BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 12 2011, pages 1630–1636
doi:10.1093/bioinformatics/btr234

Structural bioinformatics

Advance Access publication April 19, 2011

A reference dataset for the analyses of membrane protein secondary structures and transmembrane residues using circular dichroism spectroscopy

Ali Abdul-Gader†, Andrew John Miles† and B. A. Wallace*

Department of Crystallography, Institute of Structural and Molecular Biology, Birkbeck College, University of London, London WC1E 7HX, UK

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Empirical analyses of protein secondary structures based on circular dichroism (CD) and synchrotron radiation circular dichroism (SRCD) spectroscopic data rely on the availability of reference datasets comprised of spectra of relevant proteins, whose crystal structures have been determined. Datasets comprised of only soluble proteins have not proven suitable for analysing the spectra of membrane proteins.

Results: A new reference dataset, MP180, has been created containing the spectra of 30 membrane proteins encompassing the secondary structure and fold space covered by all known membrane protein structures. In addition a mixed soluble and membrane protein dataset, SMP180, has been created, which includes 98 soluble protein spectra (SP) plus the MP180 spectra. Calculations of both membrane and soluble protein secondary structures using SMP180 are significantly improved with respect to those produced, using soluble protein-only datasets. The SMP180 dataset also enables determination of the percentage of transmembrane residues, thus enhancing the information previously obtainable from CD spectroscopy.

Availability and Implementation: Reference dataset online at the DichroWeb analysis server (http://dichroweb.cryst.bbk.ac.uk); individual protein spectra in the Protein Circular Dichroism Data Bank (http://pcddb.cryst.bbk.ac.uk).

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

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

Received on February 16, 2011; revised on April 7, 2011; accepted on April 9, 2011

1 INTRODUCTION

Circular dichroism (CD) spectroscopy is an important method in structural biology with applications that include determination of protein secondary structure content, detection of protein folding and unfolding, protein stability, formation of macromolecular complexes and characterisation of ligand–protein interactions. Synchrotron radiation circular dichroism (SRCD) instruments have enhanced the technique by producing data to lower wavelengths with higher information contents and improved signal-to-noise levels relative to those obtained using conventional bench-top instruments.

The use of CD as an analytical tool for protein secondary structure analyses relies on empirical algorithms (Chen and Yang, 1971; Brahms and Brahms, 1979; Hennessey and Johnson, 1981; Wallace and Teeters, 1987; Pancoska and Keiderling, 1991; Johnson, 1999; Sreerama and Woody, 2000) and associated spectral reference datasets derived from proteins with known structures. The accuracies of such analyses are dependant upon the range of structural and spectral characteristics of the constituent proteins in the reference dataset and how representative those proteins are of the types of structures present in the proteins being analysed (Janes, 2005; Whitmore and Wallace, 2008).

Most reference databases were developed sometime ago for the analysis of soluble proteins (Johnson, 1999; Sreerama and Woody, 2000; Sreerama *et al.*, 2000). More recently, a large reference dataset of soluble proteins, designated SP175 consisting of 71 soluble protein spectra (SP) with a low wavelength limit of 175 nm [achieved using an SRCD instrument], was created based on bioinformatics considerations (Lees *et al.*, 2006a). It produces enhanced predictive accuracies for CD data in comparison with the previous reference datasets as a result of its extensive coverage of secondary structure and fold space.

Membrane proteins make up approximately one-quarter to one-third of the proteins encoded by the currently sequenced genomes and they carry out important cellular functions, making them major targets for pharmaceutical drug development. However compared to soluble proteins, membrane proteins are under-represented in the Protein Data Bank (PDB) (Berman *et al.*, 2007), because they are difficult to crystallise and they tend to be too large for NMR studies; therefore, developing other methods for their structural characterisation is important.

Analyses of membrane protein CD and SRCD spectra using reference datasets based solely on soluble proteins generally do not yield predictions as reliable as those for soluble proteins (Wallace *et al.*, 2003). It is likely that this is because membrane proteins are embedded in environments of low dielectric constant, either detergent micelles or lipid vesicles, which can influence the relative ground and excited states of the electronic transitions of the peptide backbones relative to those in aqueous solutions (Cascio and Wallace, 1995; Chen and Wallace, 1997). It has been suggested that the inclusion of membrane proteins in a soluble protein dataset

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

© The Author 2011. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

will enhance its predictive performance when analysing membrane proteins (Wallace *et al.*, 2003; Sreerama and Woody, 2004), but until now the number of available membrane protein spectra with cognate high quality crystal structures has been very limited.

In this study, we created a reference dataset designated MP180 consisting of 30 SRCD spectra of membrane proteins whose crystal structures are known, and which were examined under their crystallisation conditions. A larger dataset containing both membrane and soluble proteins, designated SMP180, was also produced. It contains the proteins from MP180 and a further 98 soluble protein spectra [including those in SP175 (Lees *et al.*, 2006a)]. SMP180 was specifically created to improve the analyses of membrane proteins with large ecto-domains. SMP180 not only produces more accurate predictions of membrane protein secondary structures than can be achieved using soluble-only protein datasets; it also produces improved secondary structure analyses of soluble proteins, especially for their beta-sheet contents.

2 METHODS

2.1 Criteria for protein selection

A list of unique membrane protein crystal structures was obtained by cross-referencing the 'The Membrane Proteins of Known 3D Structure' web site (accession date 07/06/10) (http://blanco.biomol.uci.edu/Membrane_Proteins_xtal.html) (White, 2004) with the PISCES server (http://dunbrack.fccc.edu/Guoli/PISCES_InputB.php) (Wang and Dunbrack, 2003), which defines unique polypeptides, using a $>30\%$ sequence identity cutoff; this list is referred to as the all membrane proteins (AMP) dataset. Only proteins whose structures were determined by X-ray crystallography to a resolution of 3.5 Å or better were included. It is comprised of 183 proteins and includes 286 polypeptide chains from a starting list of 613 proteins and 2685 chains. It was used to evaluate the representative nature of the membrane proteins used in this study.

Membrane protein targets for the MP180 dataset were selected on the basis of coverage of secondary structure space from alpha/beta plots (Lees *et al.*, 2006a) and fold space based on CATH architecture classifications (Orengo *et al.*, 1997), as well as availability. Coordinate files for the selected membrane proteins were obtained from the PDB. Where possible, the cognate PDB file for the structure produced by the lab that provided the sample was used. If there was more than one structure available for that protein, then the PDB file was selected based, in order, on the following criteria: (i) highest resolution; (ii) lowest R factor; and (iii) highest overall G-factor score as defined by PROCHECK (Laskowski *et al.*, 1993).

The Define Secondary Structure of Proteins (DSSP) algorithm (Kabsch and Sander, 1983) was used to calculate secondary structure contents from the PDB files via the 2Struc web server (Klose *et al.*, 2010). The 'principal' secondary structure assignments used were: helices [alpha helix (H)+3_{10} helix (G)], β-sheets (E) and other (O) (which includes all remaining secondary structure types). The 'alternate' secondary structure assignments (Sreerama and Woody, 2000) of regular alpha helix (α_R), distorted alpha helix (α_D), regular beta sheet (β_R), distorted beta sheet (β_D), turn and unordered (U) were also used. The α_D fraction includes two residues from both ends of each helix and any helix that is <4 residues in length. The α_R fraction consists of all remaining helical residues. The β_D fraction includes one residue from both ends of each β-sheet, and all residues in a β-sheet that is ≥ 2 residues in length. The β_R fraction includes all remaining β-sheet residues. The turn class is created by combining the DSSP turn (T) and bend (S) classifications. Residues with missing backbone electron density and all remaining residues not included in any of the above classes were assigned to the unordered fraction.

To assign the transmembrane (TM) fractions, the topology prediction web site MEMSAT (Jones, 2007) was used for the alpha helical-rich membrane proteins, and the transmembrane beta-barrel prediction server PROFtmb (Bigelow *et al.*, 2004) for the beta sheet-rich membrane proteins. The transmembrane helix (TMh), extramembrane helix (EMh), transmembrane sheet (TMβ) and extramembrane sheet (EMβ) fractions were separately defined in order to determine if it would be possible to discriminate between secondary structures found in transmembrane and ecto-regions of a protein. The TMh fraction included residues from membrane spanning helices. All remaining helical residues were assigned to the EMh fraction. The TMβ fraction is comprised of membrane spanning β-sheet residues. The EMβ fraction contains all remaining β-sheet residues. Any residues not in the preceding classes were included in the non-helix, non-sheet (N) fraction.

Fold classifications were obtained from the CATH database (Orengo *et al.*, 1997). CATH classifications for unassigned proteins were derived from homologous chains acquired through the sequence navigator tool on the PDBj web site (Standley *et al.*, 2008). The correlation between the secondary structure space coverage by the MP180 and AMP datasets was determined by collecting the percentage helix versus sheet secondary structures into 100 'ten percentile' groups (e.g. 0–9.99, 10–19.99, etc.) of the alpha/beta plots, followed by calculation of Pearson's correlation coefficient, r, (Pearson, 1896) for these groups.

2.2 Materials

Protein samples were, in general, obtained from the labs that had determined the crystal structures (Supplementary Tables S1, S2), ensuring that the same proteins were used for the crystallography and spectroscopic measurements, and that the spectroscopic sample conditions were consistent with the crystallisation conditions (Supplementary Table S3). Five proteins (AmtB, GlpG, ClC-ec1, BtuCD and the NpSRII + NpHtRII complex) were dialysed before measuring their spectra in order to exchange the highly absorbing NaCl in their buffers with equivalent buffers containing NaF instead (Miles and Wallace, 2006). To remove any insoluble material, all samples were centrifuged at 12 800 g for 30 s just prior to use.

The protein concentration of each sample was determined from the absorbance at 280 nm, using a Nanodrop 1000 UV/Vis spectrophotometer. Measurements were repeated five times and averaged. The absorbance values obtained had variances of $\sim 1\%$. The calculated extinction coefficients were obtained from the protein sequence using the ProtParam tool on the EXPASY web site (http://www.expasy.ch/tools/protparam.html) (Gasteiger *et al.*, 2003). Concentrations used in the CD and SRCD measurements ranged from 1.0 to 20.0 mg/ml. In general, the beta sheet-rich proteins required higher concentrations as their spectral magnitudes tend to be smaller than those of alpha helical-rich proteins. Sample and spectroscopic conditions are listed in the Supplementary Material (Supplementary Table S3).

2.3 Spectroscopic measurements

SRCD measurements were carried out on beamlines UV1 and CD1 at ISA (Denmark), on stations 3.1 and CD12 at the Synchrotron Radiation Source (SRS) Daresbury (UK) and on the DISCO beamline at the Soleil Synchrotron (France). In a number of cases, replicate measurements were carried out on more than one beamline or on an Aviv 62ds conventional CD spectrophotometer to demonstrate cross-validity, reproducibility and that the samples were unaffected by transport to the beamlines.

SRCD spectra were measured at 20° C over a wavelength range from 280 to 170 nm. The following instrumental parameters were used at UV1 and CD1: 1 nm interval, 1 nm bandwidth, 2 s averaging time. Parameters used at CD12 and DISCO were identical, except the averaging times were 1 whilst at station 3.1 the parameters were 0.2 nm interval, 1 nm bandwidth and 3 s averaging. Spectra were measured on the Aviv CD instrument under the same conditions as at UV1/CD1, except over a wavelength range from 280 to 185 nm and using an averaging time of 3 s.

Measurements were carried out using quartz cylindrical demountable Suprasil (Hellma UK Ltd) cells with pathlengths ranging from 0.0015 to 0.0050 cm, or specially designed calcium fluoride cells (Wien and

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

Wallace, 2005) with pathlengths ranging from 0.0004 to 0.0022 cm. Accurate determinations of the cell pathlengths were achieved by employing the interference fringe method, as described previously (Miles *et al.*, 2005). Pathlengths were chosen to optimise the CD signal, whilst not exceeding the high tension (HT) or high voltage signal cutoffs of instruments, as described in Miles and Wallace (2006). Three spectra of each sample and of each baseline (buffer and detergent without protein) were measured and averaged. The first and third scans were compared to establish that no beam-induced denaturation of the protein had occurred (Wien *et al.*, 2005). The averaged baseline spectrum was subtracted from the averaged sample spectrum and then smoothed with a Savitsky–Golay filter (Savitsky and Golay, 1964). Data processing was carried out using CDtool software (Lees *et al.*, 2004) and spectra were plotted in $\Delta\varepsilon$ units, using the mean residue weight values listed in Supplementary Table S3. All instruments were calibrated for both spectral magnitude and optical rotation using camphour-10-sulfonic acid (CSA) and for wavelength with either benzene vapour or a certified holmium oxide glass filter (Hellma UK Ltd) (Miles *et al.*, 2003, 2005); in the case of the SRCD beamlines, fresh CSA calibration spectra were obtained following each beam injection.

2.4 Creation of reference datasets

Thirty membrane protein spectra were used to create the MP180 (meaning: Membrane Proteins, low wavelength cutoff 180 nm) reference dataset. The combined soluble and membrane protein dataset, referred to as SMP180, was created by combining the MP180 dataset with the SP180 dataset, a modified version of the original SP175 dataset (Lees *et al.*, 2006a), where the 71 protein spectra were truncated to 180 nm to produce a wavelength range consistent with the membrane protein spectra ($SP175_{t180}$), and supplemented with an additional 27 soluble proteins (Supplementary Table S4).

Following publication, the SMP180 reference dataset will be made available as a reference dataset option on the DichroWeb (Whitmore and Wallace, 2008) analysis server (http://dichroweb.cryst.bbk.ac.uk/). In addition, as we have done for the SP175 dataset, the individual spectra will be available in the Protein Circular Dichroism Data Bank (PCDDB) (http://pcddb.cryst.bbk.ac.uk) (Whitmore *et al.*, 2010, 2011), indicated by the keywords 'MP180' and 'SMP180'.

To determine the accuracy of the secondary structure calculations, the 'leave one out' cross-validation method was employed, as previously described (Lees *et al.*, 2006a). In the predictive performance analyses, likewise the cognate membrane protein spectrum was removed from the SMP180 dataset. The analyses were carried out in MATLAB using SELMAT3 (Lees *et al.*, 2006b), a modified version of the SELCON3 algorithm (Sreerama and Woody, 2000), which enabled larger numbers of reference spectra and relaxed constraints to be used. Validation performances were quantified using Pearson's correlation coefficient (r), the root mean squared deviation (δ) and the ζ parameter (Oberg *et al.*, 2004), which represents how much better than random a prediction is and can be determined by dividing the δ value by the population standard deviation (σ_x). Principal component analyses were undertaken using the Csel2 implementation in CDtools (Lees *et al.*, 2004), which is based on the singular value deconvolution method of Hennessey and Johnson (1981). The spectra of the mostly helical proteins (as defined by their CATH classification) from the MP180 and SP180 datasets were separately analysed and the peak positions of the first basis sets calculated for each were compared.

3 RESULTS

3.1 Spectral data

The spectra of 35 membrane proteins were initially measured and of these, 30 were deemed suitable for inclusion in the MP180 dataset (Supplementary Table S1) based on the criteria of low noise levels ($<0.3\Delta\varepsilon$ between positive and negative peaks between 250 and 270 nm), high reproducibility ($<2\%$ difference between the values

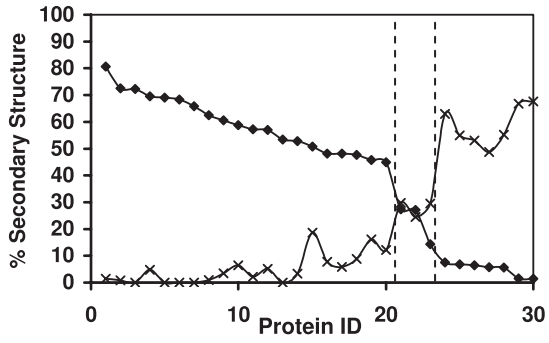

Fig. 1. Percentage secondary structure content of the MP180 proteins in the same ID order as in Supplementary Table S1: helical content [(solid diamonds) and beta-sheet content (crosses)]. Mostly helical proteins are to the left of the first vertical line, mostly sheet to the right of the second vertical line. Mixed alpha/beta proteins lie between the vertical lines.

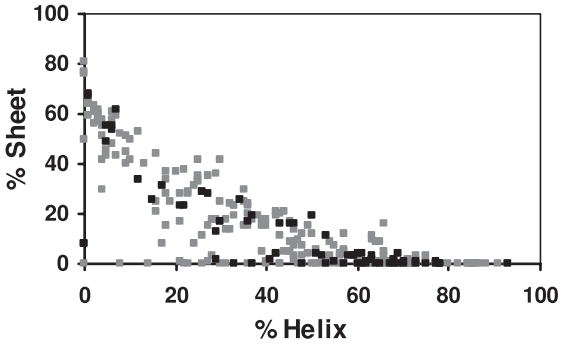

Fig. 2. Secondary structure space coverage (percentage of helix versus percentage of sheet) for the MP180 (black squares) and AMP (grey squares) polypeptides.

at ~190 nm for repeat scans) and a low wavelength cutoff of at least 180 nm (Supplementary Fig. S1).

The MP180 proteins form three groups: mainly alpha helical (>40% helical content), mainly beta sheet (>40% sheet content) and mixed alpha/beta (Fig. 1). There is a good overlap between the coverage of the AMP and the MP180 datasets (Fig. 2), with a correlation coefficient, r, of 0.91.

3.2 Characteristics of the MP180 Proteins

The MP180 dataset covers proteins with a wide range of structures and functions, making it a representative sample of the types of the membrane proteins whose crystal structures were known at the time of writing. Eighteen proteins are multimeric, 8 of which are homo-multimers and 12 are monomeric. The functional roles of the MP180 proteins include host cell recognition by bacterial pathogens, G-protein coupled receptors, electron transport within photosynthetic and respiratory pathways, flux of ions and toxic substances and active uptake of nutrients and ligands. The majority of the MP180 samples was from prokaryotic sources, with four eukaryotic proteins included.

Tryptophan residues are considered to play a significant role in anchoring a membrane protein within the lipid bilayer, so the

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

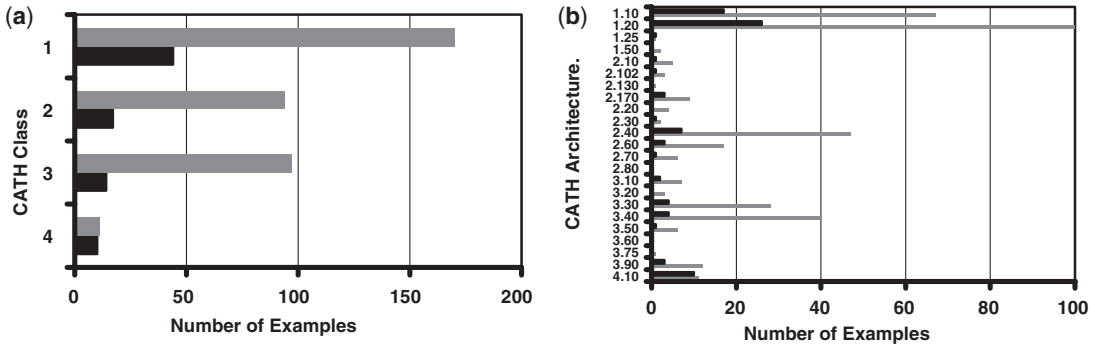

Fig. 3. CATH distributions for the MP180 (black) and AMP (grey) polypeptides: (**a**) classes and (**b**) architectures.

coverage of their content in the dataset was checked. The MP180 proteins contain between 1% and 5% tryptophan residues, with an average of 2%. Correspondingly, tryptophan residues are, on average, 2% of residues in the AMP proteins.

The percentage of residues in transmembrane segments in the MP180 proteins range from 5% to 78%, with an average of 41%. The beta sheet-rich proteins have a slightly higher average TM composition (52%), in contrast to 42% for the helix-rich proteins and 11% for the mixed alpha/beta proteins. The TM residue content for the AMP dataset ranges from 2% to 79%, with an average of 31%. Again the beta sheet-rich ones have a slightly larger value (37%) in comparison with the helix-rich (36%) and mixed alpha/beta chains (10%).

3.3 Class and fold types

The CATH class distribution for the individual MP180 polypeptides (67 chains with 85 domains) includes a larger number of mainly alpha helical domains (Class 1) than mainly beta sheet domains (Class 2) (Fig. 3a). However, there is a greater structural diversity within Class 2 as it contains seven different CATH architectures (Fig. 3b). The mixed α/β class (Class 3) is represented by five architectures, whilst Class 4 with 'few secondary structures' is represented by one architecture. Overall, the MP180 dataset contains four CATH classes and sixteen CATH architectures.

Similarly, Class 1 domains dominate the AMP dataset, with 170 examples. The Class 2 domains contain nine architectures, which again reflect their larger structural diversity. Seven architecture types are represented in Class 3 and a single architecture in Class 4. Of the twenty one AMP architectures, sixteen are represented by the MP180 polypeptide chains, hence demonstrating a good representation of CATH architectures by the MP180 dataset.

3.4 Analyses of validation and dataset predictive abilities

The self-consistency of each dataset was assessed using 'leave one out' cross-validation (Tables 1 and 2), where each spectrum is removed in turn and analysed against the rest. Cross-validation is one guide to the predictive accuracy of the dataset, which can be evaluated by the r and δ values obtained along with the ζ value. This process was carried out with both the 'principal' (Table 1) and 'alternate' (Table 2) secondary structure classification schemes. For the principal scheme, all datasets exhibited high correlations with the best values for helical secondary structures, and with SMP180 producing the best overall results. For the alternate scheme, all datasets gave high correlations for all secondary structural types, except the analyses for the turn category (probably because there

Table 1. Self-consistency check: cross-validated performances of the MP180, $SP175_{t180}$ and SMP180 datasets using the 'principal' secondary structure categories: helix ($H+G$), sheet (E) and 'other' (O)

Dataset	Parameter	Secondary structure type		
		helix ($H+G$)	β-sheet (E)	O
MP180	r	0.909	0.849	0.689
	δ	0.105	0.124	0.058
	ζ	2.388	1.870	1.354
$SP175_{t180}$	r	0.939	0.838	0.749
	δ	0.071	0.087	0.066
	ζ	2.907	1.832	1.502
SMP180	r	0.942	0.864	0.749
	δ	0.075	0.089	0.068
	ζ	2.979	1.979	1.533

Table 2. Self-consistency check: cross-validated performances of the MP180, $SP175_{t180}$ and SMP180 datasets using the 'alternative' secondary structure categories: regular helix (α_R), distorted helix (α_D), regular β-sheet (β_R), distorted β-sheet (β_D), turns and unordered (U)

Dataset	Parameter	Secondary structure type					
		α_R	α_D	β_R	β_D	Turn	U
MP180	r	0.911	0.768	0.843	0.869	−0.099	0.707
	δ	0.085	0.039	0.106	0.021	0.034	0.056
	ζ	2.415	1.533	1.841	2.004	0.778	1.376
$SP175_{t180}$	r	0.948	0.779	0.775	0.882	0.322	0.721
	δ	0.050	0.037	0.075	0.023	0.041	0.059
	ζ	3.143	1.592	1.582	2.106	1.038	1.433
SMP180	r	0.946	0.747	0.810	0.858	0.310	0.729
	δ	0.058	0.040	0.080	0.026	0.038	0.063
	ζ	3.070	1.503	1.702	1.949	1.033	1.452

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

Table 3. The predictive performances of the $SP175_{t180}$ and SMP180 datasets for the MP180 membrane proteins, and of SMP180 for the $SP175_{t180}$ soluble proteins, using the principal secondary structure categories

Dataset	Test set	Parameter	Secondary structure type		
			helix ($H+G$)	β-sheet (E)	O
$SP175_{t180}$	MP180	r	0.920	0.923	0.465
		δ	0.125	0.114	0.093
		ζ	2.035	2.056	0.876
SMP180	MP180	r	0.931	0.937	0.595
		δ	0.118	0.092	0.067
		ζ	2.166	2.547	1.224
SMP180	$SP175_{t180}$	r	0.940	0.879	0.745
		δ	0.074	0.088	0.067
		ζ	2.838	1.824	1.487

Table 4. The predictive performances of the $SP175_{t180}$ and SMP180 datasets for the MP180 membrane proteins and of SMP180 for the $SP175t_{180}$ soluble proteins, using the alternative secondary structure categories

Dataset	Test set	Parameter	Secondary structure type					
			α_R	α_D	β_R	β_D	Turn	U
$SP175_{t180}$	MP180	r	0.862	0.770	0.890	0.805	0.173	0.434
		δ	0.102	0.049	0.118	0.030	0.028	0.074
		ζ	1.972	1.203	1.665	1.442	0.945	1.003
SMP180	MP180	r	0.928	0.871	0.940	0.896	0.227	0.642
		δ	0.077	0.032	0.090	0.021	0.030	0.060
		ζ	2.702	1.886	2.224	2.022	0.945	1.302
SMP180	$SP175_{t180}$	r	0.949	0.813	0.839	0.880	0.365	0.713
		δ	0.050	0.035	0.064	0.027	0.043	0.061
		ζ	3.145	1.693	1.852	1.820	0.991	1.405

were too few examples of turns in this dataset to be meaningful for self-validation), again with regular helices being the best defined in all datasets.

To test their predictive accuracies, the MP180 test set was tested against the soluble protein-only dataset $SP175_{t180}$ and the mixed soluble/membrane protein dataset SMP180 (again with leaving out the cognate proteins) with both classification schemes (Tables 3 and 4, lines 1 and 2). For the principal scheme (Table 3), all values were better using SMP180, with only the O class below the significance cutoff level ($\zeta < 1.0$) for this dataset.

More dramatic improvements were seen with the alternative scheme (Table 4) generally used in CD analyses, with all secondary structural types being much better predicted. Thus, it is clear that SMP180 should be the preferred reference dataset for analyses of membrane proteins. To then test the value of the mixed reference dataset for the analysis of soluble proteins, similar predictive cross-valuations were done for the soluble-only test system ($SP175_{t180}$) (Tables 3 and 4, line 3). SMP180 also produces excellent results for all secondary structures in both schemes, with the beta sheet structures in both being considerably better correlated than in the

Table 5. The transmembrane/extramembrane predictive performances of the SMP180 and $SP175_{t180}$ datasets for the MP180 proteins. The categories used are transmembrane helix (TMh), extramembranous helix (EMh), transmembrane β-sheet (TMβ), extramembranous β-sheet (EMβ) and non-helix, non-sheet (N)

Dataset	Test set	Parameter	Secondary structure type				
			TMh	EMh	TMβ	EMβ	N
SMP80	MP180	r	0.814	0.497	0.855	0.530	0.562
		δ	0.175	0.162	0.172	0.136	0.069
		ζ	1.347	0.632	1.293	0.684	1.179
$SP175_{t180}$	MP180	r	–	0.513	–	0.382	0.487
		δ	0.375	0.346	0.249	0.191	0.077
		ζ	0.627	0.296	0.890	0.487	1.054

$SP175_{t180}$ self cross-correlation test. This then suggests the SMP180 reference dataset should also be the reference dataset of choice for soluble protein analyses, especially for determining beta sheet contents.

The predictive performances of the SMP180 and $SP175_{t180}$ datasets were then tested for their ability to separately identify the secondary structural types in the trans-membrane and ecto-domains of a protein (Table 5), a novel structural feature that has not been previously possible to identify from membrane protein spectra. It was expected that this type of analysis might be successful, because of the different spectral characteristics of residues present in the aqueous and hydrophobic-exposed environments (Wallace *et al.*, 2003). The SMP180 dataset produced very good predictions for both helical and sheet types of transmembrane structures (Table 5, line 1), although the extra-membranous fractions were not well-predicted. As expected, the $SP175_{t180}$ dataset did not allow identification of either the trans- or extra-membranous fractions (Table 5, line 2) (indeed the blanks on the table arise because this dataset contains no examples of transmembrane segments, which effectively results in a divide by zero). The ability to separately detect the transmembrane fractions corresponds with the observation (Supplementary Fig. S2) that separate principal component analyses of the helix-rich proteins in the MP180 and $SP175_{t180}$ datasets indicate that all three peptide backbone peaks (the $n->\pi*$ and both $\pi->\pi*$ transitions, located at ~222, and 208 and 195 nm, respectively), are shifted by ~1 nm with respect to each other in the membrane and soluble protein spectra, consistent with previously seen solvent dielectric-related shifts (Cascio and Wallace, 1995; Chen and Wallace, 1997). Thus, the SMP180 reference dataset provides additional predictive power for identifying the proportion of helical and sheet residues that are found in transmembrane segments.

4 DISCUSSION

Accurate empirical analyses of membrane protein secondary structures from CD spectra have been challenging until now due to spectral differences between soluble and membrane proteins (Park *et al.*, 1992; Wallace *et al.*, 2003). Although CD reference datasets containing membrane proteins are available (Park *et al.*, 1992; Sreerama *et al.*, 2004) the spectra were produced before there were many high quality membrane protein structures in

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

the PDB and before there were good bioinformatics tools for defining fold space. Consequently they contain proteins of limited structural diversity. Now that there are a relatively large number of good-quality membrane protein crystal structures in the PDB, we have been able to create a bioinformatics-defined membrane protein reference dataset covering a broad structural range for the specific analyses of membrane protein spectra. The ability to obtain high-quality spectral data using the SRCD spectroscopy (higher signal-to-noise levels and lower wavelengths) contributed to the improved spectral data that could be included. Furthermore, for the accuracy and utility of the new reference dataset, it was important to establish that the spectra were obtained on well- and cross-calibrated instruments, that the magnitudes determined were precisely correct (based on having accurate measurements of protein concentration and cell pathlengths) and that the measurements were done in conditions comparable to those used for the crystal structure determinations. All spectra passed the Validichro validation tests in the PCDDB (Whitmore *et al.*, 2011). For these reasons, the spectra included in these datasets are considered 'Gold standard' and thus will be made publicly available separately via the PCDDB (http://pcddb.cryst.bbk.ac.uk). It is expected that these spectra will find further use in the development of other methodologies, notably machine learning algorithms (Andrade *et al.*, 1993) [which to date have not had membrane protein spectra in their training sets, and hence have not produced suitable results for membrane proteins (Wallace *et al.*, 2003)] or for comparisons with *ab initio* calculations. It is important to note that even though the data in this study were collected by SRCD, they were cross-checked with CD spectroscopy and all of the analyses employed 180 nm cutoff values, hence making them suitable for use in analyses of conventional CD data.

The ability of the newly created membrane protein dataset, MP180, and especially of the larger combined soluble and membrane protein dataset, SMP180, to accurately predict the secondary structure contents of membrane proteins was compared with that of a modified version of the best existing soluble-only protein dataset, $SP175_{t180}$ (Lees *et al.*, 2006a). Results from the analyses of membrane protein spectra using the enlarged combined SMP180 dataset surpass the performance of both MP180 and $SP175_{t180}$ datasets. This is likely to be because membrane proteins contain both transmembrane and extra-membranous domains, which are well-represented within the combined dataset.

The ability to differentiate between transmembrane and extra-membranous secondary structures is a novel feature of the SMP180 dataset and is an important new capability for analysing membrane proteins as it provides additional structural detail that will be useful for molecular modeling of unknown structures (see the example for a voltage-gated sodium channel, NaChBac, in the Supplemental Material).

Thus significant improvements have been achieved in membrane protein CD analyses by incorporating membrane proteins and soluble proteins into a single large CD reference dataset. The increased number and diversity of proteins in SMP180 have also resulted in improved analyses of soluble proteins, especially for the sheet secondary structure components.

The observed suitability of the SMP180 dataset for analysing membrane proteins may stem from the observed shifts, which occur in spectral peak positions in membrane protein spectra in comparison with those in soluble protein spectra with similar secondary structure content, as well as other differences at the supersecondary structure level. A likely source of the former spectral differences is the different dielectric constants of the hydrophobic environment surrounding membrane proteins and the aqueous environment surrounding the soluble proteins.

In summary, the new datasets provide excellent results for both soluble and membrane proteins in secondary structure analyses, and in the transmembrane discrimination analyses of membrane proteins, thereby providing novel information for the analysis and molecular modeling of membrane proteins. Hence, they will be included in the cadre of datasets available via the DichroWeb analysis server (Whitmore and Wallace, 2008).

ACKNOWLEDGEMENTS

We are very grateful to the labs listed in Supplementary Tables S1 and S4, which generously provided the membrane and soluble protein samples, respectively, used in this study. We thank the following colleagues, students and postdocs for help with SRCD data collection: Drs R.W. Janes, J.G. Lees, F. Wien, T. Stone, Mr Ben Woollett and the late Dr P. Evans. We thank Dr Lee Whitmore (Birkbeck College) for making the dataset available as an option in the DichroWeb server. We also thank the beamline scientists, Dr Soren Vronning Hoffman (ISA), Dr F. Wien (Soleil) and Dr David Clarke (SRS), for their help.

Funding: Biotechnology and Biological Sciences Research Council (grants to B.A.W.); a Biotechnology and Biological Sciences earmarked studentship (to A.A.-G.); the ISA (Denmark) and Soleil (France) synchrotrons (SRCD beamtime grants, to B.A.W.); the SRS Daresbury (UK) synchrotron (a Programme Mode Access grant, to B.A.W. and R.W. Janes, Queen Mary, University of London).

Conflict of Interest: none declared.

REFERENCES

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

Berman,H. *et al.* (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.*, **35**, D301–D303.

Bigelow,H.R. *et al.* (2004) Predicting transmembrane beta-barrels in proteomes. *Nucleic Acids Res.*, **32**, 2566–2577.

Brahms,S. and Brahms,J. (1979) Determination of protein secondary structure in solution by vacuum ultraviolet circular dichroism. *J. Mol. Biol.*, **138**, 149–178.

Cascio,M. and Wallace,B.A. (1995) Effects of local environment on the circular dichroism spectra of polypeptides. *Anal. Biochem.*, **227**, 90–100.

Chen,Y-C. and Wallace,B.A. (1997) Secondary solvent effects on the circular dichroism spectra of polypeptides: influence of polarisation effects on the far ultraviolet spectra of alamethicin. *Biophys. Chem.*, **65**, 65–74.

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.

Gasteiger,E. *et al.* (2003). ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res.,* **31**, 3784–3788.

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

Janes,R.W. (2005) Bioinformatics analyses of circular dichroism protein reference databases. *Bioinformatics*, **21**, 4230–4238.

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

Jones,D.T. (2007) Improving the accuracy of transmembrane protein topology prediction using evolutionary information. *Bioinformatics*, **23**, 538–544.

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

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

Klose,D.P. *et al.* (2010) 2Struc: the secondary structure server. *Bioinformatics*, **26**, 2624–2625.

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

Lees,J.G. *et al.* (2006b) Novel methods for secondary structure determination using low wavelength (VUV) circular dichroism spectroscopic data. *BMC Bioinformatics*, **7**, 507.

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.

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 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.

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.* (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.

Park,K. *et al.* (1992) Differentiation between transmembrane and peripheral helices by the deconvolution of circular dichroism spectra of membrane proteins. *Protein Sci.*, **1**, 1032–1049.

Pearson,K. (1896) Mathematical contributions to the theory of evolution. III. Regression, heredity and panmixia. *Philos. Trans. R. Soc. Lond A.*, **187**, 253–318.

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

Sreerama,N. and Woody,R.W. (2000) Estimation of protein secondary structure from CD spectra: comparison of CONTIN, SELCON and CDSSTR methods with an expanded reference set. *Anal. Biochem.*, **287**, 252–260.

Sreerama,N. and Woody,R.W. (2004) On the analysis of membrane protein circular dichroism spectra. *Protein Sci.*, **13**, 100–112.

Sreerama,N. *et al.* (2000) Estimation of protein secondary structure from CD spectra: inclusion of denatured proteins with native proteins in the analysis. *Anal. Biochem.*, **287**, 243–251.

Standley,D.M. *et al.* (2008) Protein structure databases with new web services for structural biology and biomedical research. *Brief. Bioinformatics*, **9**, 276–285.

Wallace,B.A. *et al.* (2003) Analyses of circular dichroism spectra of membrane proteins. *Protein Sci.*, **12**, 875–884.

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.

Wang,G. and Dunbrack,R.L.,Jr (2003) PISCES: a protein sequence culling server. *Bioinformatics*, **19**, 1589–1591.

White,S.H. (2004) The progress of membrane protein structure determination. *Protein Sci.*, **13**, 1948–1949.

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

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

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

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

Wien,F. *et al.* (2005) VUV irradiation effects on proteins in high flux synchrotron radiation circular dichroism (SRCD) spectroscopy. *J. Synch. Radiat.,* **12**, 517–523.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016