Protein Secondary Structure Analyses from Circular Dichroism Spectroscopy: Methods and Reference Databases

Lee Whitmore, B. A. Wallace
Department of Crystallography, Birkbeck College, University of London, London WC1E 7HX, UK

Received 9 August 2007; revised 13 September 2007; accepted 17 September 2007
Published online 25 September 2007 in Wiley InterScience (www.interscience.wiley.com). DOI 10.1002/bip.20853

ABSTRACT:

Circular dichroism (CD) spectroscopy has been a valuable method for the analysis of protein secondary structures for many years. With the advent of synchrotron radiation circular dichroism (SRCD) and improvements in instrumentation for conventional CD, lower wavelength data are obtainable and the information content of the spectra increased. In addition, new computation and bioinformatics methods have been developed and new reference databases have been created, which greatly improve and facilitate the analyses of CD spectra. This article discusses recent developments in the analysis of protein secondary structures, including features of the DICHROWEB analysis webserver. © 2007 Wiley Periodicals, Inc. Biopolymers 89: 392–400, 2008.

Keywords: circular dichroism spectroscopy; protein secondary structure; analyses; reference database; bioinformatics; synchrotron radiation circular dichroism (SRCD)

This article was originally published online as an accepted preprint. The "Published Online" date corresponds to the preprint version. You can request a copy of the preprint by emailing the Biopolymers editorial office at biopolymers@wiley.com

This article is dedicated to the memory of Elkan R. Blout, in whose lab BAW was first introduced to CD spectroscopy as a postdoc.

Correspondence to: B.A. Wallace; e-mail: ubcg25a@mail.cryst.bbk.ac.uk
Contract grant sponsor: Biotechnology and Biological Sciences Research Council

INTRODUCTION

Circular dichroism (CD) spectroscopy is a powerful method in structural biology that has been used to examine proteins, polypeptides, and peptide structures since the 1960s. Because the spectra of these molecules in the far ultraviolet (UV) regions are dominated by the $n\rightarrow\pi^*$ and $\pi\rightarrow\pi^*$ transitions of amide groups (Figure 1a), and are influenced by the geometries of the polypeptide backbones, their spectra are reflective of the different types of secondary structures (and thus the ϕ, ψ angles) present. Consequently, analyses have been developed to deconvolute the various contributions arising from the different types of secondary structures present in a single molecule, thereby providing information on the overall structure of that protein. Using reference spectra derived from proteins of known structure (i.e., whose crystal structures have been determined), a wide range of different empirical algorithms has been developed, which rely on the assumption of linear independence and additivity of different components in producing the net spectrum obtained. Early methods included simple linear and nonlinear least squares analyses based on "representative reference spectra" of different secondary structural types.[1] To compensate for the lack of exact solutions (due in large part to the presence in any given protein of structural types that vary from those in the standard reference spectra) constraints were introduced to either require the calculated fractions of all the secondary structure components to be nonnegative (i.e., a requirement that the results make physical sense since proteins cannot have negative amounts of a type of structure) or that they sum to a total of one (to force the result to account for the whole structure of the protein). An alternative method[2] that normalised the sum of the values obtained to a total of 1.0 had the advantage that it did not require a precise knowledge of protein concentration, a parameter required by the other methods. More sophisticated algorithms were eventually developed, which included singular value deconvolutions,[3] parameterised fits,[4] self-consistency,[5] convex constraints,[6]

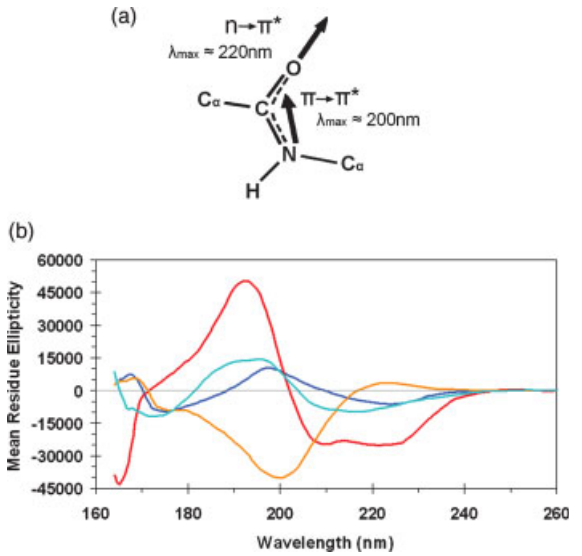

FIGURE 1 (a) Diagram of a peptide bond showing the orientation of the transition dipoles (as thick arrows) of the $n \rightarrow \pi^*$ and $\pi \rightarrow \pi^*$ transitions. (b) CD spectra of a mostly helical protein, myoglobin (in red), two mostly beta-sheet proteins, concanavalin A (blue) and beta-lactoglobulin (cyan), and a polyproline-rich protein, collagen (orange). It is clear that even though the two beta sheet proteins have virtually identical amounts of beta sheet present (46 and 45%, respectively), because they have very different folds (as indicated by their CATH classifications of 2.60.120.200 and 2.40.128.20, respectively), their spectral characteristics are very different. These spectra contain very low wavelength (VUV) data because they were obtained using SRCD. It can be seen that at higher wavelengths (above 200 nm) both sheet and helix structures produce negative peaks, with the magnitudes of the sheet spectra being much lower than those of the helical spectrum, but at low wavelengths the sheet and helical structures give rise to spectra of opposite signs; inclusion of such data substantially improves the analyses of the beta sheet components present in proteins.

matrix descriptor,[7] and neural networks.[8,9] A number of these methods have now been in standard use for more than twenty years. Recent developments[10] have been to employ new computational tools such as support vector machines, simultaneous partial least squares (SIMPLS), principal component regressions, or combinations of several such methods to improve and extend the analyses. Most all of the methods produce reasonable and consistent results.

In general the methods provide the most accurate results for helical secondary structures. This is because: (1) Helical structures tend to be very regular, having well-defined ϕ, ψ angles and thus produce very similar spectra. (2) The spectra of helical components (especially long stretches of helical amino acids) produce very intense CD signals (Figure 1b). Because beta-sheet structures tend to be more variable, with both parallel and antiparallel orientations of adjacent strands, and different twists, their ϕ, ψ angles vary considerably, as do their CD spectra (Figure 1b).[11] Furthermore, spectra of beta-sheet structures tend to be much less intense, with their negative peaks only about one-third the size of the negative peaks of an alpha-helix. One consequence of this is that when a protein contains a large amount of helix and small sheet content, the spectral contribution of the latter may be swamped out and hence the accuracy of the derived sheet content will be considerably lower. This can be mitigated against by including the very low wavelength vacuum ultraviolet (VUV) data obtainable using synchrotron radiation as a bright light source for the CD measurements (the technique then being known as synchrotron radiation circular dichroism (SRCD)); this is because the very low wavelength data for helices and sheets have opposite signs.[12] Turns, too have distinct spectra, but like other types of less common secondary structures, such as 3_{10} helices, the number of examples of each type of turn available in any given reference database may limit the accuracy of deconvolution methods. Other types of secondary or supersecondary structures that give rise to distinct spectra include polyproline II helices (Figure 1b) and coiled-coils. Finally, the remaining secondary structure, originally referred to as "random coil," but for which this nomenclature is actually inappropriate, since most such structures are neither random nor coil in nature, might better be classified as "other," that is, not canonical helix, sheet or turn. This type of secondary structure is often now also referred to as "irregular," "natively disordered" or "intrinsically disordered." However, it is not a single type of structure, but rather a grouping together of (in many cases well-defined) structures, which adopt a wide range of ϕ, ψ angles that are not those ϕ, ψ angles adopted by helix, sheets and turns. As a result, any attempt to accurately identify or quantitate them from a spectral deconvolution will be limited.

The most important variable that contributes to the success or failure of the different analyses is the reference database that is used. Obviously the wider the range of secondary structures (and ultimately protein folds) that are represented in the reference databases, the more accurate will be the result, regardless of which empirical analysis method is used. The first attempts at producing spectral examples of the various types of secondary structure utilised polylysine under different conditions to represent helical, sheet, and "random" conformations.[13] Naturally, polylysine was not a perfect example of 100% of any of these structures, but in retrospect it can be seen they provided a reasonably good first approximation. Later reference databases derived from proteins or peptides of known structure were included, in small numbers (three proteins)[14] at first, leading then to fifteen or more protein examples.[1] Sreerama and Woody[15] compiled a number

of such databases, which included between 17 and 48 spectra from several different labs[3,16,17]; these have endured as useful references for nearly a decade. They include a reasonably good coverage of different protein types (see RDB1-7 in Table I). The validity of some components in these and other reference databases[18] have been questioned[19] as they sometimes include orthologs[18] of the crystal structures rather than the cognate protein, the crystal structures they are based on may not be ideal, and there are some reported differences between spectra of identical proteins in the literature.[20] Nevertheless, the Sreerama and Woody reference databases are still the most popular databases in use, and provide very useful and reasonably accurate references for the existing analysis methods.

NEW DEVELOPMENTS IN REFERENCE DATABASES FOR CD SPECTROSCOPY

Reference databases are created using spectra of proteins whose crystal structures have been determined, and hence whose secondary structures are known. Important criteria for valid reference databases[21] include the availability of high quality X-ray crystal structures (low R and B factors, few missing residues, and good geometries including ϕ, ψ angles in allowed regions of the Ramachandran plot), purified proteins from the same biological source and under similar physical conditions (pH, salt, additives), accurate measurements of protein concentrations, and well-documented calibration conditions. With recent developments in bioinformatics that have systematised the classification of different protein folds (e.g., the CATH database[22]), it has become possible to produce databases, which are far more inclusive of the wide range of structural features now known to exist in proteins.

Recently, a large reference database, designated SP175 (for soluble proteins, data collected to 175 nm) has joined the list of possible reference databases available for CD analyses (Table I).[21] It consists of more than 70 proteins chosen to represent not only a wide range of secondary structures but also an extensive range of protein folds and architectures. The spectra for this database were collected with SRCD for maximum information content and quality; however, it has also been shown to be usable with, and to improve, conventional CD analyses.

In addition to this wide-ranging database, new narrower, focused databases are also being created with the aim of improving analyses of specific classes of proteins that are not well-analyzed by standard databases because the proteins have unusual or specific characteristics. One example of such a database is CRYST175.[20] Each of the nine proteins present in this database (Table I) belongs to the β,γ-crystallin family of eye lens proteins. These proteins have a distinctive double Greek-key fold. This particular narrow reference database provides the best results for the very limited number of proteins with such fold characteristics. Such narrow focused databases may be particularly useful for examining mutants and homologues from other species. Reference databases for other specialised protein types, such as membrane proteins[23] and coiled-coils are being added to the list of possible reference databases that can be used. Additionally, the SP170 database[21] has been produced, which includes very low wavelength VUV data down to 170 nm or below that allows the user to take advantage of the extra information obtainable on SRCD instruments.[12,24]

It is important to note that the more representative a database is of the types of structures to be found in the protein under study, the more accurate the analysis will be. In the future, such resources as the protein circular dichroism data bank (PCDDB),[25] a deposition data bank for validated published CD spectra, should ultimately contribute to producing both broader-based, as well as narrower more specialised, reference databases for CD analyses.

NEW DEVELOPMENTS IN CD ANALYSES: THE DICHROWEB SERVER

Until recently, one of the practical problems in using the various algorithms available for analyses was that they tended to accept different format data, used different reference databases, calculated different types of goodness-of-fit parameters, output the data in different formats so that comparisons between them were difficult, and limited the ability to incorporate new reference databases. The most comprehensive compilation was that included in the software package CDPRO,[15] which required the user to install the software on their own computer, and limited the number of protein components that could be accepted in any given reference database. The DICHROWEB server was developed to provide a user-friendly interface to the existing analysis programs and databases, plus access to new databases and more variable parameters and features; it enables a wide range of input formats and limits the need for preanalysis processing and conversion programs.[26,27] It includes a number of the most popular publicly-available programs, including CONTINLL,[4,28] VARSLC,[29] K2d,[9] CDSSTR,[3] and SELCON3.[10,30] It produces a wide range of output formats, graphical plots and downloadable analyses. Furthermore, it provides access to all reference databases for the various algorithms, enabling simple comparisons to be made between various combinations of algorithms and databases. An important requirement for its

Table I Comparison of Protein Components in Currently Available (Including New) Reference Databases in DICHROWEB

	RDB1	RDB2	RDB3	RDB4	RDB5	RDB6	RDB7	SP175	CRYST175
AAMY								X	
ABNG	X		X	X		X	X	X	
ACY5						X	X		
ACY9						X	X		
ADH			X	X		X	X		
ADK				X			X		
ALDO								X	
APP								X	
APRT								X	
AVDN								X	
AZU	X	X	X	X		X	X		
BAMY								X	
BBTH									X
BB1H									X
BB2H								X	X
BGAL								X	
BLAC	X	X	X	X		X	X	X	
BNJN	X	X	X	X	X	X	X		
BPTI				X			X		
CAL								X	
CAT								X	
CA1			X	X		X	X	X	
CA2								X	
CER								X	
CHYG			X	X		X	X	X	
CHYM	X	X	X	X	X	X	X	X	
CITS								X	
COLA			X	X		X	X		
CONA	X	X	X	X		X	X	X	
CPA				X			X	X	
CPHY								X	
CYTC	X	X	X	X	X	X	X	X	
DHQ1								X	
DHQ2								X	
ECOR	X		X	X		X	X		
ELAS	X	X	X	X	X	X	X	X	
E11									X
FERD								X	
FLVD	X	X	X	X	X	X	X		
GCRB	X		X	X		X	X	X	X
GDCB								X	X
GDCH								X	X
GECB								X	X
GFP			X	X		X	X		
GLOX								X	
GLUD								X	
GPB								X	
GPD	X		X	X	X	X	X		
GRS			X	X		X	X		
GSCH								X	X
HAL								X	
HGBN	X	X	X	X	X	X	X	X	
HMRT	X	X	X	X		X	X		
IFBP			X	X		X	X		
IGG								X	
INSL				X			X	X	
JAC								X	
LACF								X	
LDH	X	X	X	X	X	X	X		
LEP								X	
LLEC								X	

Table I (*Continued from the previous page*)

	RDB1	RDB2	RDB3	RDB4	RDB5	RDB6	RDB7	SP175	CRYST175
LYSM	X	X	X	X	X	X	X	X	
MGB	X	X	X	X	X	X	X	X	
MGBH								X	
MON								X	
NMRA								X	
NUCL				X			X		
OVAL								X	
OVOT								X	
OX20						X	X		
PAPN	X	X	X	X	X	X	X	X	
PARV				X			X		
PELC								X	
PGEN								X	
PGK	X		X	X		X	X	X	
PGLU					X				
PGM								X	
PYK								X	
PLA2								X	
PLEC								X	
PNMT								X	
PPSN	X	X	X	X		X	X		
PRAL	X	X	X	X	X	X	X		
PROX								X	
RHOD			X	X		X	X	X	
RIBA	X	X	X	X		X	X	X	
RUBR					X			X	
SAH								X	
SN06						X	X		
SN70						X	X		
SOD								X	
STI								X	
STRP								X	
SUBA								X	
SUBB	X		X	X	X	X	X		
SUBN	X		X	X		X	X		
SUDS	X	X	X	X		X	X		
THAU								X	
THML	X	X	X	X	X	X	X		
TNF	X		X	X		X	X		
TPI	X	X	X	X	X	X	X	X	
TRPN		X							
T4LS	X	X	X	X		X	X		
UBIQ								X	

AAMY = alpha-Amylase; ABNG = alpha-Bungarotoxin; ACY5 = Apo-cytochrome C (5C) denatured; ACY9 = Apo-cytochrome C (90C) denatured; ADH = Alcohol Dehydrogenase; ADK = Adenylate Kinase; ALDO = Aldolase; APP = Alkaline phosphatase; APRT = Aprotinin; AVDN = Avidin; AZU = Azurin; BAMY = beta-Amylase; BBTH = human beta B1 crystallin (truncated); BB1H = human beta B1 crystallin; BB2H = human beta B2 crystallin; BGAL = beta-galactosidase; BLAC = beta Lactoglobulin; BNJN = Bence Jones Protein; BPTI = Bovine Pancreatic Trypsin Inhibitor; CAL = Calmodulin; CAT = Catalase; CA1 = Carbonic Anhydrase-II (human); CA2 = Carbonic Anhydrase-II (bovine); CER = Ceruoplasmin; CHYG = alpha Chymotrypsinogen; CHYM = alpha Chymotrypsin; CITS = Citrate synthase; COLA = Colicin A; CONA = Concanavalin A; CPA = Carboxypeptidase-A; CPHY = c-Phycocyanin; CYTC = Cytochrome C; DHQ1 = Dehydroquinase-type 1; DHQ2 = Dehydroquinase-type 2; ECOR = EcoR1 Endonuclease; ELAS = Elastase; E11 = gamma crystallin E11 mutant (bovine); FERD = Ferredoxin; FLVD = Flavodoxin; GCRB = gamma B crystallin (bovine); GDCB = gamma D crystallin (bovine); GDCH = gamma D crystallin (human); GECB = gamma E crystallin (bovine); GFP = Green Fluorescent Protein; GLOX = Glucose Oxidase; GLUD = Glucose Dehydroxidase; GPB = Glycogen phosphorylase-b; GPD = Glyceraldehyde 3-P dehydrogenase; GRS = Glutathione Reductase; GSCH = gamma S crystallin (C-terminal domain) (human); HAL = Haloalkane dehydrogenase; HGBN = Hemoglobin; HMRT = Hemerythrin; IFBP = Rat Intestinal Fatty Acid Binding Protein; IGG = IgG; INSL = Insulin; JAC = Jacalin; LACF = Lactoferrin; LDH = Lactate Dehydrogenase; LEP = Leptin; LLEC = Lentil Lectin; LYSM = Lysozyme; MGB = Myoglobin (sperm whale); MGBH = Myoglobin (horse); MON = Monellin; NMRA = NmrA; NUCL = Nuclease; OVAL = Ovalbumin; OVOT = Ovotransferrin; OX20 = Ribonuclease (20C) denatured; PAPN = Papain; PARV = Parvalbumin; PELC = Pectate Lyase C; PGEN = Pepsinogen; PGK = Phosphoglycerate Kinase; PGLU = Poly Glutamic Acid; PGM = Phosphoglucomutase; PYK = Pyruvate kinase; PLA2 = Phospholipase-A2; PLEC = Pea Lectin; PNMT = Phenylethanolamine N-methyltransferase; PPSN = Pepsinogen; PRAL = Prealbumin; PROX = Peroxidase; RHOD = Rhodanasc; RIBA = Ribonuclease A; RUBR = Rubredoxin; SAH = Serum Albumin (human); SN06 = Staphylococcal Nuclease (6C) denatured; SN70 = Staphylococcal Nuclease (70C) denatured; SOD = Superoxide dismutase (Cu, Zn); STI = Soyabean Trypsin Inhibitor; STRP = Streptavidin; SUBA = Subtilisin A; SUBB = Subtilisin BPN; SUBN = Subtilisin novo; SUDS = Superoxide Dismutase; THAU = Thaumatin; THML = Thermolysin; TNF = Tumor Necrosis Factor; TPI = Triose Phosphate Isomerase; TRPN = Trypsin; T4LS = T4 Lysozyme; UBIQ = Ubiquitin.

users is that they also cite the literature for the original algorithms and databases accessed through DICHROWEB, in order that those authors receive the appropriate credit for their work. A standard goodness-of-fit parameter, the normalised root mean square (NRMSD),[31,32] provides an indication as to how closely the back-calculated spectra produced from the predicted secondary structures reproduce the experimental spectrum (also indicted graphically by plots of both spectra and the difference spectrum derived from the calculated and experimental spectra).

The DICHROWEB server was first made publicly-available in 2002,[26] and now has more than 1400 registered users from 43 countries. To date more than 110,000 deconvolutions have been performed on the server, which has recently been upgraded to cope with the increased demand. Its principal features were described in Whitmore and Wallace,[27] but since that time, many new functions have been included. This article describes subsequent developments, features and methods that are included in the present version and how they aid in improving analyses.

NEW FEATURES IN DICHROWEB

Low Wavelength Cut-Off Option

An important new option is the ability to select the lowest wavelength to use in the analysis, following the input of the full spectral data collected. The purpose of this option is to allow the user to remove unreliable data collected at the low wavelength end (i.e., data that are either too noisy, or for which the high tension (HT) or dynode reading indicates the intensity of the light reaching the detector was too low). The importance of considering such a cut-off value is described in Miles and Wallace[12] and Kelly et al.[33]

New Reference Databases

The principal new addition to DICHROWEB is the inclusion of the wide-ranging SP175 reference database of Lees et al.,[21] described above. The second new reference database available is CRYST175[20] specifically for proteins with a distinctive double Greek-key fold as found in the crystallin eye proteins.

The nine reference databases currently available in DICHROWEB each have different characteristics. Whilst there is some overlap between the protein constituents of some of the databases (Table I), because they contain different structural components and produce different spectral features, they give rise to subtly different deconvolution results. It is advised that the user attempt analyses with several of these reference databases, and use the NRMSD parameter as one guide as to which of the databases may be more suitable for the analysis of their protein (i.e., they contain representative proteins with more comparable features to those present in the query protein).

More Wavelength Range Options

The number of different types of secondary structure that can be derived from the analysis depends on the available wavelength range of the collected spectra (spectra to lower wavelengths (i.e., SRCD data) have a higher information content, and hence can be deconvoluted into more component types).[34,35] Databases that include different wavelength ranges have been a feature of DICHROWEB since its inception, restricting analyses to those compatible with the wavelength range available in that database. This has meant that most of the databases cannot be used if data are collected to only 190 nm. The addition of the SP175 database, which contains data down to 175 nm, but which has been shown to produce excellent results using only data down to 190 nm,[21] now enables more options for analyses with restricted wavelength data. However, it should be noted that because the accuracy and ability to analyse for four or more components drops off rapidly with the absence of data below 190 nm, all analyses with DICHROWEB require data to at least 190 nm (with the exception of the K2d neural network, which only requires data to 200 nm, but effectively only accurately analyses for helical components).

Magnitude Scaling

A further improvement is the inclusion of the advanced feature of magnitude scaling. This option permits multiplication of the magnitudes of the CD measurements by a scale factor that can be varied by the user from 0.51 to 1.49. The default value is 1.0. The purpose of scaling is to allow users to easily make corrections to the magnitude of the spectrum in cases where they establish that the experimental pathlength or protein concentration was incorrectly known at the time of the experiment and where it is impractical to repeat the experiment. The magnitude of the spectra has a linear relationship to the experimental pathlength and protein concentrations, so a scaling factor can be calculated for any corrections that need to be applied to either of these values. The effects of using scaling values on deconvolutions has been explored by Miles et al.,[36] who showed that scale factors that minimise the NRMSD values calculated often produce the most accurate secondary structure values.

Aids to Identification of Possible Errors/Frequently Asked Questions

In the course of operating the DICHROWEB server, a number of examples of failed deconvolutions have been observed and over time a picture has built up of common errors that occur. The errors generally fall into four categories: data entry, corrupt data files, experimental data collection problems and inappropriate reference databases. The most common types of errors are addressed as potential pitfalls/cautions in the FAQ section of the website, but are summarised briefly here.

Common data entry errors include entering the start and end wavelengths of the data the wrong way around and submitting binary- rather than ASCII- based files to the server. These are obvious when the data are displayed in the "preview" window. Other logical data entry errors are trapped at the data input stage with a checking function, which ensures that data are not sent to the server when they are out of the expected numerical bounds, or where numbers are entered into character input boxes and vise versa.

Corrupted data files can occur when the data have been moved between different computational environments or between different text reading/formatting programs, which may introduce special control characters into the data file that are invisible whilst the data are being viewed with the software that created it. Software has been introduced into DICHROWEB, which eliminates many of the common extra control characters, but unusual occurrences may still elude these data checkers.

A number of types of experimental/data collection conditions will cause data analysis problems. Common errors seen are when the data are too noisy (especially at low wavelengths) because either the sample signal is too low or the HT/dynode voltage is too high[12] or when the data are not properly zeroed (i.e., bad match between sample and baseline in wavelength regions where there should be no signal). Significant errors occur when either the cell pathlength or the protein concentration are not accurately determined, resulting in errors in magnitude of the input spectra, which will result in completely erroneous calculated values.[36] Finally, if the units are incorrect (i.e., values are calculated as mean residue ellipticity when delta epsilon units are given), the resulting calculations will be nonsensical; however, DICHROWEB provides the opportunity to convert between units where necessary. Presently validation software is being developed that can be run either in conjunction with a DICHROWEB analysis or offline (Woollett, Janes, Wallace, in preparation) to identify these and other data issues.

The final common cause of errors is the use of reference databases on noncognate or inappropriate samples. For example, the user's guide notes that all the reference databases have been derived from globular soluble proteins, so that use for any other types of samples is inadvisable. Specifically noted are peptides, which tend to have low spectral magnitudes and adopt multiple conformations in equilibrium rather than a single structure, proteins and peptides in nonaqueous solutions (i.e., membrane proteins, which exhibit different spectral characteristics because of differences in the "solvent" dielectric constant),[37,38] fibrous proteins, which tend to have different scattering properties and adopt different types of structures such as polyproline II helices, proteins with unusual supersecondary structures such as coiled-coils that are not in any of the current databases, and proteins with high amounts of "disordered" structures, that are not well represented in databases derived from crystallised proteins.

FUTURE DEVELOPMENTS FOR DICHROWEB

Since its inception, DICHROWEB has been constantly updated and new features added[27] in response to both developer-initiated ideas and requests of users.[27] We anticipate that over the next development cycle the following functionalities will be added to improve and enhance analyses:

Customised Reference Databases

An important future improvement will be the ability to create customised reference databases to target specific classes of proteins, such as membrane proteins and other special samples. This will be possible with the advent of the protein circular dichroism data bank, a deposition databank of CD spectra of proteins[25] currently under construction. With the availability of a large data bank of protein spectra produced by spectroscopists worldwide, users will be able to select spectra that can be combined to make specialised databases (much in the manner of the CRYST175 database described above). To make these data compatible with DICHROWEB, we will provide an integrated interface for DICHROWEB and the PCDDB.

Back Calculations of Spectra from Input Protein Data Bank Coordinates

Additional software to be included in DICHROWEB will calculate a CD spectrum for a known protein structure based on its crystallographic protein data bank coordinates. "Back calculation" can be useful in deciphering contributions of known components, for instance in fusion proteins, and as a means of identifying related, denatured or incorrectly folded proteins.

Matrix Calculations

The facility to undertake batch calculations using all algorithms and databases, producing a complete range of calculated secondary structures, averaged values and standard deviations for all of the calculations will provide a means of facilely showing the consistency and variation produced by the different methods. This can then be used as another means of judging the reliability of the analyses.

FUTURE POTENTIAL FOR IMPROVEMENTS IN CD ANALYSES

New techniques and technologies for data collection, especially with the advent of SRCD spectroscopy,[12] have improved the accuracy of CD analyses. Principal component analyses have shown that the information content of the spectrum increases as a function of inclusion of more low wavelength data[21,34,35] due to the additional and more complete transitions measured.[39] New reference databases that include the very low wavelength data developed using SRCD data[39] should ultimately lead to improved and expanded analyses, including information at the level of protein folds.[40] Quantitative analyses of "rarer" types of secondary structures such a 3_{10} helices and various types of turns will be enhanced as the number and breadth of proteins present in the reference databases increase. In addition, improvements in cross-calibrations between different instruments[41–43] will result in much more consistent data that can be best used with standardised databases.

CONCLUSIONS

It is important to note that whilst tools such as DICHROWEB that aid the user in analysing their CD data can result in more rapid and improved analyses, there is also the possibility that if used in a "blackbox" manner, users can produce less than ideal (or even erroneous) conclusions. Thus a number of precautions need to be considered in the use and interpretation of the results. Notably the following: (1) The amount of data must be sufficient to solve for the desired number of secondary structure components. Data that only extends to 200 nm contains at most two eigenvectors, and hence the results should only be interpreted in terms of two components (i.e., how much is helix and how much is not helix). Any interpretation of such data that attempts to deconvolute into more components than these will be an over-interpretation of the data. (2) A low value for the NRMSD or any other goodness-of-fit parameter does not always indicate it represents a correct solution. A low NRMSD value ($\leq$0.1) is a necessary but not sufficient condition for accuracy in secondary structure determination. However a high value is a good indication that either the analysis has gone wrong (often because the magnitude of the spectrum is incorrect) or the reference database is inappropriate for the characteristics of the protein being analysed. It is also important to note that some algorithms, notably CDSSTR, nearly always produce the lowest NRMSD due to the way they fit the data, but they very often are not the most correct solution.[36] (3) Reference databases derived from globular soluble proteins are not appropriate for the analysis of proteins (or peptides) in nonaqueous solutions.[38] (4) It is absolutely essential to have precisely correct concentration measurements (not just estimates from colorimetric assays) and an accurate measurement of the cell pathlength (the values cited by the manufacturers, especially for very short pathlengths, can err by 30% or more).[42] The consequence of concentration and pathlength errors is that the magnitude of the spectrum produced will err by a corresponding amount and result in incorrect analyses. Other good practice issues that can affect analyses are described in detail in Kelly et al.[33] and Miles and Wallace.[12]

In summary, with the easy availability of a wide range of empirical algorithms for secondary structure calculations, new reference databases and other data analysis tools, CD and SRCD spectroscopy should prove to be even more valuable tools in structural biology over the next decades than CD has been in the past 40 or more years since it was first used to examine protein structures.

The authors thank the members of Wallace group at Birkbeck College, University of London, and the Janes group at Queen Mary, University of London, for helpful discussions.

REFERENCES

1. Chang, C. T.; Wu, C. S.; Yang, J. T. Anal Biochem 1978, 91, 13–31.
2. Wallace, B. A.; Teeters, C. L. Biochemistry 1987, 26, 65–70.
3. Compton, L. A.; Johnson, W. C., Jr. Anal Biochem 1986, 155, 155–167.
4. Provencher, S. W.; Glockner, J. Biochemistry 1981, 20, 33–37.
5. Sreerama, N.; Woody, R. W. Anal Biochem 1993, 209, 32–44.
6. Perczel, A.; Park, K.; Fasman, G. D. Anal Biochem 1992, 203, 83–93.
7. Pancoska, P.; Janota, V.; Keiderling, T. A. Anal Biochem 1999, 267, 72–83.
8. Böhm, G.; Muhr, R.; Jaenicke, R. Protein Eng 1992, 5, 191–195.
9. Andrade, M. A.; Chacón, P.; Merelo, J. J.; Morán, F. Protein Eng 1993, 6, 383–390.
10. Lees, J. G.; Miles, A. J.; Janes, R. W.; Wallace, B. A. BMC Bioinformatics 2006, 7, 507–517.
11. Wallace, B. A.; Wien, F.; Miles, A. J.; Lees, J. G.; Hoffman, S. V.; Evans, P.; Wistow, G. J.; Slingsby, C. Faraday Discuss 2004, 17, 653–661.

12. Miles, A. J.; Wallace, B. A. Chem Soc Rev 2006, 35, 39–51.
13. Greenfield, N.; Fasman, G. D. Biochemistry 1969, 8, 4106–4116.
14. Saxena, V. P.; Wetlaufer, D. P. Proc Natl Acad Sci USA 1971, 68, 969–973.
15. Sreerama, N.; Woody, R. W. Anal Biochem 2000, 287, 252–260.
16. Sreerama, N.; Venyaminov, S. Y.; Woody, R. W. Anal Biochem 2000, 287, 243–251.
17. Pancoska, P.; Keiderling, T. A. Biochemistry 1991, 30, 6885–6895.
18. Raussens, V.; Ruysschaert, J.-M.; Goormaghtigh, E. Anal Biochem 2003, 319, 114–121.
19. Janes, R. W. Bioinformatics 2005, 21, 4230–4238.
20. Evans, P.; Bateman, O.; Slingsby, C.; Wallace, B. A. Exp Eye Res 2007, 84, 1001–1008.
21. Lees, J. G.; Miles, A. J.; Wien, F.; Wallace, B. A. Bioinformatics 2006, 22, 1955–1962.
22. Orengo, C. A.; Pearl, F. M.; Thornton, J. M. Methods Biochem Anal 2003, 44, 249–271.
23. Wallace, B. A.; Wien, F.; Stone, T. C.; Miles, A. J.; Lees, J. G.; Janes, R. W. Biophys J 2006, 90, 317a.
24. Sutherland, J. C.; Emerick, A.; France, L. L.; Monteleone, D. C.; Trunk, J. Biotechniques 1992, 13, 588–590.
25. Wallace, B. A.; Whitmore, L.; Janes, R. W. Proteins: Struct Funct Bioinformatics 2006, 62, 1–3.
26. Lobley, A.; Whitmore, L.; Wallace, B. A. Bioinformatics 2002, 18, 211–212.
27. Whitmore, L.; Wallace, B. A. Nucleic Acids Res 2004, 32, W668–W673.
28. Van Stokkum, I. H. M.; Spoelder, H. J. W.; Bloemendal, M.; Van Grondelle, R.; Groen, F. C. A. Anal Biochem 1990, 191, 110–118.
29. Manavalan, P.; Johnson, W. C., Jr. Anal Biochem 1987, 167, 76–85.
30. Sreerama, N.; Venyaminov, S. Y.; Woody, R. W. Protein Sci 1999, 8, 370–380.
31. Brahms, S.; Brahms, J. J Mol Biol 1980, 138, 149–178.
32. Mao, D.; Wachter, E.; Wallace, B. A. Biochemistry 1982, 21, 4960–4968.
33. Kelly, S. M.; Jess, T. J.; Price, N. C. Biochim Biophys Acta 2005, 1751, 119–139.
34. Toumadje, A.; Alcorn, S. W.; Johnson, W. C., Jr. Anal Biochem 1992, 200, 321–331.
35. Wallace, B. A.; Janes, R. W. Curr Opin Chem Biol 2001, 5, 567–571.
36. Miles, A. J.; Whitmore, L.; Wallace, B. A. Protein Sci 2005, 14, 368–374.
37. Chen, Y. C.; Wallace, B. A. Biophys Chem 1997, 65, 65–74.
38. Wallace, B. A.; Lees, J.; Orry, A. J. W.; Lobley, A.; Janes, R. W. Protein Sci 2003, 12, 875–884.
39. Wallace, B. A. Nat Struct Biol 2000, 7, 708–709.
40. Miles, A. J.; Wallace, B. A. Biophys J 2007, 92, 337a.
41. Miles, A. J.; Wien, F.; Lees, J. G.; Rodger, A.; Janes, R. W.; Wallace, B. A. Spectroscopy 2003, 17, 653–661.
42. Miles, A. J.; Wien, F.; Lees, J. G.; Wallace, B. A. Spectroscopy 2005, 19, 43–51.
43. Lees, J. G.; Smith, B. R.; Wien, F.; Miles, A. J.; Wallace, B. A. Anal Biochem 2004, 332, 285–289.

Reviewing Editor: Lila Gierasch