Biophysical Chemistry, 48 (1993) 19–29
Elsevier Science Publishers B.V., Amsterdam

BIOCHE 01784

Effect of spectral window size on circular dichroism spectra deconvolution of proteins

András Perczel [a,b] and Gerald D. Fasman [a,*]

[a] *Graduate Department of Biochemistry and Chemistry, Brandeis University, Waltham, MA 02254-9110 (USA)*
[b] *Department of Organic Chemistry, Eötvös University, 112 Budapest P.O. Box 32 (Hungary)*

(Received 29 March 1993; accepted in revised form 3 June 1993)

Abstract

The recently developed Convex Constraint Algorithm (CCA), which has been found to be efficient for CD spectra deconvolution of proteins, was used to examine the importance of the spectral window size and its effects on the deconvolution sensitivity. Using the "largest" protein CD spectra data base ever published (A. Toumadje, S.W. Alcorn and W.C. Johnson, Jr., *Anal. Biochem.* 200 (1992) 321–331) a systematic "spectra-truncation" was performed. The reduced spectral data sets were deconvoluted and the Pearson product-moment correlation (r), as well as the RMS values, were calculated. It was found that the spectral window size enlargement below 180 nm has only minor secondary effect on the accuracy of the deconvolution, since the r and RMS values were nearly insensitive to the broadening of the spectral window. By contrast, these values monitored the decrease of the deconvolution capacity for the analysis when the 180–200 nm spectral region was likewise eliminated. This observation harmonizes perfectly with the recent observation of S.Y. Venyaminov, I.A. Baikalov, C.-S.C. Wu and J.T. Yang (*Anal. Biochem.* 198 (1991) 250–255) and with the pioneering observation of N. Greenfield and G.D. Fasman (*Biochemistry* 8 (1969) 4108–4116), namely, by taking only the CD spectral region above 200 nm, the α-helix content may be estimated with an acceptable accuracy. It is therefore concluded that the development and the analysis of a large protein data base (e.g. dozens of spectra), incorporating as many "dissimilar proteins" as possible, may result in a better understanding of the relationship between the CD spectra and the secondary structural elements of proteins.

Keywords: Circular dichroism; Spectral window size; Proteins; Deconvolution

1. Introduction

Since 1965, when the first circular dichroism (CD) measurements investigating the helical content of globular proteins were performed by Holzwarth and Doty [1], the secondary structure determination of peptides and proteins by CD spectroscopy has been frequently attempted. Although the measured CD spectra is a function of the secondary structural elements, such as the α-helix, β-sheet, loops, β-turns etc., the CD spectra also vary with other factors (solvent (s), temperature (T), concentration (c), salt or ion concentration (I), etc.);

$$[\theta]_\lambda = f(\lambda, T, c, s, I, \ldots) \quad (1)$$

* Corresponding author. Tel.: 617-736-2370, Fax: 617-736-2376, E-mail: Bitnet FASMAN@BRANDEIS.

The mean residue ellipicity ($[\theta]_\lambda$) at a given wavelength (λ) is the explicit function of several factors. The observation, that the different secondary structural elements result in different types of CD patterns in the UV range, gives credence to the conformation determination based on CD spectra (for a review article see Ref. [2]). The conformation sensitivity of the n–π^* and π–π^* bands has been adequately demonstrated [3–6]. Theoretical calculations revealed that small torsional angle changes may result in significant alterations in the overall spectra [3–6]. By contrast, it is still not clear whether all the different secondary structures result in basically different CD spectra. In other words, two or more different sets of backbone torsion-angle combinations may accidentally result in highly similar n–π^* and π–π^* bands.

One of the main issues of CD spectral interpretation and assignment is directly related to the problem of an objective and relevant secondary structure assignment of the X-ray determined conformations. Almost all the deconvolution programs [7–9] are directly coupled with the "X-ray interpretation" problem. The fact that several different approaches [10–13] exist for the location of the secondary structural elements of the backbone of a protein reveals the "subjective" aspects of the field. The different interpretations of protein X-ray diffraction data may result in significantly different secondary structural percentages [14–16], leading to controversial CD interpretations. Using NMR spectroscopy, Wagner and others [17] recently showed that the solid and solution state conformation may be significantly different for various proteins. To avoid this and other problems, a new CD spectra deconvolution method was developed [15,16,18], which is totally independent of the crystallographic results. The Convex Constraint Algorithm (CCA) [15,16] was previously tested [18] and applied successfully for CD data deconvolution. The algorithm operates by using only measured CD curve data points and yields curves and weights simultaneously for "reconstructing" the original data set within a given tolerance (σ). The advantages of this method are that, based on the applied constraints, the shapes of the resulting pure component curves have a typical protein secondary structural appearance, namely, a number of CD bands with intensities similar to those found in native proteins. As previously discussed, the number of the expected pure components must be determined *a priori* (input data). The resulting weights and curves must be assigned followed by the deconvolution. Therefore, the deconvolution is "totally free" from any type of external data, such as X-ray or NMR parameters, but the assignment requires external information, such as the shape of theoretical calculated "pure" curves. However, this problem originates from the CD method itself (a relative method) and not from the deconvolutional procedure.

The second important problem arises from the "information content" of the standard spectral range (185–260 nm) used during the data analyses. The efforts of Johnson et al. [19] and others [20,21], to enlarge the analyzed spectral range in the shorter wavelength, seem to provide a plausible answer to the question, "How to improve secondary structure determination based on CD measurements?" The spectral domain above 260 (near UV region) nm is dominated by the "aromatic" contribution, which is also thought to be "less informative" toward conformational changes of the amide bonds. If bands below 180 are detectable and have a high conformational sensitivity, they would be expected to be useful for conformational deconvolution. Because no commercial instrument is available which can be used for such UV–CD measurements, the applicability of such an analysis is difficult. The thorough and assiduous measurements of Toumadje et al. [19] yielded a CD spectral set of 16 proteins, with known secondary structure as previously determined by X-ray diffraction studies.

During the volume minimization of the appropriate simplex (constraint c), the purest component-containing structures have key roles. In a real set of CD curves (if external data is not applied, such as X-ray or NMR), it is obviously unknown to what degree a pure component is represented in a conformational mixture. However, based on X-ray and NMR analysis, proteins generally adopt more than single secondary structural elements, meaning that only a few confor-

mational prototypes can be represented with a high weight (e.g., the α-helical content of myoglobin is $\approx 80\%$). Unfortunately, from a conformational point of view, the measured protein set does not cover a wide combination of conformational types. Only the α-helix secondary structure is represented in a large range (from 10% to 100%), but the β-turn (from 0% to 22%) and the β-sheet (from 0% to 45%) content of these proteins is highly restricted. Such a "weakness" of the data base originates from the true nature of proteins, and it makes secondary structure assignment rather problematic. Although the addition of more CD spectra of some selected proteins would be expected to improve the analyses (compare the improvement of the analyses between refs. [15] and [23], due to the addition of some β-proteins), this paper was applied to the data base selected and measured in W.C. Johnson, Jr.'s laboratory. On the other hand, the CCA method separates deconvolution from secondary structure assignment. The deconvolution part, which is totally "free" from external data, can be used for monitoring the effects of a systematic data base truncation without the more problematic problem.

Recently, Venyaminov et al. [14] investigated the effect of the data truncations in the π–π^* region (190–210 nm) of the spectra and found

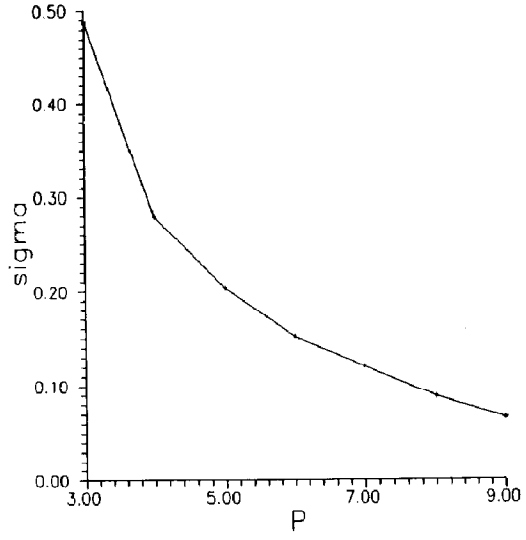

Fig. 1. The standard deviation (σ) dependence on the number of the pure components (P). Note that the inflection point located at $P = 5 \pm 1$ suggests that 5 ± 1 different secondary structures are expected, which agrees with the secondary structure determination of these proteins based on X-ray [19].

that the estimation of secondary structural elements can still be approximately correct using the "Yang deconvolution method". Simultaneously, Toumadje et al. [19] claim that in the determination procedure of some secondary structural elements, using the "variable selection method", the use of an enlarged spectral window (168–260 nm)

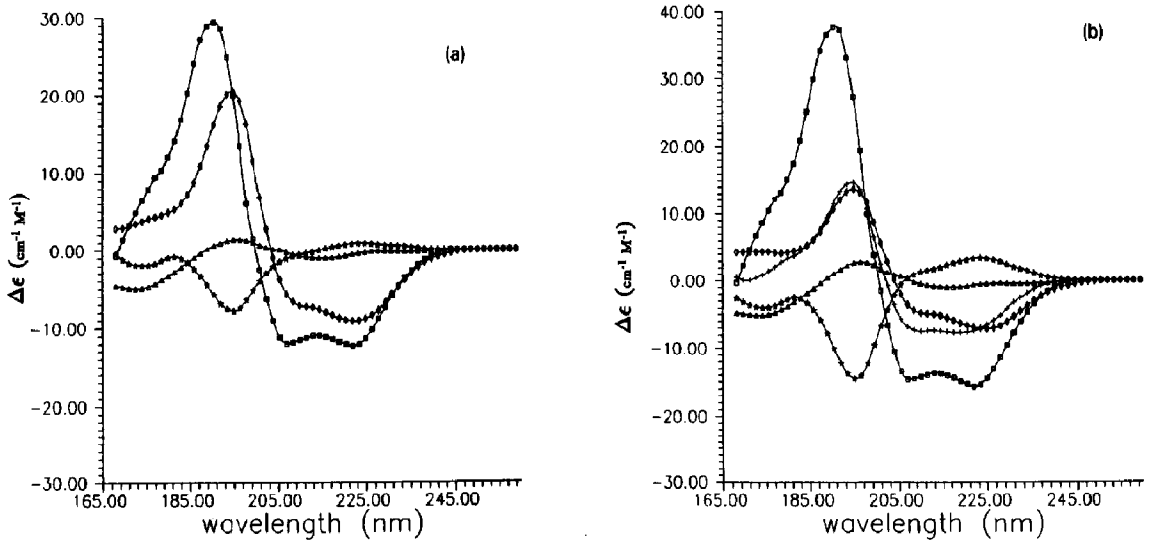

Fig. 2. The shape of the pure CD component curves for $P = 4$ (A), $P = 5$ (B) and $P = 6$ (C).

Table 1

Calculated weights for 16 proteins [a] with four pure component curves for different spectral regions [b]

Spectral region	Protein															
	Ch	Cy	El	He	LD	T4	My	Pa	Ri	Fl	GD	Pr	SN	TI	Ec	PG
168–260 nm [c]																
	10.9	0.0	9.7	2.9	14.1	4.9	38.7	18.2	8.0	0.0	9.1	0.0	2.8	14.4	25.0	95
	17.2	2.0	25.8	0.2	61.8	7.2	25.3	0.0	27.8	48.9	41.8	82.0	57.3	51.7	59.3	0
	6.2	42.4	0.0	74.3	23.9	64.3	36.0	11.5	17.0	27.2	18.7	11.0	29.3	32.7	0.0	0
	65.6	55.5	64.4	22.7	0.1	23.6	0.0	70.3	47.2	23.9	30.5	6.9	10.7	1.2	15.7	5
173–260 nm																
	5.8	0.7	3.0	11.2	17.9	14.0	43.2	13.5	3.7	0.0	7.5	0.3	3.6	17.5	24.0	100
	17.7	4.0	22.2	0.0	57.3	16.1	12.0	0.0	16.8	40.9	33.3	76.0	44.3	39.8	53.3	0
	4.7	43.3	0.0	78.9	24.6	62.6	44.8	11.5	22.6	30.7	22.3	10.5	35.5	39.0	0.1	0
	71.8	52.0	74.8	9.9	0.2	7.4	0.0	74.9	57.0	28.4	36.9	13.2	16.6	3.8	22.7	0
178–260 nm																
	12.7	0.1	9.7	0.1	16.6	10.5	36.6	19.6	4.2	0.0	8.9	2.4	2.9	13.9	26.7	100
	17.0	6.6	15.5	0.0	63.1	40.8	11.3	0.0	6.6	45.7	35.1	80.4	54.6	45.6	49.1	0
	1.7	39.4	0.0	77.8	20.1	48.7	44.4	8.4	25.1	25.9	19.0	6.1	28.2	34.6	0.0	0
	68.5	53.9	74.8	22.1	0.2	0.0	7.7	72.0	64.1	28.3	37.0	11.2	14.2	6.0	24.2	0
183–260 nm																
	10.1	1.2	6.6	1.6	17.4	15.3	38.7	19.3	1.3	0.0	8.4	0.6	5.8	15.9	25.4	100
	11.9	16.1	10.4	0.2	50.5	49.0	13.9	9.1	0.0	36.4	28.0	55.2	56.3	42.1	37.3	0
	14.0	45.0	13.7	95.2	15.1	35.7	47.2	13.9	46.0	29.6	24.3	7.4	16.1	30.2	0.0	0
	64.0	37.7	69.2	2.7	17.0	0.0	0.2	57.7	52.7	34.0	39.3	36.8	21.7	11.8	37.3	0

188–260 nm															
	16.6	0.9	13.0	0.6	15.7	8.6	34.9	22.6	7.0	0.0	10.9	1.6	2.3	11.9	25.4	100
	11.3	15.4	9.7	0.2	51.4	49.2	12.9	7.3	0.0	35.9	27.8	55.8	56.8	42.1	37.3	0
	8.1	46.2	8.2	96.5	15.7	42.1	52.0	12.4	40.5	29.8	21.8	5.3	19.1	34.0	0.0	0
	64.0	37.4	69.1	2.7	17.2	0.1	0.2	57.7	52.5	34.2	39.5	37.2	21.8	11.9	37.3	0
193–260 nm																
	2.0	3.9	0.0	30.7	17.7	16.8	47.4	0.2	15.2	0.9	6.0	4.2	0.2	21.0	19.9	100
	37.3	26.8	50.0	11.6	46.1	16.8	0.2	0.0	61.3	43.6	32.6	89.1	39.8	39.5	57.5	0
	7.5	36.9	0.0	57.7	30.9	60.4	45.7	29.2	2.2	32.7	27.9	6.7	43.1	38.0	6.8	0
	53.2	32.4	49.9	0.0	5.3	6.0	6.7	70.6	21.4	22.8	33.6	0.0	16.9	1.6	15.9	0
198–260 nm																
	26.5	13.9	23.4	22.9	14.2	10.2	44.7	30.4	23.9	5.6	17.4	0.2	0.0	14.6	25.0	100
	11.5	30.0	35.2	38.5	22.4	0.0	2.4	0.2	73.6	37.2	15.8	65.8	24.2	32.5	34.6	0
	6.7	25.2	0.0	38.5	31.7	50.9	33.3	13.6	0.1	27.1	23.2	16.6	38.0	33.6	12.1	0
	55.3	30.9	41.4	0.0	31.8	38.9	19.6	55.8	2.5	30.2	43.7	17.4	37.8	19.3	28.3	0
203–260 nm																
	0.0	13.0	8.4	35.2	10.9	0.3	39.9	9.8	35.7	7.1	6.5	8.8	0.3	21.2	16.0	98
	1.3	41.6	27.1	29.6	30.6	0.0	1.5	18.7	64.3	48.8	28.1	64.3	49.2	45.3	17.9	0
	5.9	24.9	0.0	35.1	30.9	47.2	31.5	15.7	0.1	26.7	23.5	15.8	38.7	33.2	10.0	0
	92.8	20.5	64.4	0.1	27.6	52.4	27.1	55.8	0.0	17.3	42.0	11.0	11.8	0.3	56.1	2

[a] Ch, α-chymotrypsin; Cy, cytochrome *c*; El, elastase; He, hemoglobin; LD, lactate dehydrogenase; T4, T4 lysozyme; My, myoglobin; Pa, papain; Ri, ribonuclease A; Fl, flavodoxin; GD, glyceraldehyde-3-phosphate dehydrogenase; Pr, prealbumin; SN, subtilisin novo; TI, triose-phosphate isomerase; Ec, *Eco* RI; PG, poly (L-glutamic acid).

[b] Data from Ref. [19].

[c] Used also as "reference" coefficient matrix.

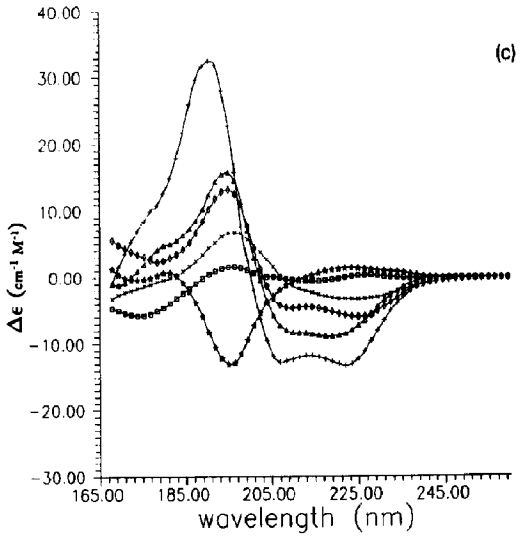

Fig. 2. Continued.

might be essential. Although there is not a direct controversy between the two statements, the main question addressed here is: "Does the enlargement of the spectral window, from the standard 185–260 nm range to the 168–260 nm range, influence the data set deconvolution?"

2. Methods

The CD spectra of the 16 proteins recorded in the 168–260 nm spectral range were taken from Ref. [19]. The Convex Constraint Algorithm [15,16,18] was applied to the overall data base (185 × 16 data points). The Pearson product-moment correlation coefficient (r) and RMS was calculated according to eqs. (2) and (3).

$$r = \left\{ \sum_{i-1}^{N} x_i y_i - \frac{1}{N} \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i \right\} \Bigg/ \left\{ \left[\sum_{i=1}^{N} x_i^2 - \frac{1}{N} \left(\sum_{i=1}^{N} x_i \right)^2 \right] \times \left[\sum_{i=1}^{N} y_i^2 - \frac{1}{N} \left(\sum_{i=1}^{N} y_i \right)^2 \right] \right\}^{\frac{1}{2}} \quad (2)$$

$$\mathrm{RMS} = \sqrt{\frac{\sum_{i=1}^{N} [x_i - y_i]^2}{N-1}} \quad (3)$$

f(x) type	f(x) form	constants	RMS (eq. 3)
exponential + linear	$f(x) = a \exp(-x) + b x + c$	a = +5.50 b = −0.02 c = +0.28	$6.0 * 10^{-3}$
quadratic + linear	$f(x) = a x^2 + b x + c$	a = +0.014 b = −0.229 c = +1.015	$2.7 * 10^{-2}$
hyperbolic + linear	$f(x) = a x^{-1} + b x + c$	a = +2.357 b = +0.020 c = −0.372	$1.3 * 10^{-1}$

Scheme 1. Analytical form of the most probable functions.

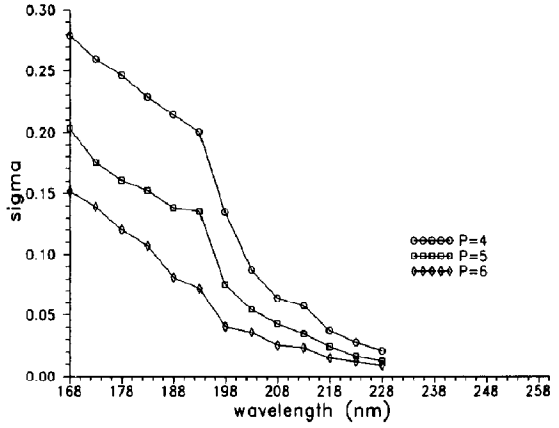

Fig. 3. The residual standard deviation [$\sigma_{168-260}(P)$, $\sigma_{173-260}(P)$, ..., $\sigma_{228-260}(P)$] dependence on the wavelength range of the data set used: (-○-) $P = 4$, (-□-) $P = 5$ and (-◇-) $P = 6$.

3. Results and discussion

It is a fundamental problem to precisely determine the number of the pure conformers (P) in a conformational mixture [18]. The fact that this is an input parameter requires one to have at least some idea of the actual magnitude of P. If one has no idea, one has to perform several runs with different P-values and determine the average standard deviations (σ) at each value of P. As previously shown [18], the dependence of $\sigma(P)$—

Table 2

Comparison of the Pearson product-moment correlation coefficients of the reference [a] weight matrix and the coefficients calculated for the truncated datasets for different spectral regions

Spectral range (nm)	Component				Total [c]
	1 [b]	2	3	4	
168–260	1.00	1.00	1.00	1.00	1.000
173–260	0.98	0.98	0.99	0.97	0.968
178–260	1.00	0.90	0.97	0.95	0.949
183–260	0.99	0.76	0.83	0.81	0.839
188–260	0.99	0.76	0.89	0.80	0.851
193–260	0.91	0.78	0.90	0.90	0.861
198–260	0.96	0.48	0.84	0.52	0.677
203–260	0.87	0.54	0.81	0.57	0.679

[a] The complete data set (168–260 nm) was used as reference matrix.

[b] Pearson "r" value between coefficients related with the first pure component curve of the reference data set and the truncated data set.

[c] Pearson "r" value between the reference matrix and the truncated matrix.

the standard deviation—on P can result in a relatively precise estimation of the optimal P-value. It was suggested that the use of the value of P, from the $\sigma(P) \rightarrow P$ function, where the exponential-like decrease (first part of the curve) turns to a linear-decrease (second part of the curve in Fig. 1) gave the optimal results. Due to uncertainties in the precise location of this point, it was recommended to discuss the deconvolution in the $P \pm 1$ region, which represent the under and the over estimation of the proper P-value. A standard deviation (σ) dependence of the number of the pure components (P) is shown in (Fig. 1). The best fit obtained by curve fitting is reported in Fig. 1, while the analytical form of the most probable functions are shown in Scheme 1.

Among the three most probable function candidates, the application of an exponential plus linear function resulted in the smallest RMS value. The change in curvature is located where the contribution of the exponential curve to the global $f(x)$ function is small enough (Fig. 1). Comparing the magnitude ratios of the exponential versus the linear part of the $f(x)$ function at $P = 5$, $P = 6$, and $P = 7$, one can determine 16%, 84%, 10%, 90%, and 3% and 97%, respectively.

The change in curvature located at $P = 5 \pm 1$ suggests that 5 ± 1 different secondary structures are expected, which harmonizes with the secondary structure determination of these proteins based on X-ray diffraction data [19]. (For the shape of 4, 5 and 6 pure component curves, see Figs. 2A, B and C.) The three sets of curves in Fig. 2 (168–260 nm), which are related to three coefficient matrices, were also used as "reference" matrices during systematic data truncation. Starting from 168 nm, with a step size of 5 nm, the measured CD data were truncated, and the resulting truncated data sets (173–260, 178–260, etc.) were also deconvoluted using $P = 4$, $P = 5$ and $P = 6$. The decreasing tendency of the residual standard deviations [$\sigma_{168-260}(P)$, $\sigma_{173-260}(P)$, ..., $\sigma_{228-260}(P)$] are plotted in Fig. 3. The deconvolution results of the differently truncated data sets yielded coefficient matrices, which are reported in Tables 1 and 4, respectively. One would assume that a larger spectral window-related coefficient matrix would be closer to the "perfect" description of the secondary structural content of the proteins than that of a smaller (truncated) data set. The resulting weights obtained from the deconvolution were used as a "reference" matrix. The reference matrix is the

Table 3

Comparison of the RMS deviations of the reference [a] weight matrix and the coefficients calculated for the truncated data sets for different spectral regions

Spectral range (nm)	Component				Total [c]
	1 [b]	2	3	4	
168–260	0.00	0.00	0.00	0.00	0.000
173–260	0.05	0.08	0.04	0.08	0.064
178–260	0.03	0.12	0.06	0.09	0.081
183–260	0.04	0.18	0.14	0.15	0.139
188–260	0.03	0.18	0.12	0.16	0.134
193–260	0.11	0.17	0.10	0.13	0.130
198–260	0.10	0.25	0.13	0.22	0.187
203–260	0.13	0.23	0.14	0.25	0.194

[a] The complete data set (168–260 nm) was used as reference matrix.

[b] Pearson "r" value between coefficients related with the first pure component curve of the reference data set and the truncated data set.

[c] Pearson "r" value between the reference matrix and the truncated matrix.

Table 4

Calculated weights for 16 proteins [a] with five pure component curves for different spectral regions [b]

Spectral region	Protein															
	Ch	Cy	El	He	LD	T4	My	Pa	Ri	Fl	GD	Pr	SN	TI	Ec	PG
168–260 nm [c]																
	6.8	0.0	4.4	0.0	14.8	11.0	33.2	15.4	0.2	0.8	7.6	0.3	6.9	14.4	21.2	79
	7.7	4.1	10.9	0.1	60.0	27.6	26.1	0.0	7.4	43.9	35.3	68.7	60.3	51.1	49.3	0
	15.8	39.4	7.0	36.2	15.0	61.2	17.5	29.7	0.0	19.0	15.3	0.2	26.9	18.9	0.2	0
	54.6	32.4	56.7	0.0	2.5	0.0	0.2	54.8	40.2	17.1	25.5	12.4	6.0	0.1	23.7	21
	15.0	24.1	20.9	63.7	7.7	0.2	22.9	0.1	52.2	19.2	16.3	18.4	0.0	15.4	5.6	0
173–260 nm																
	10.6	0.1	8.1	0.1	17.2	10.7	37.4	19.0	2.3	0.0	8.6	1.6	4.4	14.8	26.7	97
	20.7	3.9	26.2	0.1	60.2	13.2	11.6	0.0	21.2	43.5	35.8	82.1	45.0	41.1	57.5	0
	6.7	34.0	0.3	36.7	17.4	56.8	28.2	23.1	0.2	21.1	16.7	0.0	35.1	27.4	0.3	0
	50.4	37.4	50.9	0.1	0.1	11.3	0.0	57.9	32.4	19.0	25.6	5.0	15.5	2.9	15.5	2
	11.6	24.6	14.5	63.1	5.2	8.0	22.9	0.0	43.8	16.4	13.3	11.3	0.0	13.8	0.0	0
178–260 nm																
	8.4	0.1	4.7	2.7	18.2	15.6	39.5	17.2	0.0	0.1	8.0	1.0	5.5	16.3	25.2	100
	24.7	5.0	25.5	0.0	56.3	24.6	4.6	0.0	20.1	43.5	34.8	82.1	42.9	37.6	49.1	0
	3.8	33.7	0.2	36.5	20.4	51.3	33.0	23.6	0.0	21.5	17.5	0.0	37.8	30.4	5.5	0
	51.9	37.5	55.9	0.0	0.0	0.1	0.8	59.2	38.3	19.1	26.8	6.2	13.6	2.3	20.2	0
	11.2	23.7	13.7	60.7	5.2	8.4	22.1	0.0	41.6	15.8	12.9	10.8	0.2	13.3	0.0	0
183–260 nm																
	11.6	0.1	9.0	0.1	17.3	11.7	36.1	17.7	4.6	0.0	8.4	3.8	4.2	14.9	26.8	100
	12.4	12.8	13.0	0.0	54.4	44.4	9.3	0.0	8.1	38.3	28.4	68.0	55.9	43.5	42.1	0
	4.9	28.3	0.0	39.1	19.2	43.4	33.4	21.5	0.1	21.3	18.0	0.1	29.5	26.5	4.2	0
	49.4	33.0	50.4	0.0	0.2	0.0	4.7	60.6	28.4	18.8	27.0	1.0	10.4	0.8	17.4	0
	21.7	25.8	27.6	60.7	9.0	0.4	16.5	0.2	58.8	21.5	18.2	27.0	0.0	14.4	9.5	0

188–260 nm															
	17.1	0.0	15.5	0.1	15.4	4.3	32.0	18.6	12.8	0.0	9.8	6.7	0.2	11.6	27.9	100
	7.5	13.7	6.8	0.1	56.3	52.4	12.7	0.1	0.0	37.7	27.1	63.6	60.6	46.4	40.7	0
	6.4	27.5	0.0	44.1	19.3	42.3	34.3	21.1	1.7	22.0	20.0	0.0	27.7	25.4	1.5	0
	49.7	29.3	47.7	0.3	2.7	1.0	4.5	60.1	23.9	19.5	29.6	1.5	11.3	0.0	14.6	0
	19.3	29.5	30.0	55.4	6.3	0.0	16.6	0.2	61.6	20.9	13.4	28.3	0.2	16.6	15.3	0
193–260 nm																
	21.2	0.1	16.5	0.1	17.7	7.0	33.3	14.1	11.2	0.0	13.0	7.0	0.0	12.3	27.0	100
	24.5	17.7	18.4	0.0	57.2	52.4	14.0	0.0	4.4	38.4	35.9	63.0	56.8	44.3	41.1	0
	5.6	25.1	0.0	38.6	18.4	39.6	31.0	22.9	0.0	20.8	18.6	0.1	27.4	23.5	2.4	0
	48.4	29.2	47.3	0.4	2.7	0.9	4.1	63.0	24.0	20.0	29.0	1.3	12.5	0.0	15.0	0
	0.3	7.9	17.8	60.9	4.0	0.1	17.5	0.0	60.4	20.7	3.5	28.6	3.3	19.9	14.5	0
198–260 nm																
	0.0	13.7	5.4	24.3	10.6	0.1	38.0	19.2	23.5	6.7	10.1	0.1	4.6	19.6	8.3	100
	19.0	46.7	18.2	0.1	39.5	33.8	12.4	59.7	0.2	48.8	49.6	24.6	70.0	41.2	0.0	0
	36.4	7.0	15.3	49.7	24.9	63.8	47.5	0.1	0.0	6.4	12.9	7.8	4.0	14.0	40.3	0
	44.4	10.6	43.6	0.1	9.1	2.3	1.1	21.0	30.1	11.1	16.4	24.3	0.2	0.1	34.4	0
	0.2	21.9	17.5	25.8	15.9	0.0	1.0	0.0	46.2	27.0	10.9	43.2	21.3	25.0	17.0	0
203–260 nm																
	13.2	32.5	0.3	0.4	19.3	0.0	25.8	17.3	0.0	5.9	22.4	3.5	0.0	4.6	5.9	100
	9.0	59.9	1.4	0.8	43.6	22.1	0.7	23.2	1.0	43.3	44.5	42.7	52.4	29.6	0.0	0
	3.0	1.2	0.7	52.5	19.5	55.0	45.6	9.5	3.3	14.7	8.0	0.0	26.4	32.5	15.7	0
	74.7	6.4	67.0	2.9	11.9	22.7	13.8	42.7	28.9	13.0	24.9	18.2	1.7	0.0	53.1	0
	0.1	0.0	30.6	43.4	5.7	0.2	14.1	7.4	66.8	23.0	0.2	35.6	19.4	33.3	25.4	0

[a] For definitions, see Table 1.
[b] Data from Ref. [19].
[c] Used also as "reference" coefficient matrix.

Table 5

Comparison of the Pearson product-moment correlation coefficients of the reference [a] weight matrix and the coefficients calculated for the truncated datasets for different spectral regions

Spectral range (nm)	Component					Total [c]
	1 [b]	2	3	4	5	
168–260	1.00	1.00	1.00	1.00	1.00	1.000
173–260	1.00	0.91	0.94	0.94	0.98	0.946
178–260	1.00	0.90	0.89	0.96	0.97	0.937
183–260	0.99	0.95	0.87	0.94	0.98	0.943
188–260	0.96	0.94	0.86	0.92	0.95	0.923
193–260	0.96	0.92	0.87	0.92	0.93	0.914
198–260	0.89	0.35	0.45	0.81	0.57	0.600
203–260	0.87	0.51	0.63	0.77	0.69	0.691

[a] The complete data set (168–260 nm) was used as reference matrix.

[b] Pearson "r" value between coefficients related with the first pure component curve of the reference data set and the truncated data set.

[c] Pearson "r" value between the reference matrix and the truncated matrix.

coefficient matrix obtained from the largest data set (168–260 nm range). The calculated Pearson product-moment correlation coefficients, as well as the RMS values, associated with the appropriate "truncated" data set yielded coefficient matrices (the first one is the reference matrix) which are reported in Tables 2, 3 and 5, 6, respectively. If the data set is deconvoluted to five components ($P = 5$), the total (r_{total}), as well as the individual (r_i) correlation coefficients, decreases from unity with the truncation of the data base (see Tables 2 and 5). However, the claim [19] that the use of the expanded 168–180 nm region will significantly improve the deconvolution due to the introduction of "new CD bands", was not confirmed. The comparison of the r_{total}, as well as the r_i values, demonstrates that a significant alteration of the coefficient values is expected only if the region of the π–π^* transition is ignored (e.g., data base containing only the 203–260 nm range). Although the estimation of some secondary structural elements may still remain reasonable (e.g. the coefficients of the first pure curve (probably related to an α-helix)) the neglect of the spectral range below 200 nm drastically affects the reliability of the assignment. This observation harmonizes perfectly with the recent analysis of Venyaminov et al. [14] and with the well known "rule of thumb" that the α-helix, and only the α-helix, content of a protein can be roughly determined at 280 nm [22]. By contrast, the spectral range enlargement below 183 nm results in only modest, if any, improvement simply because the determined conformational weights are almost identical for coefficients calculated from the 168–260 nm, 173–260 nm, 178–260 nm and 183–260 nm regions. The observed change is not larger than that expected from a simple signal-to-noise ratio improvement by introduction of more data points.

Table 6

Comparison of the RMS deviations of the reference [a] weight matrix and the coefficients calculated for the truncated data sets for different spectral regions

Spectral range (nm)	Component					Total [c]
	1 [b]	2	3	4	5	
168–260	0.00	0.00	0.00	0.00	0.00	0.000
173–260	0.05	0.10	0.06	0.07	0.04	0.070
178–260	0.06	0.11	0.08	0.06	0.05	0.075
183–260	0.06	0.08	0.08	0.08	0.04	0.070
188–260	0.08	0.09	0.09	0.09	0.06	0.082
193–260	0.09	0.10	0.09	0.09	0.07	0.087
198–260	0.12	0.27	0.20	0.14	0.16	0.185
203–260	0.13	0.24	0.16	0.16	0.15	0.169

[a] The complete data set (168–260 nm) was used as reference matrix.

[b] Pearson "r" value between coefficients related with the first pure component curve of the reference data set and the truncated data set.

[c] Pearson "r" value between the reference matrix and the truncated matrix.

4. Conclusions

Using the CCA deconvolution method, we have shown that, for the 16 proteins reported herein, the spectral enlargement (from 183–260 nm to 168–260 nm) does not introduce a significant alteration of the coefficient matrices. Thus, the enlargement of the spectral window in this range does not significantly alter the resulting weight matrices by the deconvolution. The similarity of these coefficient matrices (Tables 1 and 3), which

is directly related to the secondary structural percentages, is higher then 0.9 ($r_{total} > 09$) and the RMS is very small. Comparing this degree of similarity with the dissimilarity of the secondary structure percentages determined by X-ray (for some cases $r = 0.5$ or smaller), it appears that spectral window enlargement is not necessarily the unique or the optimal way of improving analytical CD spectroscopy for use in protein structural determination. The development and the analysis of a larger protein data base incorporating highly "dissimilar proteins" may result in a better understanding of how CD is correlated with the secondary structural elements of proteins.

Acknowledgements

We would like to thank Professor W.C. Johnson, Jr. for supplying the CD data set on a diskette. This research was supported in part by grants from NSF (DMB-8713193) and the U.S. Army Research Office (89-K-0088).

References

1 G.M. Holzwarth and P. Doty, J. Am. Chem. Soc. 87, (1965) 218–228.
2 J.T. Yang, C.-S.C. Wu and H.M. Martinez, in: Methods in enzymology 130 (Academic Press, San Diego, CA, 1986) pp. 208–269.
3 R.W. Woody, Biopolymers 17 (1978) 1451–1467.
4 J. Applequist, Biopolymers 21 (1982) 779–795.
5 R.W. Woody, in: The peptides 7, ed. V.J. Hruby (Academic Press, San Diego, CA, 1985) pp. 15–113.
6 C.M. Manning and R.W. Woody, Biopolymers 26 (1987) 1731–1752.
7 S.W. Provencher and J. Glockner, Biochemistry 20 (1981) 33–37.
8 J.P. Hennessey and W.C. Johnson Jr., Biochemistry 20 (1981) 1085–1094.
9 P. Manavalan and W.C. Johnson Jr., Anal. Biochem. 167 (1987) 76–85.
10 M. Levitt, J. Mol. Biol 104 (1976) 59–107.
11 P. Chou and G.D. Fasman, J. Mol. Biol. 115 (1977) 135–175.
12 W. Kabsch and C. Sander, Biopolymers 22 (1983) 2577–2637.
13 C.M. Wilmot and J.M. Thorton, Protein Eng. 3 (1990) 479–493.
14 S.Y. Venyaminov, I.A. Baikalov, C.-S.C. Wu and J.T. Yang, Anal. Biochem. 198 (1991) 250–255.
15 A. Perczel, G. Tusnady, M. Hollosi and G.D. Fasman, Protein Eng. 4 (1991) 669–679.
16 A. Perczel, G. Tusnady, M. Hollosi and G.D. Fasman, Croatica Chim. Acta. 62 (1989) 189–200.
17 G. Wagner, S.G. Hyberts and T.F. Havel, Annu. Rev. Biophys. Biomol. Struct. 21 (1992) 167–198.
18 A. Perczel, K. Park and G.D. Fasman, Anal. Biochem. 203 (1992) 83–93.
19 A. Toumadje, S.W. Alcorn and W.C. Johnson Jr., Anal. Biochem. 200 (1992) 321–331.
20 S. Brahms and J. Brahms, J. Mol. Biol. 138, (1980) 149–178.
21 W.C. Johnson Jr. and I. Tinoco Jr., J. Am. Chem. Soc. 94 (1972) 4389–4390.
22 N. Greenfield and G.D. Fasman, Biochemistry 8 (1969) 4108–4116.
23 A. Perczel, K. Park and G.D. Fasman, Proteins Struct. Funct. Genet. 13 (1992) 57–69.