Published in final edited form as:
Anal Biochem. 2009 June 15; 389(2): 174–176. doi:10.1016/j.ab.2009.03.036.

CD spectroscopy has intrinsic limitations for protein secondary structure analysis

Sergei Khrapunov
Department of Biochemistry, Albert Einstein College of Medicine, 1300 Morris Park Avenue, Bronx, New York 10461, USA

Abstract

Secondary structure content (SSC) cannot be accurately calculated from circular dichroism (CD) spectra for the majority of proteins whose three dimensional structures have been solved. 'Reliable' SSC that is significantly different from random SSC can be calculated from CD spectra only for all-α proteins and all-α proteins with canonical β-strand geometry.

The two fields to which the protein CD spectroscopy is applied with well-developed methodology are folding thermodynamics [1;2] and secondary structure estimation [3]. Most thermodynamic studies rely on relative changes in CD spectra and are therefore relatively independent of calibration with structure. In contrast, the calculation of protein secondary structure content (SSC) requires strict cross-calibration/validation of experimental and reference CD spectra with reference crystallographic or NMR structural data. Following the development and dissemination of reliable CD analysis software via the internet [4;5;6;7], improvements in SSC calculations have resulted from increasing the number of proteins in the protein reference set [4], splitting the ordered fractions of regular and distorted portions [6] and expanding CD spectral analysis to wavelengths below 185 nm using vacuum ultraviolet circular dichroism spectroscopy [8;9]. Splitting the ordered fractions occurs when α-helices and β-strands are divided into regular and distorted classes [6] yielding six secondary structure classifications: regular α-helix (αR), distorted α-helix (αD), regular β-strand (βR), distorted β-strand (βD), turn and disordered.

The performances of secondary structure calculations are typically characterized by the root-mean-square deviations (*RMSD*) between the crystal and CD estimates of the secondary-structure content,

$$RMSD = \sqrt{\frac{\sum (Y_i - X_i)^2}{N}}, \quad (1)$$

where X_i and Y_i are the crystallographic and CD estimates of a given type of secondary structure, i, in N reference samples. The overall *RMSD* is determined by considering all secondary fractions collectively [4;6;8]. Lower values of *RMSD* indicate less discrepancy between the

Correspondence should be addressed to S.K. (E-mail: khraps@medusa.bioc.aecom.yu.edu).

calculated and crystallographic data. It is generally accepted that *RMSD* measures the predictive power of the method.

Joint application of splitting the ordered fractions and utilization of the lower wavelength CD data (down to 160 nm if obtainable) yields in the best accuracy [8]. Overall *RMSD* values obtained for 29 of 31 studied proteins [8] are less than the overall *RMSD* values of 0.091 – 0.098 calculated on the basis of the splitting only [6].

Two questions arise from this and similar results. First, what is the lower limit for *RMSD* in such calculations? Second, is the accuracy that is reached sufficient to make a reliable and meaningful estimation of SSC for the proteins from their CD spectra? To answer these questions we have compared the overall *RMSD* for the 31 proteins calculated from their CD spectra [8] with the overall *RMSD* value obtained for simulated SSC assuming that the main secondary structure types (α-helices, β-strands, turns and disordered) are represented equally. The simulated SSC values were assumed equal to 0.125 each for regular and distorted helices and strands (αR, αD, βR, βD) and 0.25 each for turns and unordered structure.

Table 1 lists the *RMSD* values comparing crystallographic and CD SSC estimates (*RMSD*cd) and the *RMSD* values comparing the crystallographic estimates with those obtained from simulated SSC values of SSC (0.125 or 0.25 in the particular cases; *RMSD*s). The proteins in Table 1 are grouped accordingly to their tertiary structure class: all-α, all-α and αβ combining α+β and α/β classes [10;11]. Peroxidase and xylanase are placed in the all-α and all-β groups since their helix/strand ratios are 13/2 and 1/15, respectively.

It is readily evident in the table that except for human serum albumin, only for the proteins belonging to all-α and all-β classes are the simulated *RMSD* values essentially higher than the experimental ones and values higher than the overall *RMSD* value of 0.091 estimated for 29 proteins for DSSP assignments [6]. Of the 22 proteins of the αβ class, 10 show simulated *RMSD*s lower or comparable with experimental ones while the remainder has simulated *RMSD* lower than the overall value of the DSSP assignment [6]. The only exception is insulin, presumably due to its small dimension, short structural elements and uncertainty in its intermolecular interactions in solution [12]. These properties are expected to influence the CD spectrum of insulin. Moreover individual β-sheets have variable CD spectra due to the variations in the geometry of β-structure in proteins [13]. If β-strands are within an unusual structural motif like the Pentapeptide Repeat Protein fold, the SSC calculated from CD and crystallographic data demonstrate poor correspondence [14]. Thus the analyzed proteins included in all-β class all apparently have canonical β-structure.

As follows from this analysis SSC cannot be accurately calculated from CD spectra for the vast majority of the proteins. Reliable calculation of SSC from CD spectra can be made only for all-α proteins and all-α proteins whose β-strands geometry is apparently canonical. Why does this occur and can the method be ‘repaired’? We propose there are some intrinsic limits to the application of CD spectroscopy to protein secondary structure calculation as summarized below: a) The quality of the Ramachandran plots of the reference crystallographic structures is poor for some reference CD datasets. In fact, the residues for most of the structures in some protein reference sets are under the 90% and even the 80% thresholds for the most favored region of the Ramachandran plot [15]; b) The quality of the proteins used for solution and crystallographic studies may not be consistent. Many of the reference CD spectra are obtained using commercially prepared proteins without purification [4;8]; c) The consistency of the reference CD database sets is sometimes suspect. The spectrum of some proteins differs in different databases [15]; d) There has been little cross validation of the instruments used to obtain reference and experimental CD spectra. Many reference CD spectra were obtained long ago sometimes on laboratory-specific instruments whose specifications are not documented.

Perhaps creation of a central resource of the published and cross-validated CD data files such as the proposed Protein Circular Dichroism Data Bank [16] can help solve these problems.

Lastly, the different algorithms used to calculate protein SSC from crystallographic structures give average contents for particular structures with standard deviations comparable with the RMSD values shown in the Table 1 [6;9]. Unlike the above problems this one cannot be easily fixed. Its solution requires the mutual agreement of the scientific community on the ceasing the indiscriminate use of the programs DSSP, Procheck, STRIDE, XtlSSTR and PROMOTIF in favor of one. We favor the DSSP algorithm as it is mostly used for PDB files.

We conclude that a reliable and meaningful estimation of SSC from the CD spectra can not be made for proteins with the mixed α and β elements in their structure and apparently for proteins with the noncanonical β-strand geometry. At the same time such estimations can be used as relative measures of the structural changes of the proteins at different conditions such as those observed during folding and unfolding.

Acknowledgments

The author thanks Dr. Michael Brenowitz for his constant support and generous sharing of many valuable suggestions. This study was supported by grant R01-GM079618 from the Institute of General Medical Sciences of the National Institutes of Health.

References

1. Greenfield N. Using circular dichroism collected as a function of temperature to determine the thermodynamics of protein unfolding and binding interactions. Nature Protocols 2006;1:2876–2880.
2. Greenfield N. Determination of the folding of proteins as a function of denaturants, osmolytes or ligands using circular dichroism. Nature Protocols 2006;1:2733–2741.
3. Greenfield N. Using circular dichroism spectra to estimate protein secondary structure. Nature Protocols 2006;1:2876–2890.
4. Sreerama N, Woody R. Estimation of protein secondary structure from circular dichroism spectra: comparison of CONTIN, SELCON, and CDSSTR methods with an expanded reference set. Analitical Biochemistry 2000;287:252–260.
5. Whitmore L, Wallace B. DICHROWEB, an online server for protein secondary structure analyses from circular dichroism spectroscopic data. Nucleic Acids Research 2004;32:W668–W673. [PubMed: 15215473]
6. Sreerama N, Venyaminov S, Woody R. Estimation of the number of alpha-helical and beta-strand using circular dichroism spectroscopy. Protein Science 1999;8:370–380. [PubMed: 10048330]
7. Sreerama N, Venyaminov S, Woody R. Analysis of Protein Circular Dichroism Spectra Based on the Tertiary Structure Classification. Analitical Biochemistry 2001;299:271–274.
8. Matsuo K, Yonehara R, Gekko K. Imroved estimation of the secondary structures of Proteins by vacuum-ultraviolet circular Dichroism spectroscopy. Journal of Biochemistry 2005;138:79–88. [PubMed: 16046451]
9. Wallace B, Lees J, Orry A, Lobley A, Janes R. Analyses of circular dichroism spectra of membrane proteins. Protein Science 2003;12:875–884. [PubMed: 12649445]
10. Levitt M, Chotia C. Structural patterns in globular proteins. Nature 1976;261:552–558. [PubMed: 934293]
11. Manalavan P, Johnson WJ. Sensitivity of circular dichroism to protein tertiary structure class. Nature 1983;305:831–832.
12. Grudzielanek S, Jansen R, Winter R. Solvational Tuning of the Unfolding, Aggregation and Amyloidogenesis of Insulin. Journal of Molecular Biology 2005;351:879–894. [PubMed: 16051271]
13. Sreerama N, Woody R. Computation and analysis of protein circular dichroism spectra. Methods in Enzymology 2004;383:318–351. [PubMed: 15063656]

14. Khrapunov S, Cheng H, Hegde S, Blanchard J, Brenowitz M. Solution Structure and Refolding of the Mycobacterium tuberculosis Pentapeptide Repeat Protein MfpA. Journal of Biological Chemistry 2008;283:36290–36299. [PubMed: 18977756]

15. Janes R. Bioinformatics analysis of circular dichroism protein reference databases. Bioinformatics 2005;21:4230–4238. [PubMed: 16188926]

16. Wallace B, Lee W, Janes R. The Protein Circular Dichroism Data Bank (PCDDB): A Bioinformatics and Spectroscopic Resource. Proteins 2006;62:1–3. [PubMed: 16245340]

Table 1

The accuracy of secondary structure content (SSC) calculations of proteins from their CD spectra

(**α**- all-α; **β** - all-β; **αβ** – combined α+β and α/β classes [10;11]; **RMSDcd** –overall RMSD from [8]; **RMSDs –** simulated overall RMSD (see text); **normal** – RMSDs values less or comparable with RMSD; ***italic*** - RMSDs values less than overall RMSD (0.091) estimated for DSSP assignment [6]; **bold** - RMSDs values essentially higher than RMSD)

Protein	Source	PDB code	Class	RMSDcd	RMSDs
Myoglobin	Horse heart	1WLA	α	0.03	**0.205**
Hemoglobin	Bovine blood	1G08	α	0.048	**0.201**
Cytochrome C	Horse heart	1HRC	α	0.073	**0.100**
Albumin	Human serum	1AO6	α	0.036	0.036
Peroxidase	Horse radish	1ATJ	α	0.038	**0.102**
Alfa-Lactalbumin	Bovine milk	1F6S	αβ	0.053	*0.078*
Lysozyme	Hen egg	1HEL	αβ	0.048	*0.078*
Ovalbumin	Hen egg	1OVA	αβ	0.065	0.064
RNase A	Bovine pancreas	1FS3	αβ	0.038	0.042
β-Lactoglobulin	Bovine milk	1B8E	αβ	0.046	*0.076*
Pepsin	Porcine stomach	4PEP	αβ	0.053	*0.073*
Trypsinogen	Bovine pancreas	1TGN	αβ	0.047	*0.063*
α-hymotrypsinogen	Bovine pancreas	2CGA	αβ	0.016	*0.060*
Insulin	Pig pancreas	4INS	αβ	0.103	**0.137**
Glucose isomerase	Strept. rubiginosus	1OAD	αβ	0.048	*0.083*
Lactate dehydrogenase	Bovine heart	8LDT	αβ	0.031	*0.083*
Lipase	Pseudomonas cepacia	3LIP	αβ	0.044	0.048
Transferrin	Human serum	1LFG	αβ	0.032	0.038
Conalbumin	Chicken egg white	1OVT	αβ	0.038	0.035
Thioredoxin	Escherichia coli	2TRX	αβ	0.065	0.054
Catalase	Bovine liver	7CAT	αβ	0.038	*0.055*
α-amylase	Bacillus subtilis	1BAG	αβ	0.032	0.038
Subtilisin A	Bacillus subtilis	1SBC	αβ	0.047	0.035
Papain	Papaya	8PAP	αβ	0.065	0.071
Elastase	Porcine pancreas	3EST	αβ	0.031	*0.076*
Carbonic anhydrase	Bovine erythrocytes	1G6V	αβ	0.05	0.058

Anal Biochem. Author manuscript; available in PMC 2010 June 15.

Protein	Source	PDB code	Class	RMSDcd	RMSDs
Trypsin inhibitor	Soybean	1AVU	β	0.058	**0.114**
Concanavalin A	Jack bean	2CTV	β	0.051	**0.104**
Xylanase	Trichoderma	1ENX	β	0.108	**0.158**
Avidin	Chicken egg white	1AVE	β	0.082	**0.110**