BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 16 2006, pages 1955–1962
doi:10.1093/bioinformatics/btl327

Structural bioinformatics

A reference database for circular dichroism spectroscopy covering fold and secondary structure space

Jonathan G. Lees[1,|], Andrew J. Miles[1,†], Frank Wien[1,†,~] and B. A. Wallace[1,2,*]

[1]Department of Crystallography, Birkbeck College, University of London, London WC1E 7HX, UK and [2]Centre for Protein and Membrane Structure and Dynamics, Daresbury Laboratory, Warrington WA4 4AD, UK

Received on April 11, 2006; revised on June 7, 2006; accepted on June 8, 2006

Advance Access publication June 20, 2006
Associate Editor: Alex Bateman

ABSTRACT

Motivation: Circular Dichroism (CD) spectroscopy is a long-established technique for studying protein secondary structures in solution. Empirical analyses of CD data rely on the availability of reference datasets comprised of far-UV CD spectra of proteins whose crystal structures have been determined.

This article reports on the creation of a new reference dataset which effectively covers both secondary structure and fold space, and uses the higher information content available in synchrotron radiation circular dichroism (SRCD) spectra to more accurately predict secondary structure than has been possible with existing reference datasets. It also examines the effects of wavelength range, structural redundancy and different means of categorizing secondary structures on the accuracy of the analyses. In addition, it describes a novel use of hierarchical cluster analyses to identify protein relatedness based on spectral properties alone. The databases are shown to be applicable in both conventional CD and SRCD spectroscopic analyses of proteins. Hence, by combining new bioinformatics and biophysical methods, a database has been produced that should have wide applicability as a tool for structural molecular biology.

Contact: b.wallace@mail.cryst.bbk.ac.uk

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

1 INTRODUCTION

Circular dichroism (CD) spectroscopy is a valuable method for determining the secondary structural content of proteins (Woody, 1996; Johnson, 1999; Kelly *et al*., 2005). Today it is finding additional uses in the context of structural genomics and structural biology programmes for examining the secondary structure of expressed proteins, including proteins whose structures have been predicted from modeling studies. A wide range of empirical analysis methods (Chen and Yang, 1971; Brahms and Brahms, 1979; Hennessey and Johnson, 1981; Provencher and Glockner, 1981; Wallace and Teeters, 1987; Pancoska and Keiderling, 1991; Johnson, 1999; Sreerama and Woody, 2000) have been developed that rely on the availability of a reference database of CD spectra of proteins whose structures are known. The suitability of such databases for the analyses of the structures of other proteins depends greatly on whether the database proteins encompass the range of structures (both secondary structures and fold motifs) that are present in the protein to be analyzed. Hence the variety of the proteins in the database and the consistency of data collected defines the accuracy of the methods.

The reference databases in most common use today are found in the CDPro program package (Sreerama and Woody, 2000), and are comprised of CD spectra collected by several authors over the past 30 or more years (Chen and Yang, 1971; Brahms and Brahms, 1979; Hennessey and Johnson, 1981; Pancoska and Keiderling, 1991; Sreerama and Woody, 2000). They include data in the far-ultraviolet wavelength range, with their lower wavelength limits being between 178 and 190 nm, depending on the particular database. These wavelength ranges include those practically achievable on conventional CD instruments. In recent years the availability of synchrotron radiation CD (SRCD) spectroscopy, which takes advantage of the intense light generated in a synchrotron, has made it possible to obtain even lower wavelength data. SRCD data collection has been reported well into the vacuum ultraviolet wavelength range (wavelengths <190 nm) (Wallace, 2000). The additional data contains more electronic transitions (Toumadje *et al*., 1992; Wallace and Janes, 2001; Seranno-Andres and Fulscher, 2003; Gilbert and Hirst, 2004), as well as improved signal-to-noise ratios. It has been demonstrated that a larger spectral data range (Toumadje *et al*., 1992), lower noise (Hennessey and Johnson, 1981), structural variety (Oberg *et al*., 2003), high protein structure quality (Johnson, 1999) and careful bioinformatics definitions (Janes, 2005) are important considerations for a good reference database. With this in mind we decided to generate a well-calibrated, internally consistent, wide wavelength range reference dataset containing a large variety of proteins, which effectively cover secondary structure and fold space. The latter is now possible due to advances in bioinformatics that have enabled the systematic classification of protein folds and architectures. The availability of the CATH database (Orengo *et al*., 2003; Pearl *et al*., 2005) has facilitated the selection of proteins that not only cover the range of secondary structures possible, but also a wide range of protein architectures.

[*]To whom correspondence should be addressed (at Birkbeck).
[|]Current address: School of Biological and Chemical Sciences, Queen Mary, University of London, London E1 4NS UK
[~]Current address: Synchrotron Soleil, Orme Merisiers, BP 48 St. Aubin, Gif sur Yvette, F-91192 France
[†]These authors contributed equally to this work

2 METHODS

2.1 Bioinformatics

2.1.1 Criteria for protein selection The CATH database (Orengo *et al*., 2003) was used as a guide to select proteins from a wide variety of protein families. In addition, proteins were chosen based on their secondary structures to encompass the range of secondary structures present in all proteins currently found in the Protein Data Bank (PDB) (Berman *et al*., 2000), using search tools available at the PDB website. In order to adequately cover the CATH architectures, it was realized that there should be more representatives of the high β-sheet/low α-helix type secondary structures, as these have more diverse types of structures, and consequently more diverse spectra.

2.1.2 PDB file selection and secondary structure assignments For several of the proteins used, more than one crystal structure was available in the PDB. The structure files chosen to define the secondary structures were based on criteria of highest resolution and low R and R_{free} factors (Brunger, 1992), as well as maximum correspondence to the correct protein sequence (i.e. the fewest missing residues and the same species). In cases where more than one PDB file with similar resolution was available, the protein with the lower PROCHECK G-factor (Laskowski *et al*., 1993) score was selected. The crystallographic quality parameters for each protein are listed in Supplementary Table 1S.

The DSSP algorithm (Kabsch and Sander, 1983) was used to assign secondary structures which were split into distorted and regular helix (α_{D}, α_{R}) and distorted and regular sheet (β_{D}, β_{R}) classes as defined previously by Sreerama *et al*. (1999a). That is, two residues at each end of an α-helix (H) or 3_{10}-helix (G) were assigned to the α_{D} fraction. In cases where the helix was less than four residues in length, all of the residues were assigned to the α_{D} fraction. All other α-helix or 3_{10}-helix residues were assigned to the α_{R} fraction. One residue at each end of a β-sheet (E) was assigned to the β_{D} fraction. In cases where the β-sheet was less than two residues in length all of the residues were assigned to the β_{D} fraction. Any remaining β-sheet residues were assigned to the β_{R} fractions. Two or more consecutive 'T' or 'S' assignments from DSSP were assigned to the Turn fraction. The remaining residues which do not fall into one of these categories, including any with missing backbone chain electron density in the crystal structure, were assigned to the 'Other' fraction.

The polyproline-II (PP-II) helix assignment was assigned to those residues in the core PP-II helix region (Adzhubei and Sternberg, 1993) of the Ramachandran map. This area was taken to be an ellipse centred at $\phi = -75°$, $\phi = 145°$, with major and minor axis lengths of 70° and 50°, respectively. The axis of the ellipse was set parallel to the main diagonal (top left, bottom-right) of the Ramachandran map. This method allows for PP-II helix assignment even for stretches of 1 residue in length, which have the correct phi and psi angles.

2.2 Materials

Suitable commercially-available proteins were identified as those obtainable in milligram quantities and of $\geq$95% purity as defined by SDS–PAGE. Alternatively, if the purification procedures included affinity chromatography or crystallization, the protein was considered to be sufficiently pure in the absence of SDS–PAGE information. In some of these cases, the purity of proteins was assessed in house by SDS–PAGE on PHAST gels (Orry *et al*., 2001). Furthermore, the quantitative amino acid analyses (QAA) (see below) done on each sample served as a further test of purity and a criterion for rejection. Several of the proteins targeted from the CATH database were not available from commercial suppliers but were particularly desirable because of unique secondary structures or folds. These were obtained from other labs who had expressed and purified the proteins, for the most part using the original methods used to prepare crystals. The nominal purity (and source) of each protein is listed in Supplementary Table 1S.

2.3 Data collection

2.3.1 CD instrument calibration SRCD spectra were measured at station CD12, Synchrotron Radiation Source Daresbury (SRS), UK and in most cases also on station UV1, Institute for Storage Ring Facilities (ISA), Denmark. In some instances, spectra were also obtained on U11 at the National Synchrotron Light Source (NSLS), USA, and on an Aviv 62ds conventional CD spectrophotometer. The aim of the duplicate measurements was to demonstrate that the spectra were independent of the instrument used for data collection, and this reproducibility was used as one criterion of spectral quality.

The instruments were calibrated for wavelength using a certified holmium glass filter and benzene vapour (Miles *et al*., 2005). Following each beam injection the SRCD instruments were calibrated for spectral magnitude and ratio using a solution of (+)-camphour-10-sulphonic acid (CSA). The CSA concentration was determined from its UV absorption peak at 285 nm where $\varepsilon_{285} = 34.6\ \mathrm{M}^{-1}\ \mathrm{cm}^{-1}$ (Miles *et al*., 2004).

2.3.2 Spectral measurements The proteins were dissolved in 18.2 MΩ deionized water (dH_2O) at 4°C over a period of from 2 to 24 h, depending upon their solubility. Initial concentrations were $\sim$8 mg/ml for the α-helix-rich proteins, 10 mg/ml for mixed α/β proteins and 20 mg/ml for the β-sheet-rich proteins. 0.2 ml of each solution was dialysed against 50 ml of dH_2O for 2 h to reduce any salt content, degassed and then centrifuged at $12\,000 \times g$ for 1 min to remove any undissolved material (Miles and Wallace, 2006). Before collection of the CD spectrum, many of the protein concentrations were determined using the absorbance method (Gill and von Hipple, 1989; Kelly *et al*., 2005) in the presence of 6 M guanidine hydrochloride. Immediately following data collection, duplicate 10 μl aliquots for all proteins were removed for QAA at the Protein and Nucleic Acid Chemistry (PNAC) facility of the University of Cambridge (UK). If the two QAA measurements deviated by >5%, the sample was rejected from inclusion in the database. To examine the environmental sensitivity of the spectra, a number of the proteins were also collected in either 150 mM salt, 50 mM phosphate buffer (pH 7.0) or under the crystallization buffer conditions. No detectable differences were seen for any of these proteins included in the database.

The following parameters were used on station CD12: step-size 1 nm, dwell-time 1 s, bandwidth 1 nm, temperature 4°C. Three separate scans of the protein and dialysate baseline were collected over the wavelength range from 280 to 165 nm. They were checked for reproducibility and then the matching samples and baselines were averaged and subtracted from each other. Spectra were collected in a 0.0015 cm pathlength Suprasil demountable cell (Hellma UK Ltd). Most of the spectra obtained in this cell were usable to 175 nm. Spectra of several proteins were collected to lower wavelengths using a custom-made CaF_2 cell (Hellma, Jena) with a pathlength of 0.0004 cm (Wien and Wallace, 2005). The cell pathlengths were determined using the interference fringe method (Miles *et al*., 2003). The cells were placed in a specially designed cell holder, so that the two halves of the cell were always oriented the same way relative to each other and to the beam, thus ensuring baseline consistency. On station CD12 the amide absorption at $\sim$190 nm, was restricted to values such that the high tension did not exceed the cutoff limit (Miles and Wallace, 2006). Identical parameters were used on UV1 except for the dwell time, which was set to 3 s. A flat baseline and a close match between consecutive spectra were also required. All spectral processing steps were carried out using the CDtool, v1.4, software package (Lees *et al*., 2004). The final spectra were smoothed with a third order Savitsky–Golay smoothing algorithm (Savitsky and Golay, 1964) using a smoothing window of seven data points. Spectral magnitudes are expressed $\Delta\varepsilon$ which is equal to the mean residue ellipticity [(θ)] divided by 3296.

2.4 Methods of analysis

2.4.1 Calculating information content The information content of the dataset was calculated as previously described (Toumadje *et al*., 1992). By using singular value decomposition (SVD) it is possible to represent spectral data in a matrix A in terms of three further matrices as shown by Equation (1). Matrices U and V are unitary matrices and W is a diagonal matrix. The matrix U contains the orthogonal vectors, ordered with corresponding importance for reconstructing matrix A. The matrix W has non-zero values on the main diagonal, which are the positive square of the corresponding eigenvalues. The vector components of UW are referred to as the basis spectra of the reference dataset. The matrix V^{T} contains the least squares coefficients that refit the basis CD spectra to the experimental CD spectra.

$$A = UWV^{\mathrm{T}} \tag{1}$$

It is possible to reproduce the original data in the matrix A to an error within the noise of the measurements by only including a subset of the most significant basis spectra. Within the framework of SVD analysis it is possible to calculate the variance unaccounted for (σ^2) in reconstructing the matrix A [Equation (2)], when using only the μ most significant basis spectra as follows:

$$\sigma_{\mu}^{2} = \left(\frac{1}{N(m-\mu)}\right)\sum_{i=\mu+1}^{m} s_i^2, \tag{2}$$

where, A = matrix containing the spectral data of different reference dataset proteins in its columns, σ^2 = variance unaccounted in reconstructing matrix A, s = singular values of W, μ = number of basis spectra used for reconstruction, m = the number of reference dataset spectra and N = number of spectral data points. The inherent instrument noise was estimated both visually in the raw spectra (in ellipticity units) unscaled for concentration and pathlength, and by SVD. The SVD method involved taking repeat baseline-subtracted spectra and carrying out SVD on the smoothed data. The second basis spectrum from this decomposition was taken be the instrument noise.

2.4.2 Secondary structural analyses The SELCON3 algorithm is one of the most popular methods for secondary structure prediction and is available as part of the CDPro package (Sreerama and Woody, 2000). However, the maximum number of reference protein spectra that the program would accept was 60; in addition, in some instances it did not provide a solution for the protein because of the restraints and default values available. Hence a modified version of SELCON3, named SELMAT3, was implemented in MATLAB (Matlab, 2005). By relaxing the sum, fraction, spectral and helix rules at each stage of the algorithm, it was possible to ensure that at least one solution could be obtained for every protein. Scripts for the SELMAT3 method (which require access to MATLAB software) are available on request to the corresponding author.

2.4.3 Analyses of calculation accuracies The accuracy of secondary structure calculation methods was measured using 'leave one out' cross-validation. In this method the protein spectra in the dataset were removed one after another and the remaining spectra are used to calculate the test proteins secondary structure content. The Pearsons correlation coefficient, r, and the root mean squared deviation (δ), [Equation (3)] were used to quantify the cross-validation performances. Values of r range from +1 to −1, representing perfect positive and negative correlations, respectively. When judging the performance of a method, high values of r and low values of δ indicate good performance. More recently the ζ parameter has been introduced to indicate how much better a prediction is than random (Oberg *et al*., 2004). This is defined here as the ratio of δ over the population standard deviation [Equation (4)]. Higher values indicate better predictions and ζ values of <1.0 are assumed to have little or no predictive value.

$$\delta = \sqrt{\frac{\sum_{i=1}^{n}(f_i^{\mathrm{CD}} - f_i^{X})^2}{n}} \tag{3}$$

$$\zeta = \frac{\delta}{\sigma_X} \tag{4}$$

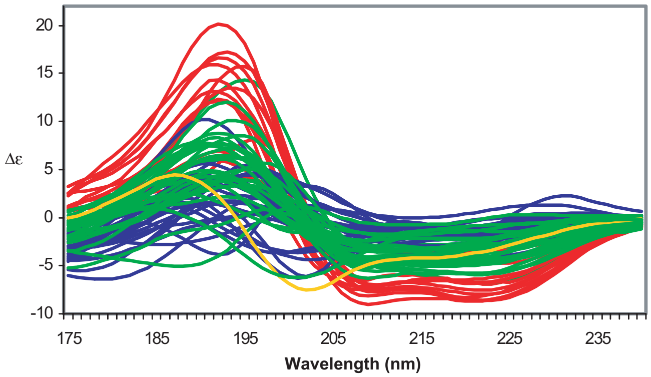

Fig. 1. CD Spectra of all component proteins in the SP175 database. The spectra of the mainly α-helical proteins are in red, the mainly β-sheet in blue, the mixed helical/sheet in green and the 'other' or 'few secondary structures' are in yellow.

In these equations, f^{CD} is the fraction of secondary structure calculated from CD, f^{X} is the fraction of secondary structure calculated from the PDB structure, n is the number of CD spectra, and σ_X is the the standard deviation of the calculated fractions of secondary structure x.

3 RESULTS

3.1 Database components

Spectra were originally collected for >90 proteins, of which 74 met the purity, crystallographic quality, structural characteristics, concentration analysis reproducibility and spectral quality criteria defined above as requirements for the database. Of these 72 have low wavelength cut-offs ≤175 nm (Fig. 1, Supplementary Table 1S), with 39 having cutoff limits ≤170 nm. These were used to produce the reference databases designated SP175 and SP170, respectively, for 'soluble protein 175 cutoff' and 'soluble protein 170 cutoff'. The corresponding PDB structures for the SP175 and SP170 proteins have average resolutions of 1.9 and 1.8 Å, respectively (Supplementary Table 1S).

3.2 CATH distribution

The CATH database categorizes protein structures at several levels of organization, designated by idenification codes in the format C.A.T.H., where C stands for class (secondary structure type), A for architecture (orientation of secondary structural features), T for topology (connectivity of secondary structures) and H for homologous superfamily (high structural similarities and functions). The most basic level is Class, where proteins are divided according to the secondary structure composition into mainly-α, mainly-β, mixed α/β and low secondary structure content classes. The CATH class distribution for the SP175 dataset has larger proportions of mainly-β and mixed α/β proteins than mainly-α proteins (Fig. 2a). This is a desirable feature for a CD reference dataset since the mainly-β structures show a greater range of spectral diversity than the α-helical structures (Wallace *et al*., 2004).

The SP175 dataset contains 72 proteins with different CATH (x.x.x.x) classifications. Of these, 19 different CATH Architectures (x.x) are represented (Fig. 2b) and 55 different Topologies (x.x.x) (Supplementary Table 1). In total, 61 of the proteins contain only a single CATH fold type (i.e. are single domain proteins). Of these single domain proteins, all of the original nine superfolds (Orengo

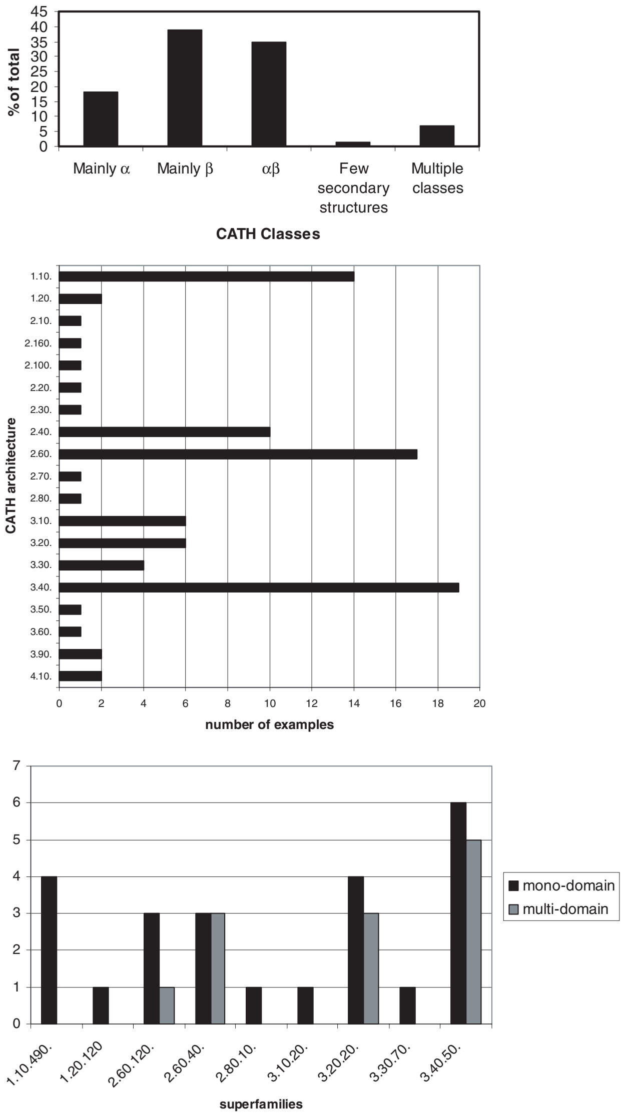

Fig. 2. Bar graphs showing CATH distributions of the component proteins in the SP175 database. (**a**) The percentages of different CATH classes. The 'multiple classes' grouping is for multidomain proteins that contain more than one type of CATH class. (**b**) The number of examples of each type of CATH architecture. (**c**) The number of examples of each of the nine CATH superfolds.

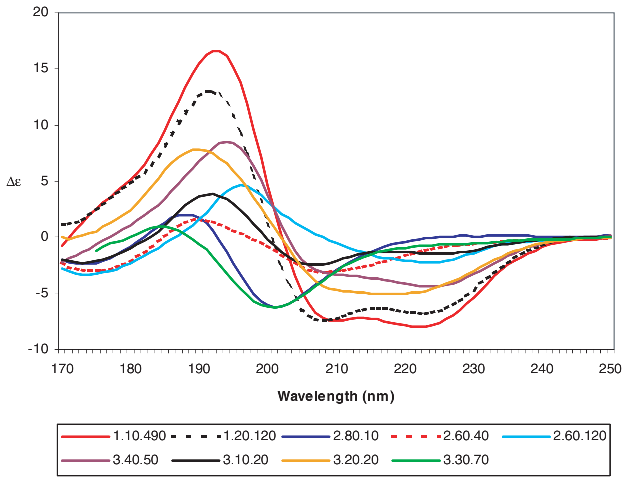

Fig. 3. Representative spectra of all nine different CATH superfolds present in the SP175 reference database.

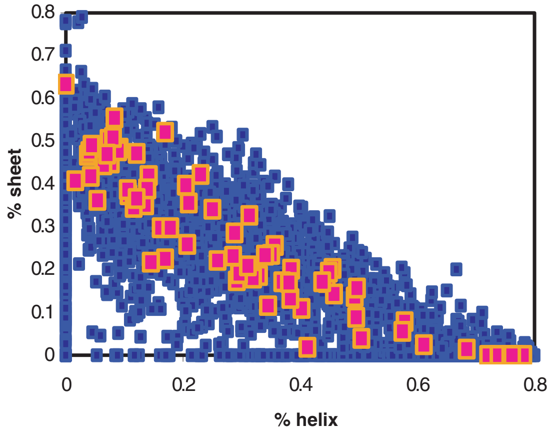

Fig. 4. Database coverage of secondary structures. Plot showing the distribution of helix (DSSP- H, G) versus β-sheet (DSSP- E, B) secondary structures for each of the proteins in the SP175 dataset (each protein is represented by one red square). For comparison, the corresponding distribution of helix versus β-sheet for all PDB structures (non-redundant at the 25% sequence identity level) is shown in blue.

et al., 1994) are represented (Fig. 2c). Example spectra of the superfolds (Fig. 3) demonstrate that they are spectrally diverse, which is a good argument for why they all need to be represented within the database.

3.3 Secondary structure distribution

The distribution of secondary structures [as assigned by calculations from the DSSP algorithm (Kabsch and Sander, 1983) on the crystal structures] was examined for the SP175 dataset (Fig. 4) and compared with the secondary structure distribution from the PDB-SELECT non-redundant (at the level of 25% identity) dataset of high quality structures (Hobohm *et al.*, 1992). The results show that the new dataset has a good coverage of the main area of α-helix/β-sheet content, as well as examples at the extremes of high α-helix and β-sheet. The dataset includes additional data points in the high β-sheet region, again a deliberate feature because of the diversity of this type of structure. Areas not covered well are the regions with low overall secondary structure content, but there are also few examples of these in the PDB, and they tend to be unique structures. Hence, conclusions about the performance of our dataset will be restricted to those proteins with α-helix and β-sheet contents that are more typical for globular proteins.

3.3.1 Information content The SP175 reference dataset contains over 95% of the variance in the first three basis spectra (Fig. 5a). However, even the eighth basis spectrum has intensity significantly larger than the estimated inherent SRCD instrument noise (in this

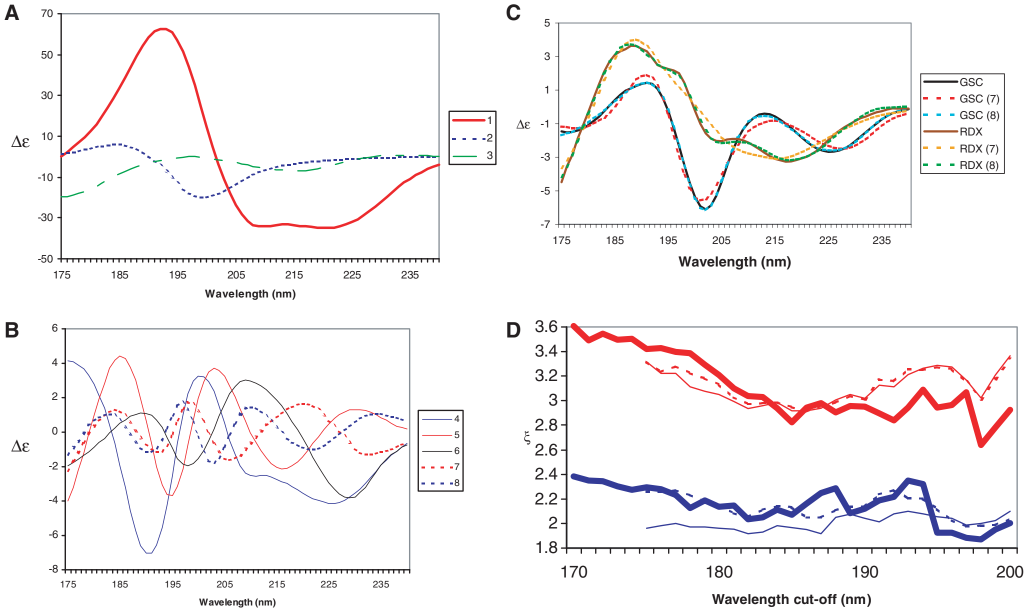

Fig. 5. Information content in the basis sets. (**a**) The first to the third basis spectra for the SP175 dataset, (**b**) The fourth to the eighth basis spectra for the SP175 dataset, (**c**) Plot showing the experimental (solid lines, black and brown, respectively) and fitted (dotted lines) CD spectra of gamma-S-crystalin C-terminus (GSC) and rubredoxin (RBX). The results with different number of basis spectra (either 7 or 8) are shown as dotted lines. It can be seen that the 7 basis spectra fits do not fully reproduce all of the spectral features, although the 8 basis spectra fits do. (**d**) The ζ values for α-helix (in red, upper curves) and β-sheet (in blue, lower curves) as a function of low wavelength cut-off for the SP170 (thick solid line), SP175 (thin solid line) and SP175-jacalin removed (thin dashed line) reference datasets by SELMAT3 method using the α-helix (H), β-sheet (E) and other (combined G, I, T, B, S, C categories) secondary structure assignments from DSSP.

Table 1. The cross-validated predictive performance of the SP175 dataset by the SELMAT3 method

Structure	SP175			SP175(nr)			CDPro29		CDPro43	
	δ	r	ξ	δ	r	ξ	δ	r	δ	r
α_R	**0.048**	**0.956**	3.42	0.049	0.954	3.34	0.054	0.946	0.053	0.941
α_D	**0.035**	**0.809**	1.70	0.037	0.776	1.58	0.052	0.717	0.044	0.776
β_R	**0.073**	**0.792**	1.64	0.083	0.725	1.43	0.087	0.646	0.086	0.663
β_D	**0.020**	**0.913**	2.44	0.023	0.891	2.20	0.034	0.742	0.031	0.743
Turn	**0.052**	0.325	1.02	0.055	0.261	0.96	0.062	**0.482**	0.073	0.367
Other	**0.050**	**0.717**	1.42	0.055	0.671	1.29	0.101	0.300	0.098	0.216

For comparison the performance measures calculated using this method are shown for previously published datasets CDPro29 and CDPro43 (Sreerama and Woody, 2000). SP175(nr) is the more stringent non-redundant cross-validation. The best performance parameters (r or δ) for each category of secondary structural type are shown in bold.

case < 0.20 Δε) (Fig. 5b). Reconstruction of the spectra using only seven basis spectra produces a fit that is within 2 SD of the noise (Fig. 5c). However, inclusion of the eighth basis spectrum is necessary for accurate reconstruction of all the spectral features (Fig. 5c). From this we can conclude that the information content of the SP175 dataset is ~8. It is noteworthy that the fourth to eighth basis spectra have significant variation at wavelengths below 190 nm, hence the low wavelength data will contribute to distinguishing between spectral (and thus structural) characteristics, and thus including the lowest wavelength data possible should benefit the analyses.

3.4 Comparison of predictive ability with existing reference datasets

The secondary structure assignment definitions we have used is the α_R, α_D, β_R, β_D and turn scheme (Sreerama *et al*., 1992a) that has been widely used to assess reference dataset prediction accuracy (Sreerama and Woody, 2000; Sreerama *et al*., 1999a; Oberg *et al*., 2004). The leave-one-out cross-validated performance of the SP175 dataset shows significant improvements over the existing CDPro reference datasets used for comparison (Table 1). The low wavelength limits of the CDPro29 and CDPro43 datasets are 178 and 190 nm, respectively. The best results—defined by either lowest δ or highest r, are noted in bold. It is seen in all categories, but particularly for the β_R and β_D and 'other' fractions, that the new dataset produces better correlations. Indeed, for all categories other than turn, the results are very good. The ξ parameter indicates that the reference dataset has virtually no predictive ability for turn structures. This is probably because the turn category is an amalgamation of a variety of secondary structural types, which have very different phi and psi angles and thus are expected to have very different spectral characteristics (Brahms and Brahms, 1979). It

Table 2. The cross-validated predictive performance of the SP175 dataset by the SELMAT3 method using an alternative secondary structure assignment scheme

Structure	SP175		
	δ	r	ξ
α-helix (H)	0.062	0.958	3.48
3_{10}-helix (G)	0.034	0.026	0.94
PP-II helix	0.038	0.729	1.44
β-sheet (E)	0.072	0.862	1.97
Other	0.063	0.640	1.29

Letters in brackets indicate the DSSP secondary structure assignment. PP-II helix was assigned as described in the methods section.

is expected that when enough examples of different turn types become available, improvements may be possible in this category too. Likewise, if 3_{10} and PP-II helices are considered separately (Table 2) from the helix and 'Other' fractions, respectively, the predictive accuracy for α-helices and β-sheet structures is comparable, but the information derived is more specific.

The alternative secondary structure assignment scheme (Table 2) shows that the α-helix fraction as defined by DSSP has very high predictive accuracy, but that 3_{10} helices cannot be predicted with any meaningful accuracy according to both the r and ξ parameters. This is probably owing to the sparse population of this secondary structural type in the database proteins (only 4.6% of the residues are defined as 3_{10} helices). The PP-II helix has a low but significant above random level of predictive accuracy according to all three parameters, suggesting this secondary structure should be separated out from the 'Other' fraction. Identifying further profitable sub-divisions of secondary structural types and fold motifs in the future may be possible using new algorithms (J. G. Lees *et al.*, manuscript in preparation); by virtue of its wide coverage of folds and structural types, the SP175 dataset will be well suited to this task.

Examining the individual proteins using the leave-one-out method revealed δ values from 0.007 (pyruvate kinase) to 0.15 (jacalin), with jacalin being considerably poorer than the next worst (monellin $= 0.11$). The regular β-sheet content of jacalin was predicted to be 15% instead of 47% found in the crystal structure. This under-prediction can be attributed to an unusual negative peak in the spectrum at 190 nm. This peak could arise from a number of sources including a large number of aromatic residues (15% of all residues are either tyr, trp or phe) in unusual environments (Sreerama *et al.*, 1999b), or an unusual twist to the sheets. Relatively mild heating of jacalin from 20 to 50°C gradually removes this negative peak and produces a classic β-I type spectrum (Sreerama and Woody, 2003) at 50°C (data not shown), and reduces the error to $\delta = 0.08$, reinforcing the idea that the unusual spectral features may be related to tertiary rather than secondary structural features. If jacalin is removed from the SP175 dataset, but is included in the cross-validation there are small improvements in the ξ parameter (Fig. 5d).

Cross-validation should emulate the type of prediction accuracy of the most general case. For a query protein we would not expect there to be a >20% chance of it sharing the same homologous superfamily (x.x.x.x) with a protein in the reference dataset. However, this is the case for several proteins in the CDPro29, CDPro43 and SP175 reference datasets. In order to assess the effects of this redundancy, the cross validation was repeated with the additional constraint that any protein sharing the same CATH homologous superfamily assignment was removed along with the test protein from the reference dataset. The results show that there is only a small drop in the performance when cross-validation is carried out under these more stringent conditions [Table 1, SP175(nr) column].

Table 3. The cross-validated predictive performance of the SP170 dataset by the SELMAT3 method

Structure	SP170		
	δ	r	ξ
α_R	0.056	0.952	3.23
α_D	0.040	0.781	1.60
β_R	0.059	0.889	2.17
β_D	0.026	0.869	2.02
Turn	0.057	0.284	0.971
Other	0.053	0.761	1.50

3.5 Effect of the low wavelength cut-off on predictive performance

Previous methods have emphasized the importance of the low wavelength cut-off on the predictive accuracy of the reference dataset (Toumadje *et al.*, 1992). Our results show that the SP170 reference dataset, which includes data to lower wavelengths, has a good secondary structure prediction performance especially for the 'Other' fraction (Table 3), despite it containing considerably fewer spectra. The effects of truncation on the SP170 reference dataset with the SELMAT3 algorithm show that the lowest wavelength cut-off of 170 nm gives the best overall performance of all of the wavelength ranges tried (Fig. 5d). However, the correspondence of performance with low wavelength cut-off (Supplementary Table 2S) was not as pronounced as previously suggested in a study based on a more limited set of spectra (Matsuo *et al.*, 2004). The results for the ζ parameter do show a gradual drop in the accuracy with increasingly low wavelength cutoff for the α-helix fraction from 170 to 200 nm. A smaller decrease is seen for the β-sheet structure. Importantly, this relative lack of sensitivity suggests that the new database should also be very useful even for conventional CD data, which does not extend to wavelengths as low as 170 nm. A similar pattern is seen for the large SP175 dataset where the low wavelength dependence is only weakly present, but is improved by removing jacalin (and hence the contributions from the unusual peak). These results suggest that there is a good advantage to collecting data to 170 nm in comparison to 175 nm, but that the SELMAT3 algorithm may not be making the best use of the low wavelength data.

3.6 Cluster analyses

Hierarchical classification was used as a new means of testing the way in which the CD spectra of SP170 dataset proteins were related in an objective manner, without reference to their sequence or structural similarities. The city-block distance metric was used with the hierarchical clustering routine using the pairwise average joining method of MATLAB (Matlab, 2005). The cophenetic correlation function for the cluster is 0.77. The results show that the spectra cluster principally on secondary structure composition (Supplementary Figure 1S), but also have a general relationship to CATH classifications. All of the large mainly-α-helical proteins, CATH class 1, (red) cluster together at a large distance from the other spectra, probably reflecting their much larger spectral magnitudes, and are well separated from the class 2 mainly β-proteins (blue and purple). At the 'architecture' level of classification, a cluster of type β-II protein spectra (STI, rubredoxin, elastase, alpha-chymotrypsin) occurs separately (blue) from the β-I proteins (purple). β-II spectra tend to resemble the spectra of disordered proteins owing to their low β-sheet/PP-II helix ratios (Sreerama and Woody, 2003). A distinct set of mainly-β proteins (streptavidin, carbonic anhydrase-I, IgG) with large minima around 175 nm relative to the ~195 nm peaks are also found to cluster together within the β-II cluster. The class 3, mixed α–β proteins (green) are clustered with some of the mixed $\alpha + \beta$ proteins separately from the mixed α/β proteins (green).

4 DISCUSSION

The SP175 dataset, designed on bioinformatics principles to cover fold and secondary structure space, clearly results in an improved performance in the prediction of the α-helix, β-sheet and 'Other' secondary structure fractions in comparison with the reference datasets currently in use for secondary structure analyses. Because it includes more protein structural types, it should result in improved analyses for a wider range of globular proteins. However it is important to note that it may not be useful for other structural types such as fibrous proteins and peptides (but neither are any of the existing reference databases). In addition, it should be noted that all the spectra reported are for proteins at a temperature of 4°C. Whereas protein spectra generally do not vary significantly up to ~37°, their denaturation points do differ and the user should consider this parameter in any conclusions that they make using the SP175 (or any) reference dataset.

The low wavelength cut-off results show that the optimal secondary structure analysis is found when the maximal wavelength range (170 nm) of data is utilized, but that the wavelength-dependence of the types of the analyses reported here are not large. Indeed, if the spectra have been accurately calibrated and the larger SP175 dataset is used, a good prediction of helix content can still be obtained for data truncated to 200 nm. Furthermore, even sheet and other secondary structures can be reasonably well predicted using only data to 190 nm from the new datasets. Consequently, the results from this study suggest that using this database, even with data truncated at higher wavelengths (such as that obtained on a conventional CD instrument) can still produce good performances. However, identification of other features, such as detected by the cluster analyses may only be possible when the lower wavelength data are available, and the development of alternative algorithms, including neural networks, may find the low wavelength data more valuable than the types of algorithms used here which were developed primarily for analyses in the higher wavelength region. It is important however, for future developments, that the SP175 database provides for the first time a comprehensive reference database for SRCD data. In summary, this new database should become a valuable tool for analyses of both CD and SRCD spectra of proteins.

The SP175 dataset will be publicly available [upon publication of this paper] as an alternative reference database in the DICHROWEB calculation server (Lobley *et al*., 2002; Whitmore and Wallace, 2004) located at http://www.cryst.bbk.ac.uk/cdweb, and it will be deposited in the Protein Circular Dichroism Data Bank (PCDDB) (Wallace *et al*., 2006) located at http://pcddb.cryst.bbk.ac.uk.

ACKNOWLEDGEMENTS

The authors thank Dr Robert W. Janes for helpful discussions, and the following for their help and advise at the SRCD beamlines: Dr Soren Vronning Hoffman (ISA), Dr John Sutherland and John Trunk (NSLS), and Dr David Clarke and Alan Brown (SRS). The authors thank Peter Sharratt of the University of Cambridge Protein and Nucleic Acids Centre (PNAC) Facility for the quantitative amino acid analyses, and the following for generously supplying us with protein samples: Drs Jennifer Martin, Christine Slingsby, Robert Evans, Radka Chaloupkova, Robert Woody, Jeremy Lakey and Paul Evans. The authors also thank Dr Curtis Johnson for helpful communications on SVD methodologies. We thank Dr L. Whitmore (Birkbeck College) for adding the SP175 dataset to the Dichroweb server.

This work was supported by grants from the BBSRC to B.A.W. and Dr Robert W. Janes (Queen Mary, University of London). J.G.L. was supported in part by a BBSRC CASE studentship with GlaxoSmithKline and A.J.M. was supported in part by a studentship from the MRC. Beamtime access to the SRS Daresbury has been enabled by a Programme Mode Access grant to B.A.W. and Dr R.W. Janes. Beamtime access at ISA was enabled by the European Community—Research Infrastructure Action under the FP6 'Structuring the European Research Area' Programme to Soren Pape Moller (Aarhus University). Beamtime at the National Synchrotron Light Source, Brookhaven National Laboratory was supported by the US Department of Energy, Division of Materials Sciences and Division of Chemical Sciences, under Contract No. DE-AC02-98CH10886. Funding to pay the Open Access publication charges was provided by the BBSRC.

Conflict of Interest: none declared.

REFERENCES

Adzhubei,A.A. and Sternberg,M.J. (1993) Left-handed polyproline II helices commonly occur in globular proteins. *J. Mol. Biol.*, **229**, 472–493.

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

Brunger,A.T. (1992) Free R Value: a novel statistical quantity for assessing the accurracy of crystal structures. *Nature* **355**, 472–475.

Brahms,S. and Brahms,J. (1979) 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.

Gilbert,A.T.B. and Hirst,J.D. (2004) Charge-transfer transitions in protein circular dichroism spectra. *J. Mol. Struct. (THEOCHEM.)*, **675**, 53–60.

Gill,S.C and von Hipple,P.H. (1989) Calculation of protein extinction coefficients from amino acid data. *Anal. Biochem.*, **182**, 319–326.

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

Hobohm,U. *et al*. (1992) Selection of representative protein datasets. *Protein Sci.*, **1**, 409–417.

Janes,R.W. (2005) Bioinformatics analyses of circular dichroism protein reference databases. *Bioinformatics*, **21**, 4230–4239.

Johnson,W.C.Jr. (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.

Laskowski,R.A. *et al*. (1993) PROCHECK—a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr*., **26**, 283–291.

Lees,J.G. *et al*. (2004) CDtool—an integrated software package for circular dichroism spectroscopic data processing, analysis, and archiving. *Anal. Biochem*., **332**, 285–289.

Lobley,A. *et al*. (2002) DICHROWEB: an interactive website for the analysis of protein secondary structure from circular dichroism spectra. *Bioinformatics*, **18**, 211–212.

MATLAB (2005). [7.0] MathWorks.

Matsuo,K. *et al*. (2004) Secondary-structure analysis of proteins by vacuum-ultraviolet circular dichroism spectroscopy. *J. Biochem. (Tokyo)*, **135**, 405–411.

Miles,A.J. and Wallace,B.A. (2006) Synchrotron radiation circular dichroism spectroscopy of proteins and applications in structural and functional genomics. *Chem. Soc. Rev*., **35**, 39–51.

Miles,A.J. *et al*. (2003) Calibration and standardisation of synchrotron radiation circular dichroism (SRCD) amplitudes and conventional circular dichroism (CD) spectrophotometers. *Spectroscopy*, **17**, 1–9.

Miles,A.J. *et al*. (2004) Redetermination of the extinction coefficient of camphor-10-sulfonic acid, a calibration standard for circular dichroism spectroscopy. *Anal. Biochem*., **335**, 338–339.

Miles,A.J. *et al*. (2005) Calibration and standardisation of synchrotron radiation and conventional circular dichroism spectrometers. Part 2: factors affecting magnitude and wavelength. *Spectroscopy*, **19**, 43–51.

Oberg,K.A. *et al*. (2003) Rationally selected basis proteins: a new approach to selecting proteins for spectroscopic secondary structure analysis. *Protein Sci*., **12**, 2015–2031.

Oberg,K.A. *et al*. (2004) The optimization of protein secondary structure determination with infrared and circular dichroism spectra. *Eur. J. Biochem*., **271**, 2937–2948.

Orengo,C.A. *et al*. (1994) Protein superfamilies and domain superfolds. *Nature*, **372**, 631–634.

Orengo,C.A. *et al*. (2003) The CATH domain structure database. *Methods Biochem. Anal*., **44**, 249–271.

Orry,A. *et al*. (2001) Synchrotron radiation circular dichroism spectroscopy: vacuum ultraviolet irradiation does not damage protein integrity. *J. Synchrotron Radiat*., **8**, 1027–1029.

Pancoska,P. and Keiderling,T.A. (1991) Systematic comparison of statistical analyses of electronic and vibrational circular dichroism for secondary structure prediction of selected proteins. *Biochemistry*, **30**, 6885–6895.

Pearl,F.M. *et al*. (2005) The CATH domain structure database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. *Nucleic Acids Res*., **33**, 247–251.

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

Savitsky,A. and Golay,M.J.E. (1964) Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem*., **36**, 1627–1639.

Serrano-Andres,L. and Fulscher,M.P. (2003) Theoretical study of the electronic spectroscopy of peptides. III. Charge-transfer transitions in polypeptides. *J. Am. Chem. Soc*., **120**, 10912–10920.

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. (2003) Structural composition of beta(I)- and beta(II)-proteins. *Protein Sci*., **12**, 384–388.

Sreerama,N. *et al*. (1999a) Estimation of the number of alpha-helical and beta-strand segments in proteins using circular dichroism spectroscopy. *Protein Sci*., **8**, 370–380.

Sreerama,N. *et al*. (1999b) Tyrosine, phenylalanine, and disulfide contributions to the circular dichroism of proteins: circular dichroism spectra of wild-type and mutant bovine pancreatic trypsin inhibitor. *Biochemistry*, **38**, 10814–10822.

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

Wallace,B.A. (2000) Synchrotron radiation circular dichroism spectroscopy as a tool for investigating protein structures. *J. Synchrotron Radiat*., **7**, 289–295.

Wallace,B.A. and Janes,R.W. (2001) Synchrotron radiation circular dichroism spectroscopy of proteins: secondary structure, fold recognition and structural genomics. *Curr. Opin. Chem. Biol*., **5**, 567–571.

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.

Wallace,B.A. *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.

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**, 668–673.

Wien,F. and Wallace,B.A. (2005) Calcium fluoride micro cells for synchrotron radiation circular dichroism spectroscopy. *Appl. Spectrosc*., **59**, 1109–1113.

Woody,R.W. (1996) Theory of circular dichroism of proteins. In Fasman,G.D. (ed), *Circular Dichroism and the Conformational Analysis of Biomolecules*. Plenum Press, New York. pp. 25–67.