

Structural bioinformatics

## Bioinformatics analyses of circular dichroism protein reference databases

Robert W. Janes

School of Biological and Chemical Sciences, Queen Mary, University of London, Mile End Road, E1 4NS, UK

Received on August 2, 2005; revised on September 21, 2005; accepted on September 23, 2005

Advance Access publication September 27, 2005

### ABSTRACT

**Motivation:** Circular dichroism (CD) spectroscopy has become established as a key method for determining the secondary structure contents of proteins which has had a significant impact on molecular biology. Many excellent mathematical protocols have been developed for this purpose and their quality is above question. However, reference database sets of proteins, with CD spectra matched to secondary structure components derived from X-ray structures, provide the key resource for this task. These databases were created many years ago, before most CD spectrophotometers became standardized and before it was commonplace to validate X-ray structures prior to publication. The analyses presented here were undertaken to investigate the overall quality of these reference databases in light of their extensive usage in determining protein secondary structure content from CD spectra.

**Results:** The analyses show that there are a number of significant problems associated with the CD reference database sets in current use. There are disparities between CD spectra for the same protein collected by different groups. These include differences in magnitudes, peak positions or both. However, many current reference sets are now amalgamations of spectra from these groups, introducing inconsistencies that can lead to inaccuracies in the determination of secondary structure components from the CD spectra. A number of the X-ray structures used fall short on the validation criteria now employed as standard for structure determination. Many have substantial percentages of residues in the disallowed regions of the Ramachandran plot. Hence their calculated secondary structure components, used as a foundation for the reference databases, are likely to be in error. Additionally, the coverage of secondary structure space in the reference datasets is poorly correlated to the secondary structure components found in the Protein Data Bank. A conclusion is that a new reference CD database with cross-correlated, machine-independent CD spectra and validated X-ray structures that cover more secondary structure components, including diverse protein folds, is now needed. However, that reasonably accurate values for the secondary structure content of proteins can be determined from spectra is a testament to CD spectroscopy being a very powerful technique.

**Contact:** r.w.janes@qmul.ac.uk

### 1 INTRODUCTION

Circular dichroism (CD) spectroscopy has become an invaluable research technique, used by many labs worldwide, for gaining information about protein structure, dynamics and interactions both with other proteins and with ligands. This is possible because different types of secondary structures give rise to characteristic CD spectra, which differ in their peak positions and intensities, and to a first approximation a spectrum can be considered to arise from the weighted sum of these components. The information content available from CD is wavelength-range dependent and analyses of spectral data can determine the number of independent eigenvectors needed to reconstruct the original spectrum. For data down to wavelengths of  $\sim 190$  nm this number is between three and four (Hennessey and Johnson, 1981). However, because secondary structure components are not independent of each other (Pancoska *et al.*, 1992), solutions for a greater number of components than there are independent eigenvectors can be found (Hennessey and Johnson, 1981; Wallace and Janes, 2001).

There are a number of methods that have been developed for deconvoluting CD spectra into the calculated secondary structure components present in the protein. These include as examples, linear least-squares (Chen and Yang, 1971; Brahms and Brahms, 1980), parameterized fit (Provencher and Glöckner, 1981), singular-value decomposition (Hennessey and Johnson, 1981), non-linear least-squares (Wallace and Teeters, 1987) and self-consistent variable selection methods (Sreerama and Woody, 1993; Johnson, 1999; Sreerama and Woody, 2000). These methods are based on very sound mathematical approaches. They have been enhanced and refined over the years, and because of this they can yield reasonable results for the calculation of secondary structure content. Many of these methods are to be found in DICHROWEB (Lobley and Wallace, 2001; Lobley *et al.*, 2002; Whitmore and Wallace, 2004), a package designed to aid in the determination of secondary structure content and used world wide. Such is the wide-scale use of CD spectroscopy in research that the creation of the Protein Circular Dichroism Data Bank (PCDDDB) has been proposed, which will act as a repository and resource for CD spectra and associated data (Wallace *et al.*, 2005).

Synchrotron radiation circular dichroism (SRCD), first developed in 1980 (Sutherland *et al.*, 1980), has recently become a potentially

valuable tool for substantially extending the wavelength range of available data due to the increased photon flux of the source over conventional CD (cCD) machines at the lower wavelength limits. For data collected over the full SRCD range, down to  $\sim 160$  nm, the information content rises to at least seven or eight eigenvectors. These data may be deconvoluted into as many as 12 different secondary, and perhaps supersecondary, structure types thereby enabling a much more detailed resolution of structural features than has been possible from a cCD source (Wallace and Janes, 2001).

Empirical determination of the secondary structure components from CD spectral data employs reference databases. These are either a combination of CD spectra from a set of proteins with known secondary structure content, obtained from their X-ray crystallography structures, or principal component spectra derived from a set of individual spectra. Examples of these databases are Chang *et al.* (1978), Bolotina *et al.* (1980a,b), Brahms and Brahms (1980), Provencher and Glöckner (1981), Compton and Johnson (1986), Pancoska and Keiderling (1991) and Sreerama *et al.* (2000). These reference datasets were created early in the development of CD spectroscopy as a technique for proteins, and significantly more recent databases are for the most part combinations of older ones, and include no or limited new protein secondary structure types and few, if any, new protein constituents. For the major reference databases available, the lowest wavelength data included are to 178 nm. Of note, for SRCD measurements, while a higher number of resolvable secondary structure types should potentially be determinable and with a greater degree of accuracy than is possible from cCD sources, currently no databases are capable of covering down to the full wavelength range available to this technique.

The work presented here analyses the current CD reference databases for the quality of the CD data and X-ray structures used, for their breadth of secondary structure types covered and their effectiveness at covering fold space.

## 2 METHODS

### 2.1 CD spectra

CD spectral data were obtained from reference databases at the CDPRO (Sreerama *et al.*, 2000) program website (<http://lamar.colostate.edu/~sreeram/CDPro/>). Additional spectra were obtained from the Brahms and Brahms (1980) reference dataset (provided by Prof. Jon B. Applequist, personal communication) and the Supplementary data (Pancoska and Keiderling, 1995). An SRCD spectrum for  $\gamma$ -crystallin came from Paul Evans and Dr Christine Slingsby. As some original spectra were collected at non-integral wavelengths, an in-house program was used to interpolate these to integral wavelengths for comparison purposes. This did not alter the spectral characteristics in any way however (data not shown). Brahms and Brahms spectra were reported in mean residue ellipticity ( $\theta$ ) values and were therefore scaled to match the delta epsilon ( $\Delta\epsilon$ ) units used in CDPRO by dividing by 3298.

### 2.2 X-ray structure data

The X-ray structure data in Tables 1 and 2 are derived from Pancoska and Keiderling (1995), (their original Table 1 set of structures) and *Exp32* reference set from Sreerama *et al.* (2000). Original Protein Data Bank (PDB) (Bernstein *et al.*, 1977; Berman *et al.*, 2000) files were used for subsequent analyses, even when these had been superseded, as the data within the reference databases are still derived from this original material. Atomic co-ordinates were from PDB files (<http://www.pdb.org/>)

or from the archive site for obsolete structures (<http://pdobobs.sdsc.edu/index.cgi>).

### 2.3 X-ray structure analyses

The resolution was obtained from the PDB files. Structural fold information was from the CATH (class, architecture, topology and homologous superfamily) protein topology website (Orengo *et al.*, 1997; Pearl *et al.*, 2000, 2005). The DSSP (definition of secondary structure of proteins) program (Kabsch and Sander, 1983) was used (<http://bioweb.pasteur.fr/seqanal/interfaces/dssp-simple.html>) to assign secondary structure content to these proteins. Percentages of secondary structure, derived from the DSSP output, were determined using an in-house program as were the percentages and numbers of 'missing' (undetermined) residues in these structures. These missing residues were checked against the sequence data of the native protein in each case (<http://us.expasy.org/srs5/>). PROCHECK (Laskowski *et al.*, 1993) was used to derive the percentages of residues in fully, additionally and generously allowed and disallowed regions of the Ramachandran plot.

### 2.4 Evaluating the correlation coefficient of alpha helix and beta sheet content of all PDB proteins against *Exp32*

The values for percentage alpha helical against beta sheet content of all proteins in the PDB were binned into 'ten percentile tranches' (0–9.99, 10–19.99, etc.), not including any nucleic acid material but leaving in all homologue proteins. The reasoning was that any of these proteins could have their CD spectra recorded and they were therefore eligible for inclusion. The *Exp32* set for proteins were binned in a similar way to enable a direct comparison. To quantify the coverage of 'secondary structure space' of the *Exp32* set compared with the whole PDB, the standard Pearson  $r^2$  correlation coefficient was calculated. To create an idealized set of data maximizing the coverage of these secondary structure components for a reference set containing only 32 proteins, each bin of PDB data was reduced by an overall scaling term such that the total of proteins then approximated to 32. Rounding these new values to the nearest integer, and rounding one value of  $>0.49$  manually to one, created the desired idealized 32 protein set.

### 2.5 Evaluating the correlation coefficient of fold space

In a manner similar to that for the secondary structure components, the fold space of single-domain proteins was obtained from the CATH database for all proteins in the PDB. These data were correlated with those from the *Exp32* set of proteins and a hypothetical set of proteins was also generated to characterize the quality of fold space coverage (in fact comprising 31, as there was one protein unclassified in the *Exp32* set).

## 3 RESULTS AND DISCUSSION

### 3.1 Quality of CD spectra in reference databases

The CD spectra used in many of the current reference databases are derived from amalgamations of previously created datasets from different groups, with the aim of broadening the secondary structure types represented by the proteins in these sets. As examples, *Exp32* used as a stand-alone set in SELCON3 (Sreerama and Woody, 1993; Sreerama *et al.*, 1999) and as part of many sets in CDPRO (Sreerama *et al.*, 2000, <http://lamar.colostate.edu/~sreeram/CDPro/>), and analysed here, contains 29 CD spectra from Johnson (a gift as stated in Sreerama *et al.*, 1999) and 3 from Sreerama *et al.* (1999). The set CDDATA56, currently the largest used in CDPRO containing 56 CD spectra, is a combination from Johnson (Sreerama

**Table 1.** Set of reference proteins used by Pankoska *et al.*, 1995

PDB Code	Res Å	CATH Code	%H	%E	%T	%G	%B	%S	%O	%M	%F	%A	%Ge	%D
4adh <sup>a</sup>	2.40	3.90.180.10 3.40.50.720 <sup>b</sup>	21.12	20.58	14.70	3.74	2.13	11.49	26.20	0	78.3**	17.2	1.9	2.5*
1ca2	2.00	3.10.200.10	8.20	28.90	12.50	8.20	1.17	14.06	26.95	1.15	88.4*	11.6	0	0
2cga	1.80	2.40.10.10	7.34	32.04	14.89	6.12	3.06	9.38	27.14	0	85.3*	12.3	0.9	1.4*
5cha	1.67	2.40.10.10	9.32	32.62	12.50	2.54	2.75	11.65	28.60	2.07	84.8*	15	0.2	0
3can	2.40	2.60.120.200	0.00	40.50	9.28	0.00	0.00	19.83	30.37	0	69.2**	24	3.8	2.9*
1cyt <sup>d</sup>	2.00	1.10.760.10	43.20	0.00	13.10	0.00	1.94	9.22	32.52	0	76.2**	19.2	4.1	0.6*
3est	1.65	2.40.10.10	5.41	34.16	17.08	5.41	3.33	7.08	27.50	0	87.9*	12.1	0	0
2grs <sup>a</sup>	2.00	3.50.50.60(2) 3.30.390.30 <sup>c</sup>	27.11	18.65	10.41	2.16	1.95	17.35	22.34	3.55	72.9**	20.3	4.3	2.5*
1hco	2.70	1.10.490.10	52.96	0.00	18.81	9.75	0.00	6.62	11.84	0	77.1**	20.9	1.6	0.4*
1rei	2.00	2.60.40.10	0.00	49.06	14.01	2.80	0.46	10.74	22.89	0	87.9*	9.9	0	2.2*
4ldh <sup>a</sup>	2.00	3.40.50.720 3.90.110.10 <sup>b</sup>	33.73	11.24	14.28	3.03	2.43	10.63	24.62	0	65.2**	23.5	6.1	4.8*
7lyz	2.50	1.10.530.10	30.23	7.75	20.93	9.30	3.10	12.40	16.27	0	80.5*	17.7	1.8	0
1mbn	2.00	1.10.490.10	77.12	0.00	9.80	0.00	0.00	1.96	11.11	0	83.9*	15.3	0.7	0
8pap <sup>a</sup>	2.80	3.90.70.10	23.11	16.50	8.49	1.41	1.88	18.39	30.18	0	75.6**	23.3	0.6	0.6*
1rhd	2.50	3.40.250.10	27.64	10.92	16.38	2.04	2.38	10.92	29.69	0	77.1**	20.1	2	0.8*
1rn3 <sup>a</sup>	1.45	3.10.130.10	17.74	38.70	11.29	3.22	2.41	10.48	16.12	0	81.7*	18.3	0	0
1rns <sup>a</sup>	2.00	3.10.130.10	17.74	39.51	7.25	3.22	2.41	10.48	19.35	13.88	67.3**	23.9	7.1	1.8*
1sbt	2.50	3.40.50.200	30.18	17.81	15.27	0.00	1.81	10.18	24.72	0	74.3**	21.7	3.5	0.4*
2sod	2.00	2.60.40.200	0.66	36.75	12.08	1.15	2.31	20.86	26.15	0	66.5**	24.4	4.4	4.7*
2tln <sup>a</sup>	2.30	3.10.170.10 1.10.390.10 <sup>b</sup>	29.74	15.18	14.55	3.79	0.94	15.50	20.25	0	77.0**	18.1	4.4	0.4*
1tim	2.50	3.20.20.90	43.72	16.80	6.07	2.22	0.80	9.31	21.05	0	77.3**	17.8	3.6	1.4*
3pti <sup>a</sup>	1.50	4.10.410.10 <sup>d</sup>	13.79	24.13	6.89	6.89	1.72	17.24	29.31	0	87.0*	13	0	0
3ptn	1.70	2.40.10.10	7.17	32.28	14.79	2.69	2.69	15.24	25.11	0	86.2*	13.8	0	0

Some of these are now used for SELCON3 and CDPRO as part of the set containing 56 proteins. The table columns are PDB code, resolution of the structure, CATH code, percentages alpha helix (%H), beta sheet (%E), turn (%T), 3<sub>10</sub>-helix (%G), bridge (%B), bend (%S) and other (%O) [previously called random coil], as defined by DSSP (Kabsch and Sander, 1983), missing residues (%M), and percentages of residues in the fully (%F), additionally (%A), generously (%Ge) allowed and disallowed (%D) regions of the Ramachandran plot. Note that pi-helix (%I) is not included as no proteins had this secondary structure component. The asterisks are used as flags within PROCHECK to indicate potential problems with the structures. For the fully allowed (%F) region, residue percentages <90% have one asterisk, whilst <80% have two asterisks. In the disallowed (%D) region, residue percentages above zero have one asterisk, and those >5% have two asterisks.

<sup>a</sup>Indicates that these PDB files have now been superseded by superior structure files.

<sup>b</sup>This is a two domain protein.

<sup>c</sup>This is a three domain protein.

<sup>d</sup>Refers to 4pti which supersedes 3pti.

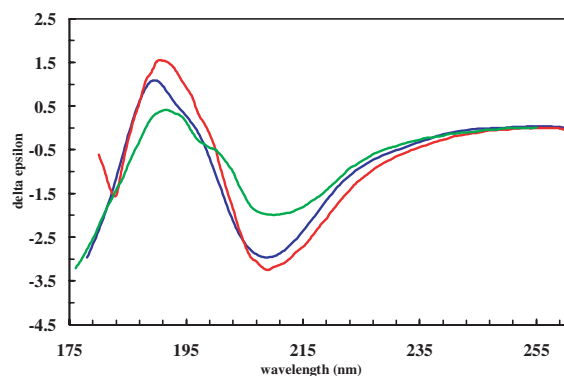
*et al.*, 1999) (29), Sreerama *et al.* (1999) (3), Yang *et al.* (1978) (a gift as stated in Provencher and Glöckner, 1981) (6), Pancoska and Keiderling (1991) (5) and membrane proteins from Park *et al.* (1992) (13). An inherent problem is that the spectra used were often obtained from individual non-commercial CD machines, or machines modified to collect CD data, with, at the time, limited cross-reference calibration. CD spectra for the same protein from the different sets used in these amalgamated databases have different spectral features in a number of cases, as illustrated in Figures 1 and 2. Figure 1 shows the CD spectra of superoxide dismutase from three original databases (Johnson, 1999; Pancoska and Keiderling, 1995 and Brahms and Brahms, 1980). The spectra are considerably different from each other, and yet are from the same source material. In addition, each of these spectra is equated to secondary structure components from the same PDB file (2sod). Only one of these spectra is now used in the current datasets, but it is unclear why one was chosen over the rest and which one is the 'actual' spectrum of the protein? Figure 2a–c show the spectra from Brahms and

Brahms (1980) compared with that from Johnson (1999). Here the differences are not so pronounced as in Figure 1, being either wavelength shifts in the spectra, magnitude shifts in the peaks, ratio differences between peaks or a combination of these, but they are nevertheless serious when being used as the basis for empirical calculations of secondary structure content for novel proteins. Figure 2d compares a spectrum of  $\gamma$ -crystallin from a database set with an SRCD spectrum of the same protein (Evans *et al.*, 2004). Although the spectra have comparable characteristics there is a 10 nm shift between them, the SRCD spectrum being down-wavelength from that in the database. Whilst CD spectra from single-source reference databases were possibly 'internally consistent' when used as an isolated set, when they became components within an amalgamated set, their differences, as exemplified here, created problems with consistency within these combined sets. To ensure consistency, cross-calibration and cross-checking on a diverse range of machines are of vital importance to remove possible machine bias within the data (Miles *et al.*, 2003, 2005).

**Table 2.** *Exp32*—a set of reference proteins used in SELCON3 and many other amalgamation sets<sup>a</sup>

PDB Code	Res Å	CATH code	%H	%E	%T	%G	%B	%S	%O	%M	%F	%A	%Ge	%D
4mbn	2.00	1.10.490.10	74.50	0.00	5.88	5.88	0.00	1.96	11.76	0	92.0	7.3	0.7	0.0
2mhb	2.00	1.10.490.10	67.24	0.00	9.05	8.71	0.00	3.83	11.14	0	89.9*	8.1	2.0	0.0
2hmz	1.66	1.20.120.50	64.60	0.00	8.62	5.53	0.00	4.20	17.03	0	88.0*	11.1	1.0	0.0
2lzm	1.70	1.10.530.40	66.46	8.53	7.31	0.00	0.60	7.31	9.75	0	95.3	4.7	0.0	0.0
3tim	2.80	3.20.20.90	39.35	15.46	10.84	5.42	1.60	5.42	21.88	0.40	89.1*	10.4	0.5	0.0
6ldh	2.00	3.40.50.720	39.81	16.10	8.51	3.95	0.91	9.72	20.97	0	85.3*	10.9	2.7	1.7*
1lys	1.72	1.10.530.10	31.00	6.20	24.03	10.85	4.65	7.75	15.50	0	90.3	8.8	0.4	0.4*
8tln	1.60	3.10.170.10	37.10	16.35	10.37	4.08	0.62	13.52	17.92	0	87.8*	11.9	0.0	0.4*
5cyt	1.50	1.10.760.10	40.77	0.00	16.50	0.00	1.94	8.73	32.03	0	91.8	8.1	0.0	0.0
3pgk	2.50	3.40.50.1260	34.45	11.08	4.09	0.00	0.48	22.89	26.98	0	59.9**	20.3	10.6	9.2**
1eri	2.70	3.40.580.10	28.35	18.77	15.32	5.36	1.53	9.57	21.07	5.43	86.2*	12.5	0.9	0.4*
1fx1	2.00	3.40.50.360	29.25	21.76	17.00	2.72	1.36	11.56	16.32	0.67	84.8*	12.0	0.8	2.4*
1sbt	2.50	3.40.50.200	30.18	17.81	15.27	0.00	1.81	10.18	24.72	0	74.3**	21.7	3.5	0.4*
3gpd	3.50	—	24.85	20.80	10.62	2.54	1.04	14.97	25.14	0	62.8**	21.5	7.5	8.2**
9pap	1.65	3.90.70.10	23.11	16.98	10.84	2.83	4.24	12.26	29.71	0	88.4*	11.6	0.0	0.0
2sbt	2.80	3.40.50.200	21.45	13.81	17.45	0.00	0.00	15.63	31.63	0	59.6**	30.2	7.6	2.7*
3rn3	1.45	3.10.130.10	17.74	33.06	14.51	3.22	2.41	10.48	18.54	0	87.0*	13.0	0.0	0.0
2psg	1.80	2.40.70.10	10.81	38.64	11.35	9.72	1.08	8.64	19.72	0	90.8	8.6	0.6	0.0
1beb	1.80	2.40.128.20	9.93	42.62	11.21	7.37	0.00	12.17	16.66	3.70	87.0*	10.9	0.4	1.8*
5cha	1.67	2.40.10.10	9.32	32.62	12.50	2.54	2.75	11.65	28.60	1.69	84.8*	15.0	0.2	0.0
1azu	2.70	2.60.40.420	11.29	25.80	13.70	0.00	2.41	20.16	26.61	1.56	60.6**	31.2	5.5	2.8*
3est	1.65	2.40.10.10	5.41	34.16	17.08	5.41	3.33	7.08	27.50	0	87.9*	12.1	0.0	0.0
4gcr	1.47	2.60.20.10	2.87	45.97	8.04	6.32	2.29	12.06	22.41	0	91.4	8.6	0.0	0.0
2pab	1.80	2.60.40.180	7.01	50.00	12.28	0.00	0.43	10.08	20.17	10.24	81.8*	15.7	1.5	1.0*
2ctv	1.95	2.60.120.200	0.00	46.41	11.81	3.79	0.84	14.34	22.78	0	87.5*	12.5	0.0	0.0
1rei	2.00	2.60.40.10	0.00	49.06	14.01	2.80	0.46	10.74	22.89	0.93	87.9*	9.9	0.0	2.2*
1tnf	2.60	2.60.120.40	0.00	44.73	8.11	1.97	1.09	16.88	27.19	3.18	67.2**	25.6	4.7	2.6*
2sod	2.00	2.60.40.200	0.66	36.75	12.08	1.15	2.31	20.86	26.15	0	66.5**	24.4	4.4	4.7*
2abx	2.50	2.10.60.10	0.00	10.81	1.35	0.00	0.00	30.40	57.43	0	14.8**	36.9	23.8	24.6**
1col	2.40	1.10.490.30	75.12	0.00	6.59	3.04	0.00	1.26	13.95	3.43	95.2	4.8	0.0	0.0
1ema	1.90	2.40.155.10	3.55	50.66	15.11	6.22	1.33	8.44	14.66	4.20	93.2	6.8	0.0	0.0
1lfc	1.19	2.40.128.20	11.45	58.77	13.74	0.00	0.00	3.05	12.97	0	93.2	6.8	0.0	0.0

<sup>a</sup>The headings to the columns in Table 2 are as described for Table 1.



**Fig. 1.** Superoxide dismutase spectra of the same protein from three CD reference databases Johnson (blue), Pancoska and Keiderling (red) and Brahms and Brahms (green). These spectra are equated to the same secondary structure data derived from PDB file 2sod.

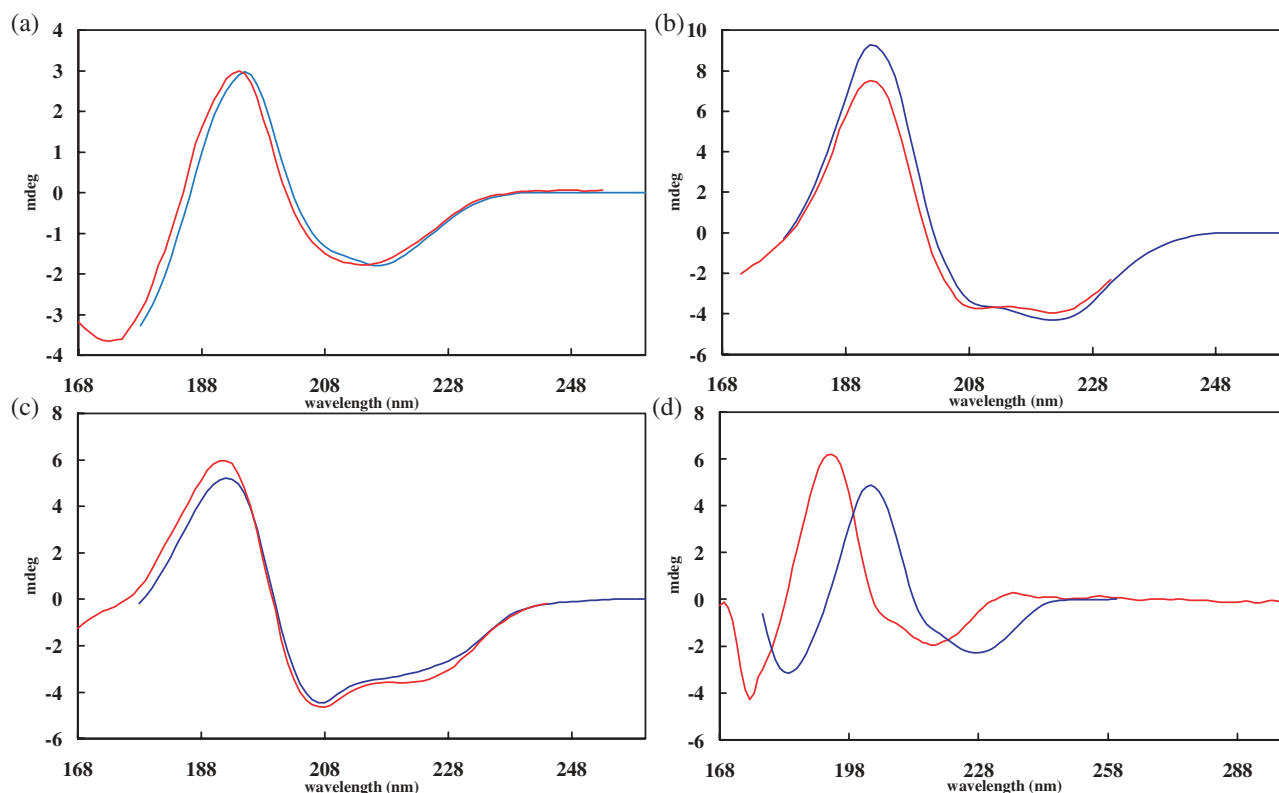
### 3.2 Wavelength range of CD reference spectra

CD spectra from different reference databases were collected over different wavelength ranges, as illustrated in Figures 1 and 2. The information content is directly proportional to the wavelength

range: the shorter the range the less the information available. The data from Johnson (1999) for example have a wavelength range 178–240 nm, while those of the CDDATA56 set, which incorporates the Johnson data, are only over 190–240 nm as other contributing groups collected over a shorter range. This reduction in range represents a significant difference in the available information content, decreasing rather than increasing the number of secondary structure components that can be derived from the data (Wallace and Janes, 2001). In addition, none of the current reference sets covers the range obtainable by SRCD, down to ~160 nm, and any new reference database would need to address this current shortcoming.

### 3.3 Quality of X-ray structures in the reference databases

The majority of the X-ray structures used in current reference databases were taken from the limited number available in the PDB (Bernstein *et al.*, 1977; Berman *et al.*, 2000) in the early 1980s, and some are from earlier. This has an inherent and serious weakness associated with it. Many of these structures were determined long before any systematic refinement protocols, checking and validation programs like PROCHECK (Laskowski *et al.*, 1993) were available.



**Fig. 2.** (a)–(c) CD spectra of three proteins from current reference database sets in CDPRO (blue) in comparison with the spectra of the same protein from a set now not used (Brahms and Brahms, 1980) (red). The proteins are (a) prealbumin, (b) lactate dehydrogenase and (c) lysozyme. The spectra in (d) are of  $\gamma$ -crystallin from the reference database sets used in CDPRO (blue) and SRCD (red) spectral data recorded by Evans *et al.* (2004).

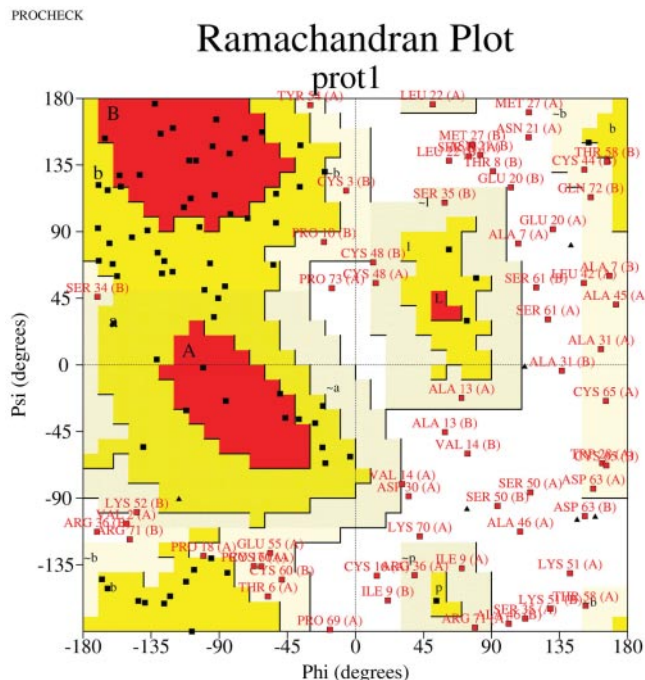
With limited validation, to a lesser or greater extent flaws do exist in a number of these structures which went undetected at that time. However, they are still used either within a stand-alone set or as part of an amalgamation set for determining secondary structure content from a CD spectrum. Data on some of these reference set proteins are presented in Tables 1 and 2 as illustrative examples. The Tables give the PDB code for the given protein at the time of their database inception, CATH database code, resolution of the structure, percentage secondary structure content, percentage residues missing (undetermined) in the structure and percentage residues in the fully, additionally and generously allowed and disallowed regions of the Ramachandran plot, as defined by PROCHECK. Another issue is that it is assumed that protein crystal and solution structures are the same despite the environments being markedly different. Any structural differences that might result from different conditions, e.g. concentrations, salts, pH, etc., could also compound the database inaccuracies in secondary structure determination from CD data.

### 3.4 Ramachandran plot quality of structures

For the Ramachandran plot, PROCHECK defines a threshold for well-resolved, accurate structures as >90% of their residues being located in the most favoured region, and this is flagged should the value fall below this level. If the value <80% then a double flag is issued to draw attention to potentially more serious problems within the structure. Table 1 is a reference dataset from Pancoska *et al.* (1995), and some members of this set are now used in amalgamation

sets in CDPRO. All 23 (100%) proteins of this set are under the 90% threshold for the most favoured region. Additionally, 13 of 23 structures (57%) of these proteins are under the 80% threshold. Of these 5 proteins are now used in amalgamated datasets (2cga, 4adh, 1ca2, 2grs and 1rhd) of which 3 (60%) are below the 80% threshold. In Table 2, the *Exp32* set, 22 of 32 structures (69%) have less than the 90% threshold for residues in the most favoured region. Of these, 8 (25%) are doubly flagged, as being <80%, substantially less than optimal. Failing to reach these thresholds indicates there may be some degree of error in their determined conformations, which means that secondary structure contents derived from them must also be in error. Many of these structures would not be publishable today given the validation procedures now employed. Figures 3 and 4 illustrate some of the problems associated with the structures comprising the reference databases. Here, only 14.8 and 59.9% of residues are located in the most favoured region of the Ramachandran plot. Indeed, in the first case, 24.6% of the residues are in the disallowed region of the plot, indicating a larger number of residues wholly incorrect than correct.

X-ray structures solved at low resolution can potentially have regions where errors arise from an inability to follow accurately the electron density. Some structures with less than optimal resolution are in the reference sets. In Table 2 the five structures with the lowest percentage amino acids in the fully allowed region are all solved at a resolution of 2.5 Å or worse. Clearly, the more incorrect the conformation the more incorrect the percentages of secondary structure types derived, and this questions their reliability for use



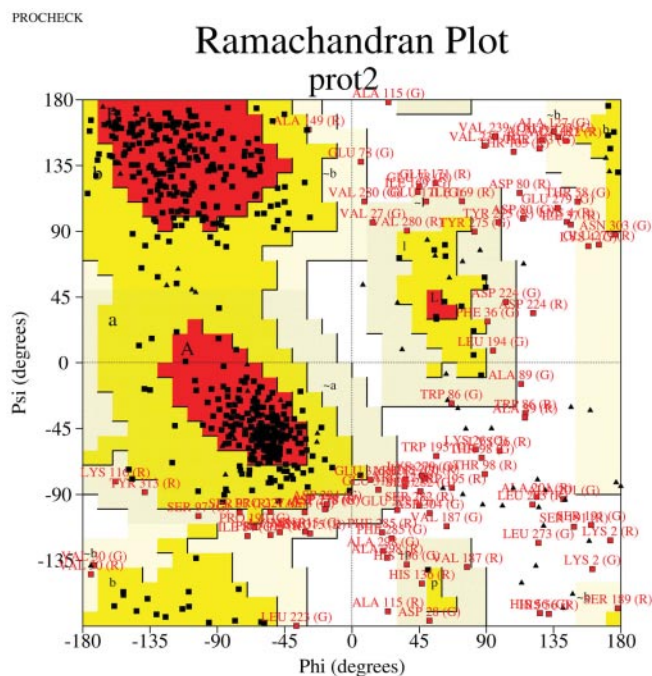
**Fig. 3.** A Ramachandran plot output, modified from PROCHECK (Laskowski *et al.*, 1993) of 'Prot1', a protein used in current reference databases. The areas marked are fully (red), additionally (yellow), generously (fawn) allowed and disallowed (white) regions for amino acid  $\phi/\psi$  angles. Shown in red lettering are those residues in the generously allowed and disallowed regions.

in the reference datasets. Only a few of the structures have serious mistakes, nevertheless, inclusion of small errors will introduce a degree of inaccuracy, which in turn will lead to erroneous calculation of secondary structure components for empirical determination from a CD spectrum.

### 3.5 Coverage of secondary structure space

The numbers of X-ray structures used within the reference databases are limited, especially so when compared with those in the PDB. For their optimal utilization it would be important for the sets accurately to reflect the secondary structures present in the PDB. Figure 5 shows as an example, a plot of the coverage of alpha helical against beta sheet content for (a) all protein structures in the PDB, (b) the *Exp32* set and (c) a theoretical idealized set also containing 32 proteins. The *Exp32* reference set does not cover the same secondary structure space as found in the PDB. Correlating the two sets of data, as described in the Methods, gives a value for  $r^2$  of 0.55. This is significantly lower than it could be and is again reflective of the limits imposed in having minimal numbers of protein structures available at the inception of these databases. By comparison, the  $r^2$  term is 0.95 for the idealized set, indicating the coverage of secondary structure space is more extensive, even for such a small dataset.

Increasing the number to that in CDDATA56 (not incorporating the membrane proteins, so this becomes a 43 protein set) gains little in the secondary structure coverage, the  $r^2$  value now becoming

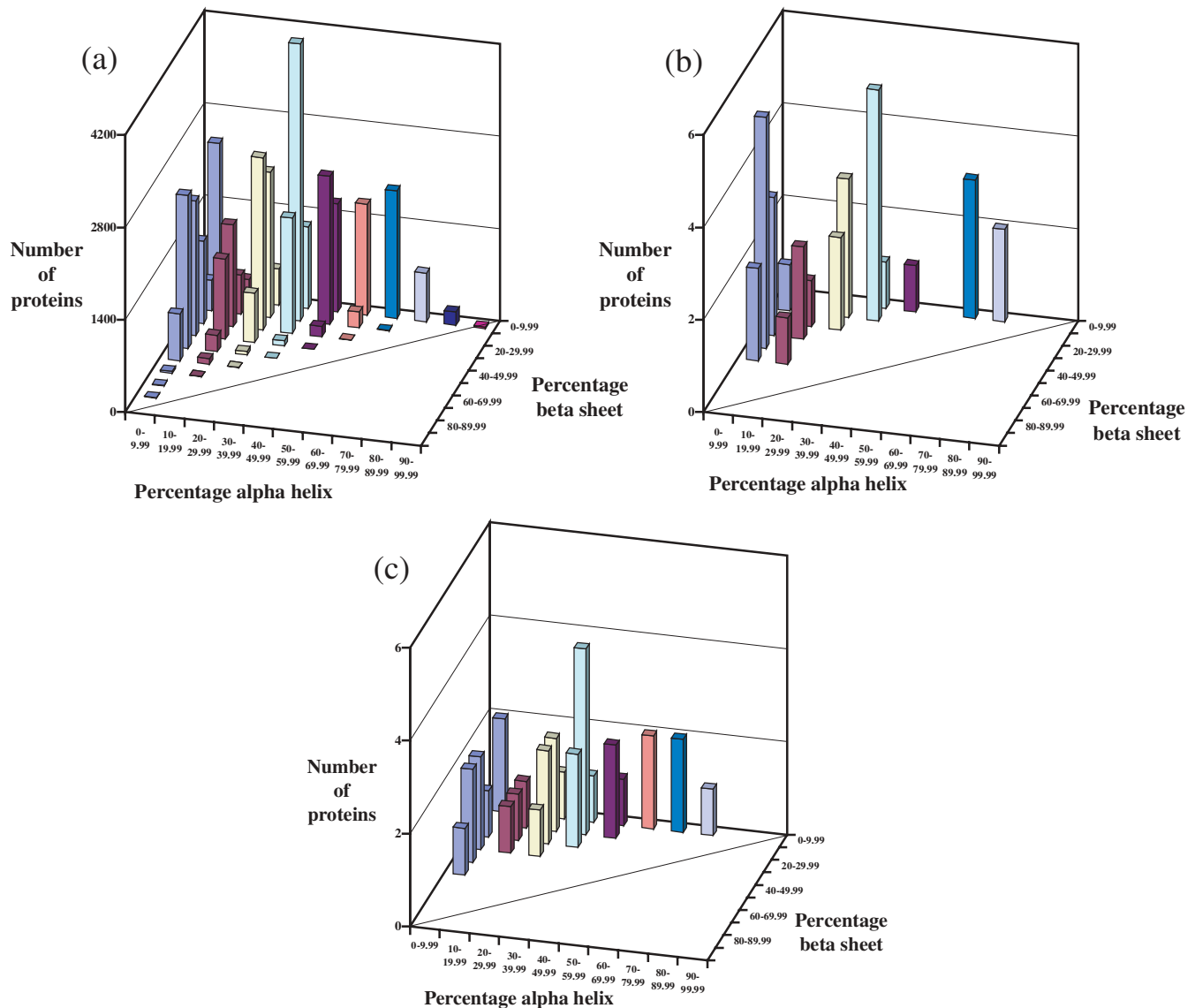


**Fig. 4.** A Ramachandran plot output (modified from PROCHECK) of 'Prot2', another protein in the current reference databases. Refer Figure 3 for a description of what is depicted.

0.62. However, increasing the number of available proteins in the reference set even allows improvement for an idealized set, the  $r^2$  becoming 0.97. Optimizing this correlation with the PDB is one way of ensuring that for a given number of proteins within a reference dataset, they will be as representative as is possible of the whole PDB.

### 3.6 Coverage of fold space

The maximum number of structures in the current reference databases used is 56, and this includes 13 membrane protein structures together with their related spectra. Whether this is an advisable move is a matter of debate (Wallace *et al.*, 2003; Sreerama and Woody, 2004). With SRCD sources the increased information content may enable analysing for the fold of a protein, therefore it is interesting to consider how broad based the coverage of fold space in the current reference databases is? This was not an issue at their inception, where interest lay only in the secondary structure content of the proteins, but it would be an important factor to consider for any new reference databases. Table 2 gives the CATH entry code for the structures in *Exp32*. Figure 6 shows these data in relation to those for all single-domain proteins in the PDB and to a hypothetical set of proteins with the same number as that in *Exp32* (31 here as 1 was unclassified in this set). While some of the more populated CATH classes are well represented in the *Exp32* set, others that should be present to ensure a good coverage of fold classes are clearly lacking. The  $r^2$  correlation coefficient is 0.81 for *Exp32* in relation to the entire PDB. In contrast, the hypothetical set of proteins with the same number of components as in *Exp32* has



**Fig. 5.** Plot of alpha helix against beta sheet content for (a) all proteins in the PDB compared with (b) those proteins found in the CD reference set *Exp32* used in SELCON3, and as part of most other reference databases, and (c) for an idealized reference dataset also with 32 proteins, with high correlation to the PDB, as described in the text. Beyond the diagonal base line represents possible regions for alpha/beta content.

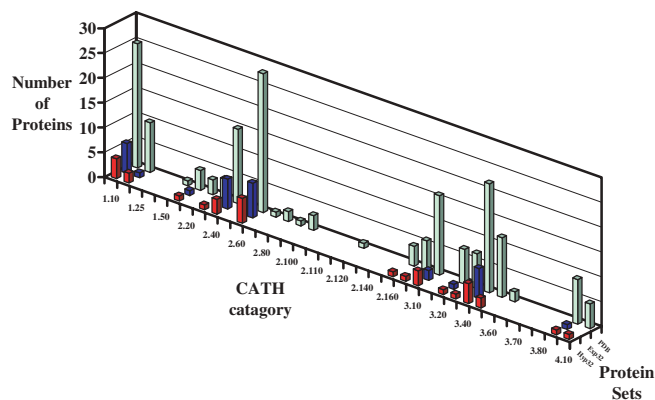
an  $r^2$  of 0.97 demonstrating that a better coverage of fold space is possible.

In Table 1, there are four structures present that are multi-domain proteins, containing two or more recognized CATH topologies in their structures, and two of these are used in amalgamated sets in CDPRO. Again, this was not an issue at the time of database inception. However, inclusion of multi-domain proteins into future reference database sets, which might be aimed at analysing for fold recognition (Wallace and Janes, 2001), would lead to difficulties in interpretation of such fold classes and so single-domain proteins would be the optimum to be used in such sets.

### 3.7 Completeness of structures

The *Exp32* set of structures in Table 2 has 11 (34%) of them that are incomplete, having undetermined regions, maybe from inherent

structural flexibility. These missing residues are predominantly lost from the N- and C-termini, and two have >5% of missing structure (5.43% for 1eri and 10.24% for 2pab) representing 15 and 18 residues, respectively. How missing structural content is accounted for in each of the databases is not always clear, especially because definitions pertinent to secondary structure features are also in their infancy and so totals counted for such features are sometimes not the same as those from current calculations. The main method assumes missing residues can be considered as ‘other’ (previously referred to as ‘random coil’) and thus adds them to that component’s total. This is feasible only if the number of missing residues is insufficient to form any type of secondary structure component. Highly flexible regions with proportionately larger numbers of missing residues might prevent determination of other more structured areas containing secondary structure



**Fig. 6.** Population of CATH fold space for single-domain proteins in the PDB (light blue) in comparison with data for the *Exp32* set (blue) and with a hypothetical set of proteins, labelled *Hyp32* (red) with the same number of proteins as that of the *Exp32* set.

components that would be missed as a result. Another method might be to ignore the missing content and to take the known portion of determined structure as being the total content. Whichever of these two methods were to be used, neither is valid as it is making unsupported assumptions in one case and introducing direct errors in the other, and hence both are unsatisfactory. Indeed, there is no satisfactory answer for dealing with missing residues in protein structures being used for a CD reference dataset other than to use only 'complete' proteins, structures whose entire length of chain is resolvable. It should be put into perspective, however, that at the time of inception of the databases many limitations hampered the selection of structures.

## 4 CONCLUSIONS

Despite the internal errors associated with the CD spectra and X-ray structures found in many reference databases, reasonably accurate values for secondary structure content of proteins can be determined from CD data. This is particularly true for mainly alpha helical proteins, likely due to the lack of variance in the geometry of this secondary structure component in different proteins. Accurate secondary structure determinations for many unknown proteins are also possible because the reference databases contain several of the most popular protein folds. Determination of beta sheet-containing proteins is usually less accurate, due to both the lesser intensity of CD signal from this component relative to that from alpha helices and the greater diversity of topologies of such a component found within proteins. Other less common secondary structural components, such as  $3_{10}$  and PPII helices, also tend to be less accurately determined because of their limited representation within the reference databases.

CD spectroscopy is used to determine the secondary structure content of proteins and many excellent mathematical approaches have been developed for this procedure. All rely on reference databases to obtain accurate values for calculating this content, but the databases themselves must have minimal errors otherwise this accuracy could be compromised. From the analyses presented there are some problems associated with the current CD reference databases, both with the CD spectra and with the X-ray structures

used. Some of the CD spectra are potentially erroneous representations of the referenced protein, with restrictions in their wavelength range covered, and many of the X-ray structures are not of the highest quality because of limitations in structure validation procedures. Also the structures are limited in their range of secondary structure types represented and coverage of secondary structure space. These analyses suggest that there is an urgent need to create a new, more comprehensive CD reference database containing cross-validated CD spectra collected and cross-checked on a number of different cCD spectrophotometers and SRCD beamlines to ensure machine independence. These should be for proteins with X-ray structures that have a broad coverage both of different secondary structure types and of secondary structure space, whose quality has been assured by the many available validation programs. Additionally, with the number of SRCD facilities increasing worldwide and improvements in cCD machine optics, any new reference database should extend to lower wavelength limits to enable analyses of these extra data. In summary, the recommendations for the content of a future reference database would be as follows: to contain ~80 proteins; with complete X-ray structures (i.e. those with no missing residues) whose quality has been confirmed by programs such as PROCHECK; a broad range of secondary structure and fold types represented; created with SRCD spectra to achieve low wavelength data (at least 170 nm), matching the CD data to the protein organism/sequence of the X-ray structure; and with full calibration/validation of these CD spectra. Such a database would enhance the quality and accuracy of secondary structure component determination, ensuring that CD spectroscopy remains a very powerful technique.

## ACKNOWLEDGEMENTS

I thank Prof. B. A. Wallace for many useful discussions. I thank Prof Jon B. Applequist for the CD data from some of the database sets, and Paul Evans and Dr Christine Slingsby for the SRCD spectral data for  $\gamma$ -crystallin. I also thank Dr Alison Cuff for provision of the single-domain proteins CATH data. This work was supported by a BBSRC grant (B19312).

*Conflict of Interest:* none declared.

## REFERENCES

- Berman, H.M. *et al.* (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.
- Bernstein, F.C. *et al.* (1977) The protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.*, **112**, 535–542.
- Bolotina, I.A. *et al.* (1980a) Determination of the secondary structure of proteins from the circular-dichroism spectra. 1. Protein reference spectra for alpha structure, beta structure and irregular structure. *Mol. Biol.*, **14**, 701–709.
- Bolotina, I.A. *et al.* (1980b) Determination of the secondary structure of proteins from the circular-dichroism spectra. 2. Consideration of the contribution of beta-bends. *Mol. Biol.*, **14**, 709–715.
- Brahms, S. and Brahms, J. (1980) Determination of protein secondary structure in solution by vacuum ultraviolet circular dichroism. *J. Mol. Biol.*, **138**, 149–178.
- Chang, T.C. *et al.* (1978) Circular dichroic analysis of protein conformation: inclusion of beta-turns. *Anal. Biochem.*, **91**, 13–31.
- 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.
- Compton, L.A. and Johnson, W.C., Jr (1986) Analysis of protein circular dichroism spectra for secondary structure using a simple matrix multiplication. *Anal. Biochem.*, **155**, 155–167.

- Evans, P. *et al.* (2004) The P23T cataract mutation causes loss of solubility of folded gammaD-crystallin. *J. Mol. Biol.*, **343**, 435–444.
- Hennessey, J.P., Jr and Johnson, W.C., Jr (1981) Information content in the circular dichroism of proteins. *Biochemistry*, **20**, 1085–1094.
- 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.
- Laskowski, R.A. *et al.* (1993) PROCHECK—a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.*, **26**, 283–291.
- Lobley, A. and Wallace, B.A. (2001) DICHROWEB: a website for the analysis of protein secondary structure from circular dichroism spectra. *Biophysical J.*, **80**, 373a.
- Lobley, A. *et al.* (2002) DICHROWEB: an interactive website for the analysis of protein secondary structure from circular dichroism spectra. *Bioinformatics*, **18**, 211–212.
- Miles, A.J. *et al.* (2003) Calibration and standardisation of synchrotron radiation circular dichroism and conventional circular dichroism spectrophotometers. *Spectroscopy*, **17**, 653–661.
- 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.
- Orengo, C.A. *et al.* (1997) CATH—a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.
- 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.
- Pancoska, P. *et al.* (1992) Relationships between secondary structure fractions for globular proteins. Neural network analyses of crystallographic datasets. *Biochemistry*, **31**, 10250–10257.
- Pancoska, P. *et al.* (1995) Comparison of and limits of accuracy for statistical analyses of vibrational and electronic circular dichroism spectra in terms of correlations to and predictions of protein secondary structure. *Protein Sci.*, **4**, 1384–1401.
- Pearl, F.M. *et al.* (2000) Assigning genomic sequences to CATH. *Nucleic Acids Res.*, **28**, 277–282.
- Pearl, F. *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**, D247–D251.
- Provencher, S.W. and Glöckner, J. (1981) Estimation of globular protein secondary structure from circular dichroism. *Biochemistry*, **20**, 33–37.
- Sreerema, 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.
- Sreerema, 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.
- Sreerema, N. *et al.* (1999) Estimation of the number of helical and strand segments in proteins using CD spectroscopy. *Protein Sci.*, **8**, 370–380.
- Sreerema, N. *et al.* (2000) Estimation of protein secondary structure from circular dichroism spectra: inclusion of denatured proteins with native proteins in the analysis. *Anal. Biochem.*, **287**, 243–251.
- Sutherland, J.C. *et al.* (1980) Versatile spectrometer for experiments using synchrotron radiation at wavelengths greater than 100 nm. *Nucl. Instrum. Methods*, **172**, 195–199.
- 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.* (2005) *Proteins*, in press.
- 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.