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

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 data sets of virtually any format and returning results as high quality graphical output to the user.

Results: CAPITO (CD Analysis and Plotting Tool) is a novel web server based tool for analysing and plotting CD data. It allows reliable estimation of secondary structure content utilising different approaches. CAPITO accepts multiple CD data sets 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.

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

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 very much 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. Prior to 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 to 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 (Pribič, 1994), convex constraint analysis (Perczel et al., 1991), neural network based analysis (Böhm et al., 1992; Andrade et al., 1993) 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).

\[
[\Theta]_\lambda = \sum f_n S_{\lambda n} + noise \tag{1}
\]

The CD spectrum of a given protein can be represented by the molar ellipticity \([\Theta]_\lambda\) as a function of wavelength \(\lambda\), where \(f_n\) is the fraction of each secondary structure \(n\), and \(S_{\lambda n}\) is the ellipticity at each wavelength of each \(n^{th}\) 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 above mentioned 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 (Wallace et al., 2006; Lees 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 (Whitmore and Wallace, 2004, 2008; Raussens et al., 2003; Louis Jeune et al., 2012) 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 service based tool combining different methods for estimating secondary structure

*to whom correspondence should be addressed
content and analysing far-UV CD data based on a selected set of far-UV CD data as available from the PCDD.

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, Carbonic Anhydrase II), pH 7.5, or exchanged into pure water (Lysozyme, Ubiquitin, 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 an 1 mm quartz cuvette. The instrument was calibrated with d-10-camphorsulfonic 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 data sets

For this study we employed the PCDDB data set 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 data set 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 data set contains 107 entries (Suppl. 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 data sets containing the CD values for λ = 200 nm and 222 nm were utilized as published (Uversky, 2002).

2.3 CAPITO Input

Spectral data in millidegrees or mean residue CD extinction coefficient (Δε) 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 (Δε in M⁻¹ cm⁻¹) or mean residue ellipticity (θ in deg cm² dmol⁻¹). Optionally, the amino acid sequence can be submitted as one-letter code for prediction of secondary structure utilizing an implemented Chou-Fasman-algorithm (Chou and Fasman, 1978).

2.4 CAPITO Output

CAPITO provides for the spectral data converted into either Δε or θ as a graph (for review see Seeraram and Woody (2004); Kelly et al. (2005); Greenfield (2006)). In addition, the spectral values at 200 nm 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 below. Of all CD spectra in our reference data set 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 sulfur-containing side chains, the length of α-helices, and the twist in β-sheets (Johnson, 1999). A large reference data set is necessary to cover all these features for analysis. In principle, the number of CD spectra in a reference data set defines the number of structural features that can be determined. Based on considerations of Hennessey and Johnson (1981; 1992; 1999), the number of different secondary structural elements significantly depends upon 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 data set derived basis spectra

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

\[ \theta_{\lambda} = f_h[\theta]_{h,\lambda} + f_b[\theta]_{b,\lambda} + f_i[\theta]_{i,\lambda} \]  

(2)

If the relative proportion for the secondary structural elements is known (from a 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 employed for solving the f's from a system of equations for our reference data set. Solving the matrix (eq. 3) by least square fitting for each selected protein j in the reference data set, a calculated [θ]_{h,b,i} for each secondary structure element is returned over the wavelength range of 180 to 240 nm (Suppl. Table S2). [θ] for each secondary structure
Fig. 1. Calculated CD spectra for a content of 100% helix ([Θ]helical), β-strand ([Θ]β-strand) and irregular ([Θ]irregular), respectively, based on the solution of the matrix (eq. 3) and employing all 107 proteins in the reference data set (Suppl. Table S1).

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 (Figure 2). To extract information about the secondary structure content from a CD curve, the three basis spectra ([Θ]h,b,i) as derived from matrix (eq. 3) were used in eq. 2 at defined wavelengths as follows. For prediction of the helical content the calculated molar ellipticity of a pure helix [Θ]h at 220 nm was used. By varying the f_h, f_b and f_i (eq. 4), with the constraint that the sum of the f’s is equal to one (eq. 5), the difference between the calculated and the observed [Θ]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 eq. 2 is used. Based upon the maxima found in the linear regression analysis above (Figure 2), as a measure for the accuracy of prediction, the programme extracts helical content at 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) (Suppl. eq. 1) 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 data set, the secondary structure content was calculated as described above and compared with deposited data within the PCDDB. This prediction was repeated for all spectra within our reference data set. The calculated secondary structure contents are plotted against the actual values (Figure 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 out of 107 validated proteins are beyond the margin of 20 % deviation. The

(eq. 5) was calculated using eq. 2.

\[ \sum_{i=1}^{3} f_i = 1 \]
For all of the proteins in the reference data set 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.

Cross-validation of the basis spectra method: The secondary structure content for each of the proteins in the reference data set 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.

The prediction of β-strand secondary structure is somewhat less accurate than for helical content and shows that 8 proteins of our reference data set 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 4 proteins of our reference data set 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 to 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 & 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 data set 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 employed two methods to evaluate a query CD spectrum against our reference data set. For either method, the query CD spectrum is compared to each CD spectrum of the reference data set. 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 one-dimensional 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.

In order 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 eq. 6.

\[
AD_i = \sqrt{\frac{1}{\lambda} \sum_n \left( [\Theta]_{Q,\lambda} - [\Theta]_{Ref,i,\lambda} \right)^2}
\]

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 \(\lambda\), and \([\Theta]_{Ref,i,\lambda}\) is the molar ellipticity of the protein \(i\) in the reference data set at the same wavelength \(\lambda\), respectively. In extension of the rotational strength approach presented by Klose et al. (2012) the AD approach here evaluates in one nm steps over the wavelength range \(\lambda\) 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 (though 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 (Suppl. eq. 1) 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 data set. 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 was calculated and plotted against the respective wavelengths.
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 data base 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 to 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 Materials and Methods) and processed the recorded CD data with the CAPITO web server. Our reference data set contains 5 (β-Amylase, Carbonic Anhydrase II, Cytochrome C, Lysozyme and Ubiquitin) out of 6 of the selected proteins used here. As seen in Figure 5a (and also Suppl. Figure S4), our matching based prediction allows for the identification of proteins present in a reference data set - even under buffer conditions slightly different from the ones in the reference data set. All tested proteins were identified as indicated by the best score (Figure 5a and Suppl. Figure S4).

To compare our method of secondary structure prediction we employed our recorded and processed spectra to a selection of other available programmes and web services (Sreerama and Woody, 2000; Böhm et al., 1992; Raussens et al., 2003; Louis Jeune et al., 2012). 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 in order to compare the accuracy of CAPITO with other available programs. 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 programs that rely on a protein data set, CAPITO is not suitable for evaluating helical and β-strand conformation content for long homopolymers such as poly-L-lysine as our reference data set does not include such homopolymers.

3.4 Monitoring protein folding under different conditions

Different folding states of polypeptide chains are characterised 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 nm and 208 nm and an intensive maximum near 195 nm (Figure 1). The “double wavelength” plot (θ222 versus θ208) as described (Uversky, 2002) allows the direct visualisation of the folding state of a protein for different conditions (e.g. temperature, pH, buffer type, ionic strength). Here we utilise 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
Table 1. Estimation of structural contents in comparison

<table>
<thead>
<tr>
<th>Protein</th>
<th>SELCON3</th>
<th>CDSSTR</th>
<th>CONTINLL</th>
<th>CDNN2.1 (1)</th>
<th>Raussem (2)</th>
<th>K2D3 (3)</th>
<th>CAPITO (4)</th>
<th>PDB (5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ubiquitin</td>
<td>h: 15.5</td>
<td>17.6</td>
<td>26.3</td>
<td>14.3</td>
<td>29.7</td>
<td>10.1</td>
<td>11 (4-25)</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>b: 27.6</td>
<td>28.1</td>
<td>20</td>
<td>42.9</td>
<td>16.3</td>
<td>28.6</td>
<td>30 (34-45)</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td>i: 52.8</td>
<td>53.7</td>
<td>53.7</td>
<td>48.2</td>
<td>47.6</td>
<td>61.3</td>
<td>52 (41-54)</td>
<td>41</td>
</tr>
<tr>
<td>Lysozyme</td>
<td>h: 41.4</td>
<td>41.8</td>
<td>41.9</td>
<td>34</td>
<td>30.9</td>
<td>32.6</td>
<td>25 (31-50)</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>b: 9.1</td>
<td>10.4</td>
<td>8.9</td>
<td>12.9</td>
<td>16.3</td>
<td>17.5</td>
<td>14 (4-21)</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>i: 49.4</td>
<td>47.4</td>
<td>49.2</td>
<td>55.8</td>
<td>46.8</td>
<td>49.3</td>
<td>51 (46-49)</td>
<td>50</td>
</tr>
<tr>
<td>Cytochrome C</td>
<td>h: 36.3</td>
<td>40.9</td>
<td>36.6</td>
<td>31.8</td>
<td>25</td>
<td>29.4</td>
<td>28 (31-50)</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td>b: 14.1</td>
<td>12.3</td>
<td>11.3</td>
<td>12.1</td>
<td>22.7</td>
<td>17.9</td>
<td>15 (4-21)</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>i: 51.8</td>
<td>47.2</td>
<td>52</td>
<td>62.3</td>
<td>47.5</td>
<td>52.7</td>
<td>53 (46-48)</td>
<td>49</td>
</tr>
<tr>
<td>β-Amylase</td>
<td>h: 9.6</td>
<td>9.6</td>
<td>7.6</td>
<td>10.3</td>
<td>0.9</td>
<td>2.5</td>
<td>0 (12-16)</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>b: 35.8</td>
<td>41.8</td>
<td>32.3</td>
<td>32.6</td>
<td>34.6</td>
<td>37.9</td>
<td>37 (30-37)</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>i: 55.9</td>
<td>52.4</td>
<td>60</td>
<td>54.9</td>
<td>51.5</td>
<td>59.6</td>
<td>52 (48-54)</td>
<td>55</td>
</tr>
<tr>
<td>GB1</td>
<td>h: 42.9</td>
<td>45.4</td>
<td>42.5</td>
<td>40.3</td>
<td>40</td>
<td>34.1</td>
<td>39 (29-44)</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>b: 13.6</td>
<td>13.5</td>
<td>15.2</td>
<td>15.4</td>
<td>13.4</td>
<td>19.8</td>
<td>13 (12-30)</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td>i: 43.1</td>
<td>41.3</td>
<td>42.3</td>
<td>45</td>
<td>42.7</td>
<td>46.1</td>
<td>37 (39-54)</td>
<td>33</td>
</tr>
</tbody>
</table>

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). 1) Bohm et al. (1992), 2) Raussens et al. (2003), 3) Louis Jeune et al. (2012). 4) For CAPITO the result of the basis spectra based method is given. In parenthesis the range of the three best hits based on the area difference method. 5) Secondary structure content obtained through the PDB web site using the PDB ID 1UBI (Ubiquitin), 193L (Lysozyme), 1HRC (Cytochrome C), 1FA2 (β-Amylase), 1V9E (Carbonic Angydrase 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.

Double wavelength plot it can be concluded that GB1 shows a well folded state at lower temperatures, 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 circular dichroism spectra. It allows the simultaneous evaluation of multiple data sets. Hence, it is suitable for the investigation of a protein in question under different conditions (temperature, pH, buffer solvent, 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 to 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

PB and CW are supported by the Leibniz Graduate School on Ageing and Age-Related Diseases (LGSA). We are grateful to G. Peier and F. Haubensak for help with the IT infrastructure. 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.

REFERENCES

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

<table>
<thead>
<tr>
<th>Program</th>
<th>Standards</th>
<th>Wavelength</th>
<th>helical</th>
<th>β-strand</th>
<th>irregular</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>P</td>
<td>σ</td>
<td>P</td>
</tr>
<tr>
<td><strong>Linear Regression - unconstrained</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLR(3)</td>
<td>4 peptides</td>
<td>178-240</td>
<td>0.91</td>
<td>0.13</td>
<td>0.43</td>
</tr>
<tr>
<td>MLR</td>
<td>4 peptides</td>
<td>200-240</td>
<td>0.92</td>
<td>0.14</td>
<td>0.74</td>
</tr>
<tr>
<td><strong>Linear Regression - constrained</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G&amp;F(2)</td>
<td>poly-L-lysine</td>
<td>208-240</td>
<td>0.92</td>
<td>0.13</td>
<td>0.61</td>
</tr>
<tr>
<td>LINCOMB(3)</td>
<td>4 peptides</td>
<td>178-240</td>
<td>0.93</td>
<td>0.11</td>
<td>0.58</td>
</tr>
<tr>
<td>LINCOMB</td>
<td>17 proteins</td>
<td>178-240</td>
<td>0.94</td>
<td>0.09</td>
<td>0.62</td>
</tr>
<tr>
<td><strong>Singular value decomposition</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVD(4)</td>
<td>17 proteins</td>
<td>178-240</td>
<td>0.98</td>
<td>0.05</td>
<td>0.68</td>
</tr>
<tr>
<td><strong>Convex constraint algorithm</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CCA(5)</td>
<td>17 proteins</td>
<td>178-260</td>
<td>0.96</td>
<td>0.10</td>
<td>0.62</td>
</tr>
<tr>
<td><strong>Ridge regression</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CONTIN(6)</td>
<td>17 proteins</td>
<td>178-260</td>
<td>0.93</td>
<td>0.11</td>
<td>0.56</td>
</tr>
<tr>
<td><strong>Variable selection</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VARSLC(7)</td>
<td>17 proteins</td>
<td>178-260</td>
<td>0.97</td>
<td>0.07</td>
<td>0.81</td>
</tr>
<tr>
<td><strong>Variable selection - self-consistent method</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SELCON(8)</td>
<td>17 proteins</td>
<td>178-260</td>
<td>0.95</td>
<td>0.09</td>
<td>0.84</td>
</tr>
<tr>
<td>SELCON</td>
<td>33 proteins</td>
<td>178-260</td>
<td>0.93</td>
<td>0.09</td>
<td>0.91</td>
</tr>
<tr>
<td><strong>Neural network analysis</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDNN2(9)</td>
<td>17 proteins</td>
<td>178-260</td>
<td>0.93</td>
<td>0.10</td>
<td>0.73</td>
</tr>
<tr>
<td>K2D2(10)</td>
<td>19 proteins</td>
<td>200-240</td>
<td>0.95</td>
<td>0.09</td>
<td>0.77</td>
</tr>
<tr>
<td><strong>this work</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAPITO</td>
<td>107 proteins</td>
<td>178-260</td>
<td>0.96</td>
<td>0.11</td>
<td>0.80</td>
</tr>
</tbody>
</table>


**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 data set are depicted in black, other temperatures as indicated in the inset. CD spectrum of GB1 after heat treatment and cooled down to 4 °C is shown in orange. Lower panel: CD values at λ= 200 nm plotted versus the values at λ= 222 nm to deduce the folding state of GB1 at the respective temperatures.


