*., 2006). With the advent of the Protein Circular Dichroism Data Bank (PCDDB), a public repository for far ultraviolet (far-UV) and synchrotron radiation CD spectral data and their associated experimental metadata, the number of publicly available CD spectra increased enormously (Lees *et al*., 2006; Wallace *et al*., 2006; Whitmore *et al*., 2011). In the PCDDB, each entry contains sequence and experimental information for the respective protein and includes the PDB code for proteins of which 3D structures are available.

During the recent past, different web services (Louis Jeune *et al*., 2012; Raussens *et al*., 2003; Whitmore and Wallace, 2004, 2008) or programmes (Böhm *et al*., 1992; Johnson, 1999) for analysing of CD data and estimating secondary structure content became available. Recently, Janes and co-workers (Klose *et al*., 2012) launched the tool DichroMatch for matching spectra against reference data. Here, we describe a novel web server-based tool combining different methods for estimating secondary structure content and analysing far-UV CD data based on a selected set of far-UV CD data as available from the PCDDB.

*To whom correspondence should be addressed.

2 METHODS

2.1 CD Data collection and processing

Lysozyme (chicken), cytochrome C (horse), β-amylase (sweet potato) and carbonic anhydrase II (bovine) were purchased from Sigma Aldrich at the highest purity available. Recombinant ubiquitin (human) was provided by Thomas Seiboth (FLI Jena) and the β1 immunoglobulin-binding domain of protein G (GB1) was expressed and purified as described (Bellstedt *et al*., 2012). All proteins were dissolved in 50 mM borate (cytochrome C, β-amylase and carbonic anhydrase II), pH 7.5, or exchanged into pure water (lysozyme, ubiquitin and GB1) using NAP-5 columns (GE Healthcare). The protein concentrations were in the 10 μM range and verified spectrophotometrically at 280 nm with extinction coefficients calculated using ProtParam (http://web.expasy.org/protparam/). CD spectra were collected on a JASCO J-710 CD spectropolarimeter at 4°C in a 1 mm quartz cuvette. The instrument was calibrated with D-10-camphorsulphonic acid. Each CD spectrum represents the average of 10 accumulated scans at 100 nm/min with a 1 nm slit width and a time constant of 1 s for a nominal resolution of 1.7 nm. Data were collected between 185 and 260 nm with the appropriate buffer and solvent background subtraction. No further zeroing was applied (after background subtraction) because none of the six proteins we used for the experimental part exhibited, for our chosen 10 μM concentration range, a CD between 260 and 320 nm—as tested in preliminary experiments. We tested a whole range of scan rates and time constants and did not notice significant changes in CD and the outcome of the CAPITO analysis.

2.2 Reference datasets

For this study, we used the PCDDB dataset of October 2012 as a well-calibrated, wide wavelength range reference dataset containing a large number of proteins, which effectively cover a large combination of secondary structures and fold space (Lees *et al*., 2006). That database does not include structures of oligopeptides. From this dataset, only entries linked to an existing PDB code were selected. For multiple entries referring to the same PDB code, only the spectral data recorded at the lowest temperature were used. Our selected dataset contains 107 entries (Supplementary Table S1). Note that for each PCDDB entry, the values for α, 3_{10} and π-helix are summarized as helical (h), β-strand (b) also includes β-bridge and bonded turn, bend, loop and irregular are combined as irregular (i), respectively.

In addition, as reference for significantly unfolded and pre-molten globule states, 95 datasets containing the CD values for $\lambda = 200$ and 222 nm were used as published (Uversky, 2002).

2.3 CAPITO input

Spectral data in millidegrees or mean residue CD extinction coefficient ($\Delta\epsilon$) or mean residue ellipticity ([Θ]), respectively, can be submitted in different data formats as text (txt) file: AVIV 60DS, Aviv, Aviv CDS, BP (Wallace and Teeters, 1987; Whitmore and Wallace, 2004) and Jasco. Example files are available through the CAPITO web page. The user also has the possibility to manually enter or copy/paste spectral data, where wavelength and CD data are separated by a blank or a tab stop with one wavelength per row. In addition, it is possible to upload CD data collected with either smaller or larger step size than 1 nm. The default input data dimension is in millidegrees. Following input of additional experimental parameters such as protein concentration, cuvette pathlength and the number of amino acids, millidegrees are converted to either mean residue CD extinction coefficient ($\Delta\epsilon$ in M^{-1} cm^{-1}) or mean residue ellipticity ([Θ] in deg cm^2 $dmol^{-1}$). Optionally, the amino acid sequence can be submitted as one-letter code for prediction of secondary structure using an implemented Chou-Fasman-algorithm (Chou and Fasman, 1978).

2.4 CAPITO output

CAPITO provides for the spectral data converted into either $\Delta\epsilon$ or [Θ] as a graph (for review see Greenfield, 2006; Kelly *et al*., 2005; Sreerama and Woody, 2004). In addition, the spectral values at 200 versus 222 nm are plotted for an estimate of the folding state of the protein in question. The prediction of the secondary structure elements is realised via extraction of information from a calculated set of basis spectra and a matching-based approach as described later in the text. Of all CD spectra in our reference dataset, the three curves best matching the submitted query are plotted as well. All graphs can be downloaded as high-quality portable network graphic (png) files.

3 RESULTS AND DISCUSSION

One of the most widely used applications of CD is the estimation of protein secondary structure content from far-UV CD spectra. Not only the relative proportion of secondary structure (e.g. helical, β-strand and others) provides for a characteristic contribution to the far-UV CD spectrum of a protein but also aromatic and sulphur-containing side chains, the length of α-helices and the twist in β-sheets (Johnson, 1999). A large reference dataset is necessary to cover all these features for analysis. In principle, the number of CD spectra in a reference dataset defines the number of structural features that can be determined. Based on considerations of Hennessey and Johnson (1981) and Johnson (1992, 1999), the number of different secondary structural elements significantly depends on the shortest wavelength used in a CD spectrum. For example, a lower spectral limit set to 190 nm reduces information content so that three to four different structural elements can be safely predicted. As using lower wavelengths might be impractical, in particular for biochemists, we restrict the evaluation of the CD data entered into CAPITO to three structural elements: the combination of α, 3_{10} and π-helix as helical content (h), β-strand (b) also includes β-bridge and bonded turn, bend and loop are included in the structural feature irregular (i), respectively.

3.1 Reference dataset derived basis spectra

The optical activity of individual secondary structural elements is assumed to be additive and can be expressed as given in Equation (2). At any particular wavelength λ, the sum of $f's$ is equal to 1 and all $f's \geq 0$ for a constrained approach.

$$[\Theta]_{\lambda} = f_h[\Theta]_{h,\lambda} + f_b[\Theta]_{b,\lambda} + f_i[\Theta]_{i,\lambda} \tag{2}$$

If the relative proportion for the secondary structural elements is known (from an X-ray or NMR spectroscopy-based protein structure) and the corresponding CD spectrum is at hand, it is possible to calculate the ellipticity for any given wavelength within the range of the CD spectrum. The least-square method was used for solving the $f's$ from a system of equations for our reference dataset. Solving the matrix [Equation (3)] by least-square fitting for each selected protein j in the reference dataset, a calculated $[\Theta]_{h,b,i}$ for each secondary structure element, is returned over the wavelength range of 180–240 nm (Supplementary Table S2). [Θ] for each secondary structure element is plotted in Figure 1 against the wavelength.

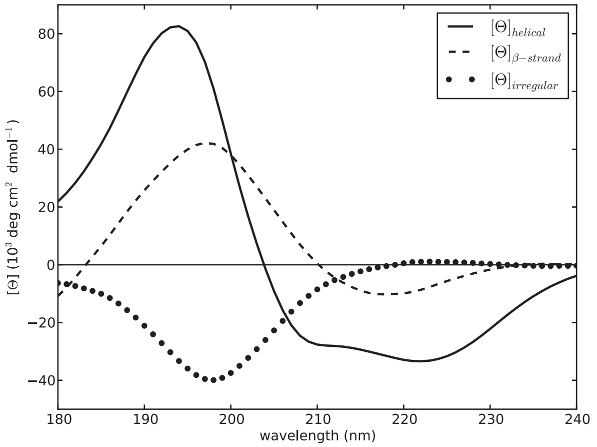

Fig. 1. Calculated CD spectra for a content of 100% helix ($[\Theta]_{helical}$), β-strand ($[\Theta]_{\beta-strand}$) and irregular ($[\Theta]_{irregular}$), respectively, based on the solution of the matrix [Equation (3)] and using all 107 proteins in the reference dataset (Supplementary Table S1)

$$\begin{pmatrix} [\Theta]_{180} \\ [\Theta]_{181} \\ [\Theta]_{182} \\ . \\ . \\ . \\ [\Theta]_{240} \end{pmatrix}_{[j,n]} = \begin{pmatrix} [\Theta]_h \\ [\Theta]_b \\ [\Theta]_i \end{pmatrix}_{\lambda} \times \begin{pmatrix} f_{h,j} & f_{b,j} & f_{i,j} \\ f_{h,j+1} & f_{b,j+1} & f_{i,j+1} \\ f_{h,j+2} & f_{b,j+2} & f_{i,j+2} \\ & . & \\ & . & \\ & . & \\ f_{h,n} & f_{b,n} & f_{i,n} \end{pmatrix} \quad (3)$$

With this approach, we derived three basis CD spectra (helical, β-strand and irregular) using our reference dataset. For the helical basis spectrum, the presence of one positive band at 194 nm and two major negative bands at 210 and 222 nm are most evident. This is consistent with basis CD spectra derived by others (Brahms and Brahms, 1980; Chen and Yang, 1971; Chen *et al.*, 1974, 1972; Hennessey and Johnson, 1981; Perczel *et al.*, 1992b; Reed and Reed, 1997; Sreerama and Woody, 1993; Toumadje *et al.*, 1992). The β-strand basis spectrum is characterized by a less intensive positive band at 197 nm and a negative band at 217 nm. A negative band at 197 nm and a weak positive band at 223 nm are features of the irregular basis spectrum. The irregular structure closely resembles that of polypeptides in extended poly-L-proline II-like structures (Woody, 1992). Also the irregular structure fits well the CD curve for poly(Pro-Lys-Leu-Lys-Leu)$_n$ in salt-free solution as model compound for unordered conformation (Brahms and Brahms, 1980). A plot overlay of different basis spectra in comparison with those calculated in this work are given in the Supplementary Data (Supplementary Figs S1–S3).

In a simple model, the ellipticity of one secondary structure content is related to certain wavelengths. For identifying wavelengths best representing secondary structure content, a constrained approach was performed. With the three basis spectra ($[\Theta]_{h,b,i}$) as derived from solving the aforementioned matrix [Equation (3)], every possible combination of f's [Equation (4)] with the boundary condition [Equation (5)] was calculated using Equation (2).

$$f_{h,b,i} = 0.00, 0.01, 0.02, \ldots, 1.00 \quad (4)$$

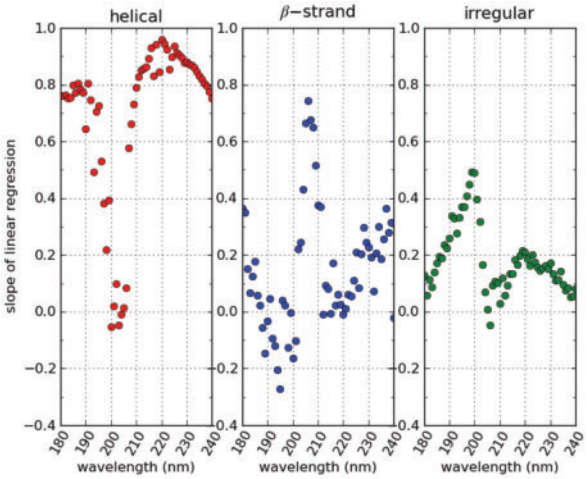

Fig. 2. For all of the proteins in the reference dataset, the secondary structure content was calculated, and the resulting slope of the linear regression of the auto-correlation—as a measure of accuracy of the prediction—is plotted against the respective wavelength

$$f_h + f_b + f_i = 1 \quad (5)$$

Subsequently, the minimum of the square difference between the calculated $[\Theta]_{cal}$ and the observed molar ellipticity $[\Theta]_{obs}$ in the reference dataset was determined. This was carried out for all proteins of the reference dataset for each wavelength in the range of 180–240 nm. For each wavelength, the best matching combination of f's was plotted against the experimental f values to assess the correlation between calculated and experimental data points, and a linear regression analysis was performed (Figs 2 and 3).

For a perfect correlation, the actual slope of the linear regression of the auto-correlation would be one, and the closer the slope is approaching unity, the higher the quality of the prediction of secondary structure elements for a given wavelength.

The results of the linear regression analysis are plotted against the wavelength. Maxima for the prediction of secondary structure contents are found for helical at 220 nm, for β-strand at 206 nm and for irregular at 199 nm, respectively (Fig. 2). To extract information about the secondary structure content from a CD curve, the three basis spectra ($[\Theta]_{h,b,i}$) as derived from matrix [Equation (3)] were used in Equation (2) at defined wavelengths as follows. For prediction of the helical content, the calculated molar ellipticity of a pure helix $[\Theta]_h$ at 220 nm was used. By varying the f_h, f_b and f_i [Equation (4)], with the constraint that the sum of the f's is equal to one [Equation (5)], the difference between the calculated and the observed $[\Theta]_{220nm}$ of a query protein was minimized. The f_h resulting from this solution represents the helical content of the respective query protein. The whole procedure was then independently repeated at 206 nm, and the resulting f_b at this wavelength is considered as the predicted β-strand content. The irregular content (f_i) was calculated at 199 nm in the same manner. In summary, for the prediction of the content of each of the three secondary structure elements, Equation (2) is used. Based on the maxima found in the linear regression analysis aforementioned (Fig. 2), as a measure for the accuracy of prediction, the programme extracts helical content at

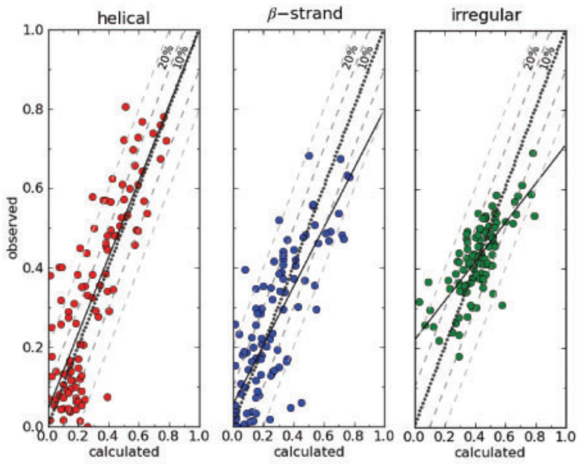

Fig. 3. Cross-validation of the basis spectra method: the secondary structure content for each of the proteins in the reference dataset was calculated and plotted against the secondary structure content available in the PCDDB. Helical content was obtained at 220 nm, β-strand at 206 nm and irregular at 199 nm. The hatched line indicates 10 and 20% deviation from the ideal auto-correlation (dotted line). Solid line represents the linear regression of the calculated auto-correlation

220 nm, β-strand at 206 nm and irregular at 199 nm only (neglecting the respective contributions of the other two elements at those wavelengths). Hence, the sum for the three secondary structural elements can differ from 100%. Based on the calculated secondary structural content, optionally a theoretical CD curve can be back calculated and plotted against the query curve, and the goodness of fit between the two curves is presented as normalized root-mean-square deviation (NRMSD) (Supplementary Equation S1) to the user, where an ideal fit would approach an NRMSD of zero (Mao *et al.*, 1982). To evaluate the quality of the model, a cross-validation was carried out. For each reference protein in the dataset, the secondary structure content was calculated as described earlier in the text and compared with deposited data within the PCDDB. This prediction was repeated for all spectra within our reference dataset. The calculated secondary structure contents are plotted against the actual values (Fig. 3). From this, it becomes evident that this procedure delivers a reliable estimate of helical, β-strand and irregular secondary structure content for a given protein.

As illustrated in Figure 3, the cross-validation provides the most accurate results for helical secondary structure. Only 7 of 107 validated proteins are beyond the margin of 20% deviation. The prediction of β-strand secondary structure is somewhat less accurate than for helical content and shows that eight proteins of our reference dataset are not within the range of 20% deviation. Irregular structures (neither helix nor β-strand) adopt a wide range of backbone angles exhibiting CD spectra of heterogeneous character. This hampers the accurate estimation of such secondary structure elements. With our approach, only four proteins of our reference dataset are not in the range of 20% deviation for the prediction of irregular structures. The irregular structure content in our reference data is not spread over the range observed for helical or β-strand content. In such a situation, outliers have a more significant influence on the linear regression, resulting in a slope further away from approaching unity.

This approach appears somewhat less accurate in comparison with other methods [SELCON3 (Sreerama and Woody, 1993), CDSSTR (Johnson, 1999) or CONTIN (Provencher and Glöckner, 1981), Table 1], but the results and their validation strongly suggest that the input of measured values at three wavelengths (220 nm for helical, 206 nm for β-strand and 199 nm for irregular) are sufficient to describe the information contained within a given CD spectrum.

3.2 Matching-based prediction: nearest-neighbour and area difference method

Under the general assumption that proteins with similar secondary structure content give rise to comparable CD curves, we tested alternative prediction methods. If the reference dataset effectively covers a large combination of secondary structures and fold space, matched reference proteins should comprise the actual secondary structure content of a query protein. To test this hypothesis, we used two methods to evaluate a query CD spectrum against our reference dataset. For either method, the query CD spectrum is compared with each CD spectrum of the reference dataset. As standard pattern recognition method, we used a k-nearest-neighbour algorithm (Cover, 1968; Kowalski and Bender, 1972) in the following two approaches.

For the nearest-neighbour approach, for each wavelength within the range of 180–240 nm, the 25 best matching reference curves defined by closest proximity were determined. Here, proximity is defined as the 1D distance between the query and the reference CD spectra at each wavelength. Subsequently, the frequency (N) of a given reference protein among the 25 nearest neighbours to the query protein is used for ranking.

To assess the lowest area difference (AD), the best matching reference curves for the range of 180–240 nm were selected. AD was defined as shown in Equation (6).

$$\mathrm{AD}_i = \sum_{\lambda} \sqrt{\left(([\Theta]_{Q,\lambda} - [\Theta]_{Ref,\lambda,i}) \times 1\,nm\right)^2} \qquad (6)$$

AD (in deg cm^2 dmol^{-1} nm) represents the CD curve area difference between query (Q) and the reference (Ref) protein, where $[\Theta]_{Q,\lambda}$ is the molar ellipticity of the query protein at wavelength λ, and $[\Theta]_{Ref,\lambda}$ is the molar ellipticity of the protein i in the reference dataset at the same wavelength λ, respectively. In extension of the rotational strength approach presented by Klose *et al.* (2012), the AD approach here evaluates in 1 nm steps over the wavelength range λ the difference in area between a reference and a query curve rather than the area of a given CD curve per se. It, hence, also includes an (although indirect) evaluation of the shape of the query CD curve. The AD output provides the best ranked CD reference curves with the smallest area difference. The NRMSD (Supplementary Equation S1) between the best matching reference CD spectra and the query CD spectra is calculated and shown for comparison. AD and NRMSD are always calculated over the same wavelength range as provided by the query dataset. However, the NRMSD is neither used for matching nor for ranking of identified hits.

To validate our approach for extracting structural information by matching query and reference CD spectra an auto-correlation

Table 1. Estimation of structural contents in comparison

Protein	SELCON3	CDSSTR	CONTINLL	CDNN2.1[a]	Raussens[b]	K2D3[c]	CAPITO[d]	PDB[e]
Ubiquitin								
h	15,5	17,6	26,3	14,3	29,7	10,1	11 (4–25)	25
b	27,6	28,1	20	42,9	16,3	28,6	30 (34–45)	34
i	52,8	53,7	53,7	48,2	47,6	61,3	52 (41–54)	41
Lysozyme								
h	41,4	41,8	41,9	34	30,9	32,6	25 (31–50)	40
b	9,1	10,4	8,9	12,9	16,3	17,5	14 (4–21)	10
i	49,4	47,4	49,2	55,8	46,8	49,3	51 (46–49)	50
Cytochrome C								
h	41,3	43,5	44	31,6	24,7	26,3	31 (13–48)	40
b	5,2	10,1	3,1	12,7	19,2	23,4	25 (2–34)	1
i	50,6	47,8	52,9	58,2	46,3	50,3	46 (49–53)	59
β-Amylase								
h	36,3	40,9	36,6	31,8	25	29,4	28 (31–50)	38
b	14,1	12,3	11,3	12,1	22,7	17,9	15 (4–21)	13
i	51,8	47,2	52	62,3	47,5	52,7	53 (46–48)	49
CA-II								
h	9,6	9,6	7,6	10,3	0,9	2,5	0 (12–16)	15
b	35,8	41,8	32,3	32,6	34,6	37,9	37 (30–37)	30
i	55,9	52,4	60	54,9	51,5	59,6	52 (48–54)	55
GB1								
h	42,9	45,4	42,5	40,3	40	34,1	39 (29–44)	25
b	13,6	13,5	15,2	15,4	13,4	19,8	13 (12–30)	42
i	43,1	41,3	42,3	45	42,7	46,1	37 (39–54)	33

Note: The CD spectra of the indicated proteins were recorded and processed. The resulting data are analysed with different programmes. Helical content is represented as h, β-strand content as b and irregular as i. SELCON3 (Sreerama and Woody, 1993), CDSSTR (Johnson, 1999) and CONTINLL (variant of CONTIN (Provencher and Glöckner, 1981) are provided in the CDPro software package (Sreerama and Woody, 2000).

[a]Böhm *et al.* (1992), [b]Raussens *et al.* (2003), [c]Louis Jeune *et al.* (2012). [d]For CAPITO, the result on the basis of spectra-based method is given. In parenthesis, the range of the three best hits is provided based on the area difference method. [e]Secondary structure content obtained through the PDB website using the PDB ID 1UBI (ubiquitin), 193L (lysozyme), 1HRC (cytochrome C), 1FA2 (β-amylase), 1V9E (carbonic angydrsae II, CA-II) and 2LGI (β1 immunoglobulin-binding domain of protein G, GB1), respectively. For K2D3 and PDB, the helical and β-strand contents were subtracted from 100 to calculate the irregular content. All values are given as percentage. *Of note*: SELCON, CDSSTR, CONTINLL and K2D2 are used by DichroWeb (Whitmore and Wallace, 2004). Here, we have used the most recent versions of these packages for comparing their results with the results returned by CAPITO.

was performed. For this purpose, single-reference spectra were removed in turn from the reference dataset, and matching was performed with the remaining spectra against the structure content as derived from the removed spectrum. The structural contents of the three best matching reference proteins are combined to give a range for f_h, f_b and f_i. In this approach, the secondary structure content of a protein in question is not defined by single calculated value but by a range of secondary structural content derived from the three best matching reference proteins. This prediction was repeated for all the spectra within the reference dataset.

Figure 4 depicts the cross-validation for the structural content predicted by the nearest-neighbour method (a) and by the area difference method (b), respectively. Both methods deliver a reliable and comparable estimate of helical, β-strand and irregular content for a given protein. Only a few outliers in the prediction of helical and β-strand content are not within the margin of 20% deviation. Surprisingly, for irregular content, all validated proteins fall within the boundary of 20% deviation. The range of secondary structural content present in the three best matches covers in most cases the actual content of the query protein. A consequence of this is that an increasing number of protein structures in the reference database will further restrict the range covered by the three best matches and, hence, lead to an increased accuracy of the secondary structure content for a given protein as predicted by CAPITO.

3.3 Validation by protein identification and comparison with other programmes

Whitmore *et al.* (2011) mentioned the idea of identifying proteins based on their CD spectral characteristics, i.e. if a protein is deposited in a database, it should be possible to identify this protein based on its CD curve. Here, we recorded CD spectra of freshly prepared solutions of six different proteins (see Section 2) and processed the recorded CD data with the CAPITO web server. Our reference dataset contains five (β-amylase, carbonic anhydrase II, cytochrome C, lysozyme and ubiquitin) of six of the selected proteins used here. As seen in Figure 5a (and also Supplementary Fig. S4), our matching-based prediction allows for the identification of proteins present in a reference dataset—even under buffer conditions slightly different from the ones in the reference dataset. All tested proteins were identified as indicated by the best score (Fig. 5a and Supplementary Fig. S4).

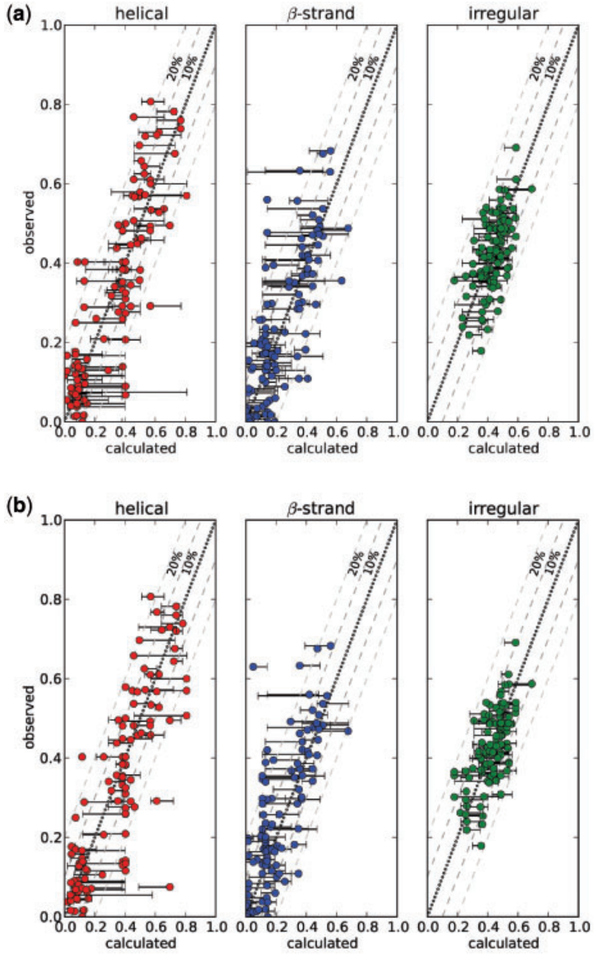

Fig. 4. Cross-validation of the nearest-neighbour method (**a**) and area difference method (**b**). The secondary structure content for each of the proteins in the reference dataset was calculated and plotted against its actual secondary structure content. The top-ranked hit is plotted as dot, error bars connect second- and third-ranked hit. The hatched line indicates 10 and 20% deviation from the ideal auto-correlation (dotted line)

To compare our method of secondary structure prediction, we used our recorded and processed spectra to a selection of other available programmes and web services (Böhm *et al.*, 1992; Louis Jeune *et al.*, 2012; Raussens *et al.*, 2003; Sreerama and Woody, 2000). The results (Table 1) were compared with secondary structure assignments deposited in the Protein Data Bank, which in turn are based on the DSSP programme (Kabsch and Sander, 1983). A standard set of reference CD data was tested to compare the accuracy of CAPITO with other available programmes. The standard set consisted of 16 proteins and poly-L-glutamic acid (Sreerama and Woody, 1993). CAPITO returns a good correlation coefficient for the helical and *β*-strand secondary structure elements found in the X-ray structure of the 16 test proteins (Table 2). As most programmes that rely on a protein dataset, CAPITO is not suitable for evaluating helical and *β*-strand conformation content for long homopolymeric peptides such as poly-L-lysine, as our reference dataset does not include such homopolymers.

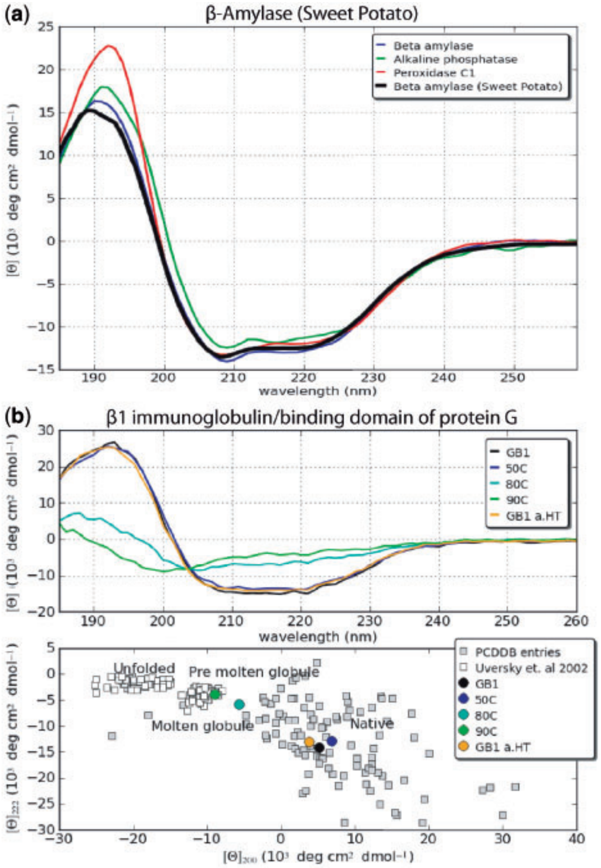

Fig. 5. Experimental validation: (**a**) a CD spectrum of *β*-amylase (sweet potato) was recorded and analysed with CAPITO (black curve). As an example, the graphical output for the area difference method is shown. The best-matching CD spectra of the reference dataset are *β*-amylase (blue), the second best alkaline phosphatase (green) and the third best hit peroxidase C1 (red), respectively. (**b**) Recorded CD spectra for GB1 at different temperatures. Upper panel: CD curve for the starting temperature of 4°C is depicted in black, other temperatures as indicated in the inset. CD spectrum of GB1 after heat treatment and cooled down to 4°C (GB1 a.HT) is shown in orange. Lower panel: CD values at $\lambda = 200$ nm plotted versus the values at $\lambda = 222$ nm to deduce the folding state of GB1 at the respective temperatures

3.4 Monitoring protein folding under different conditions

Different folding states of polypeptide chains are characterized by specific shapes of their far-UV CD spectrum. For example, unfolded polypeptides or proteins containing mainly irregular structural elements show a spectral minimum in the vicinity of 200 nm and an ellipticity close to zero in the vicinity of 222 nm. Helical proteins show a characteristic double minimum at 222 and 208 nm and an intensive maximum near 195 nm (Fig. 1). The 'double wavelength' plot ($[\Theta]_{222}$ versus $[\Theta]_{200}$) as described (Uversky, 2002) allows the direct visualization of the folding state of a protein for different conditions (e.g. temperature, pH, buffer type and ionic strength). Here, we use this plotting routine within CAPITO to carry out an assessment of the GB1 folding state as a function of temperature. As shown in Figure 5b (lower panel) from the double wavelength plot, it can be concluded that GB1 shows a well folded state at lower temperatures,

Table 2. Comparisons of methods of analysing protein secondary structure content from CD data

Program	Standards	Wavelength	Helical		β-Strand		Irregular	
			P	σ	P	σ	P	σ
Linear regression—unconstrained								
MLR[a]	4 peptides	178–240	0.91	0.13	0.43	0.21	0.07	0.16
MLR	4 peptides	200–240	0.92	0.14	0.74	0.16	0.23	0.16
Linear regression—constrained								
G&F[b]	poly-L-lysine	208–240	0.92	0.13	0.61	0.18	ND	ND
LINCOMB[c]	4 peptides	178–240	0.93	0.11	0.58	0.15	0.61	0.11
LINCOMB	17 proteins	178–240	0.94	0.09	0.62	0.14	0.21	0.13
Singular value decomposition								
SVD[d]	17 proteins	178–240	0.98	0.05	0.68	0.12	0.22	0.10
Convex constraint algorithm								
CCA[e]	17 proteins	178–260	0.96	0.10	0.62	0.18	0.39	0.18
Ridge regression								
CONTIN[f]	17 proteins	178–260	0.93	0.11	0.56	0.15	0.58	0.08
Variable selection								
VARSLC[g]	17 proteins	178–260	0.97	0.07	0.81	0.10	0.60	0.07
Variable selection—self-consistent method								
SELCON[h]	17 proteins	178–260	0.95	0.09	0.84	0.08	0.77	0.05
SELCON	33 proteins	178–260	0.93	0.09	0.91	0.07	0.53	0.09
Neural network analysis								
CDNN2.1[i]	17 proteins	178–260	0.93	0.10	0.73	0.11	0.82	0.05
K2D2[j]	19 proteins	200–240	0.95	0.09	0.77	0.10	ND	ND
This work								
CAPITO	107 proteins	178–260	0.96	0.11	0.80	0.13	ND	ND

Note: Table 2 lists the correlation coefficient (P) and the mean-square errors (σ) between the calculated and the observed contents of each secondary structure. The table was abstracted from Greenfield (1996).

[a]Perczel *et al.* (1992a), [b]Greenfield and Fasman (1969), [c]Perczel *et al.* (1992a), [d]Hennessey and Johnson (1981), [e]Perczel *et al.* (1991), [f]Provencher and Glöckner (1981), [g]Manavalan and Johnson (1987), [h]Sreerama and Woody (1993), [i]Böhm *et al.*, (1992) and [j]Andrade *et al.* (1993).

which is stable even at higher temperatures. At 80°C, a transition from the native folded state to the molten globule is observed. The temperature shift from 80°C to 90°C changes the GB1 fold from a molten globule towards a pre-molten globule state. This observation is consistent with the previous determined midpoint of denaturation at ~75°C (Alexander *et al.*, 1992; Minor and Kim, 1994). Although, this does not replace a detailed analysis (e.g. melting curve), it allows for a quick and coarse estimation of a transition point enabling analysis of unstable proteins, which may not withstand a time-consuming detailed analysis.

4 CONCLUSION

We have developed CAPITO, a novel web server-based analysis tool for interpreting CD spectra. It allows the simultaneous evaluation of multiple datasets. Hence, it is suitable for the investigation of a protein in question under different conditions (temperature, pH, buffer solvent and mutations). Our approaches (basis spectra and matching-based method) to extract secondary structure information from a CD spectrum take advantage of a recent significant increase in the availability of well-calibrated far-UV CD spectra linked to available tertiary structures. The accuracy of our methods in predicting α-helical, β-strand or irregular content is reliable compared with other frequently used programmes or web services. In summary, we here provide a freely accessible, user-friendly and robust tool for the analysis of CD spectra.

ACKNOWLEDGEMENT

The authors are grateful to G. Peiter and F. Haubensak for help with the IT infrastructure.

Funding: The FLI is a member of Science Association ‘Gottfried Wilhelm Leibniz (WGL)’ and is financially supported by the Federal Government of Germany and the State of Thuringia. P.B. and C.W. are supported by the Leibniz Graduate School on Ageing and Age-Related Diseases (LGSA).

Conflict of Interest: none declared.

REFERENCES

Alexander,P. *et al.* (1992) Thermodynamic analysis of the folding of the streptococcal protein G IgG-binding domains B1 and B2: why small proteins tend to have high denaturation temperatures. *Biochemistry*, **31**, 3597–3603.

Andrade,M.A. *et al.* (1993) Evaluation of secondary structure of proteins from UV circular dichroism spectra using an unsupervised learning neural network. *Protein Eng.*, **6**, 383–390.

Bellstedt,P. *et al.* (2012) Solid state NMR of proteins at high MAS frequencies: symmetry-based mixing and simultaneous acquisition of chemical shift correlation spectra. *J. Biomol. NMR*, **54**, 325–335.

Berman,H.M. *et al.* (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.

Böhm,G. *et al.* (1992) Quantitative analysis of protein far UV circular dichroism spectra by neural networks. *Protein Eng.*, **5**, 191–195.

Brahms,S. and Brahms,J. (1980) Determination of protein secondary structure in solution by vacuum ultraviolet circular dichroism. *J. Mol. Biol.*, **138**, 149–178.

Chen,Y.H. and Yang,J.T. (1971) A new approach to the calculation of secondary structures of globular proteins by optical rotatory dispersion and circular dichroism. *Biochem. Biophys. Res. Commun.*, **44**, 1285–1291.

Chen,Y.H. *et al.* (1972) Determination of the secondary structures of proteins by circular dichroism and optical rotatory dispersion. *Biochemistry*, **11**, 4120–4131.

Chen,Y.H. *et al.* (1974) Determination of the helix and beta form of proteins in aqueous solution by circular dichroism. *Biochemistry*, **13**, 3350–3359.

Chou,P.Y. and Fasman,G.D. (1978) Empirical predictions of protein conformation. *Annu. Rev. Biochem.*, **47**, 251–276.

Cover,T.M. (1968) Estimation by the nearest neighbor rule. *IEEE Trans. Information Theory*, **14**, 50–55.

Greenfield,N. and Fasman,G.D. (1969) Computed circular dichroism spectra for the evaluation of protein conformation. *Biochemistry*, **8**, 4108–4116.

Greenfield,N.J. (1996) Methods to estimate the conformation of proteins and polypeptides from circular dichroism data. *Anal. Biochem.*, **235**, 1–10.

Greenfield,N.J. (2006) Using circular dichroism spectra to estimate protein secondary structure. *Nat. Protoc.*, **1**, 2876–2890.

Hennessey,J. Jr and Johnson,W. Jr (1981) Information content in the circular dichroism of proteins. *Biochemistry*, **20**, 1085–1094.

Johnson,W. Jr (1992) Analysis of circular dichroism spectra. *Methods Enzymol.*, **210**, 426–447.

Johnson,W.C. (1999) Analyzing protein circular dichroism spectra for accurate secondary structures. *Proteins*, **35**, 307–312.

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

Kelly,S.M. *et al.* (2005) How to study proteins by circular dichroism. *Biochim. Biophys. Acta*, **1751**, 119–139.

Klose,D.P. *et al.* (2012) DichroMatch: a website for similarity searching of circular dichroism spectra. *Nucleic Acids Res.*, **40**, W547–W552.

Kowalski,B.R. and Bender,C.F. (1972) The K-nearest neighbor classification rule (pattern recognition) applied to nuclear magnetic resonance spectral interpretation. *Anal. Chem.*, **44**, 14051411.

Lees,J.G. *et al.* (2006) A reference database for circular dichroism spectroscopy covering fold and secondary structure space. *Bioinformatics*, **22**, 1955–1962.

Louis Jeune,C. *et al.* (2012) Prediction of protein secondary structure from circular dichroism using theoretically derived spectra. *Proteins*, **80**, 374–381.

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

Mao,D. *et al.* (1982) Folding of the mitochondrial proton adenosinetriphosphatase proteolipid channel in phospholipid vesicles. *Biochemistry*, **21**, 4960–4968.

Minor,D. Jr and Kim,P.S. (1994) Measurement of the beta-sheet-forming propensities of amino acids. *Nature*, **367**, 660–663.

Perczel,A. *et al.* (1991) Convex constraint analysis: a natural deconvolution of circular dichroism curves of proteins. *Protein Eng.*, **4**, 669–679.

Perczel,A. *et al.* (1992a) Analysis of the circular dichroism spectrum of proteins using the convex constraint algorithm: a practical guide. *Anal. Biochem.*, **203**, 83–93.

Perczel,A. *et al.* (1992b) Deconvolution of the circular dichroism spectra of proteins: the circular dichroism spectra of the antiparallel beta-sheet in proteins. *Proteins*, **13**, 57–69.

Pribicĉ,R. (1994) Principal component analysis of Fourier transform infrared and/or circular dichroism spectra of proteins applied in a calibration of protein secondary structure. *Anal. Biochem.*, **223**, 26–34.

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

Raussens,V. *et al.* (2003) Protein concentration is not an absolute prerequisite for the determination of secondary structure from circular dichroism spectra: a new scaling method. *Anal. Biochem.*, **319**, 114–121.

Reed,J. and Reed,T.A. (1997) A set of constructed type spectra for the practical estimation of peptide secondary structure from circular dichroism. *Anal. Biochem.*, **254**, 36–40.

Sreerama,N. and Woody,R.W. (1993) A self-consistent method for the analysis of protein secondary structure from circular dichroism. *Anal. Biochem.*, **209**, 32–44.

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.

Sreerama,N. and Woody,R.W. (2004) Computation and analysis of protein circular dichroism spectra. *Methods Enzymol.*, **383**, 318–351.

Toumadje,A. *et al.* (1992) Extending CD spectra of proteins to 168 nm improves the analysis for secondary structures. *Anal. Biochem.*, **200**, 321–331.

Uversky,V.N. (2002) Natively unfolded proteins: a point where biology waits for physics. *Protein Sci.*, **11**, 739–756.

Wallace,B.A. and Teeters,C.L. (1987) Differential absorption flattening optical effects are significant in the circular dichroism spectra of large membrane fragments. *Biochemistry*, **26**, 65–70.

Wallace,B.A. *et al.* (2006) The protein circular dichroism data bank (PCDDB): a bioinformatics and spectroscopic resource. *Proteins*, **62**, 1–3.

Whitmore,L. and Wallace,B.A. (2004) DICHROWEB, an online server for protein secondary structure analyses from circular dichroism spectroscopic data. *Nucleic Acids Res.*, **32**, W668–W673.

Whitmore,L. and Wallace,B.A. (2008) Protein secondary structure analyses from circular dichroism spectroscopy: methods and reference databases. *Biopolymers*, **89**, 392–400.

Whitmore,L. *et al.* (2010) The protein circular dichroism data bank, a Web-based site for access to circular dichroism spectroscopic data. *Structure*, **18**, 1267–1269.

Whitmore,L. *et al.* (2011) PCDDB: the protein circular dichroism data bank, a repository for circular dichroism spectral and metadata. *Nucleic Acids Res.*, **39**, D480–D486.

Woody,R.W. (1992) Circular dichroism and conformation of unordered poly-peptides. *Adv. Biophys. Chem.*, **2**, 31–79.