Biophysical Chemistry 7 (1977) 223–227
© North-Holland Publishing Company

EQUIVALENCY OF LINEAR LEAST SQAURES CURVE FITTING AND RECIPROCAL FUNCTIONS IN PROTEIN CIRCULAR DICHROIC SPECTRA ANALYSIS *

Gordon WILLICK and Michael ZUKER
Division of Biological Sciences, National Research Council of Canada,
Ottawa, Canada K1A 0R6

Received 12 May 1977

Two methods for the analysis of the circular dichroism spectra of proteins for determination of secondary structure have been examined. These are the linear curve fitting of the data, minimized in the least squares sense, and the method of reciprocal functions proposed by C.C. Baker and I. Isenberg, Biochemistry 15 (1976) 629. It is shown that the use of these two methods give results that are identical, providing the same set of reference spectra are used in each case, and, therefore, that no new information is obtained by the use of either one over the other.

1. Introduction

In recent years a number of methods have been suggested for analyzing the far UV circular dichroism (CD) spectra of proteins in terms of the principle known secondary structure elements, the α-helix and the β-sheet. The simplest approximation is to assume that the observed spectra can be represented by a linear combination of the α-helix, β-sheet, and a remaining portion, usually assumed to be a type of spectra associated with a random coil polypeptide:

$$[\theta]_{\text{obs}} = f_H[\theta]_H + f_\beta[\theta]_\beta + f_R[\theta]_R \ , \qquad (1)$$

where $[\theta]_{\text{obs}}$ = observed ellipticity on a mean residue weight basis and $[\theta]_H$, $[\theta]_\beta$, $[\theta]_R$ are the ellipticities associated with the α-helix, β-sheet, and random coil, respectively. f_H, f_β, and f_R are the fractions of each of these components.

These latter are the quantities to be determined. Often a further constraint is imposed that $f_R = 1 - f_H - f_\beta$. This last contraint may be explicit, or, in a sense implicit in the choice of reference spectra, i.e. the $[\theta]_H$, $[\theta]_\beta$, and $[\theta]_R$.

An early proposal for a set of reference spectra was that of poly (L-lysine), which under appropriate conditions exists in the α-helix, β-sheet, and random form [1]. More recently, Chen et al. [2] have proposed the use of proteins with crystallographically known structures to determine the set of reference spectra. They chose five proteins, which had structures known to 2 Å resolution and which had varying α-helical and β-sheet content, and accurately measured their CD spectra. In addition, they attempted to take into account the effect of the chain length of the α-helix on its contribution. However, we shall restrict ourselves here to analysis of the spectra using eq. (1) and the reference spectra of Chen et al. [2].

More recently, Baker and Isenberg [3] have proposed an analysis of the CD data using reciprocal functions and integration of these with the observed ellipticities to determine the f_H, f_β, f_R. In their paper they have made certain statements that imply that the use of these functions is in some ways superior to curve fitting procedures. We shall here show that, in fact, their method is identically equivalent to a linear, least squares curve fitting procedure, and therefore provides no new information.

2. Materials and methods

Two of the proteins used in this study were acidic proteins from the large subunit of bacterial ribosomes. The L_7 protein from E. coli and the corresponding

* NRCC number 16195.

protein L_{20} from the as yet unidentified moderate halophile H'X' were obtained by previously published procedures [4]. Concentrations were determined by amino acid analysis. The spectra are reported in units of deg. cm^2 per decimean residue weight (dMRW), where the dMRW for L_7 was taken as 98.1, for H'X' L_{20} as 103, and for rabbit skeletal tropomyosin as 119.7. The circular dichroic spectra were taken with the Cary 61 circular dicrograph operating at its ambient temperature, 27°.

3. Theory

We write the observed CD spectra, $[\Theta]$, in terms of a linear combination of the three reference spectra:

$$[\Theta] = \sum_{i=1}^{3} f_i\,[\theta]_i \ ,$$

where the $[\theta]_i$ are the reference spectra for the α-helix, β-sheet, and random coil [2], and the f_i, the coefficients to be determined, represent the fractions of these three components. We have imposed no constraint on the value of Σf_i.

Two methods have been used to fit the data. The first, a curve-fitting method, minimizes the residuals $[\Theta]_{\mathrm{obs}} - [\Theta]_{\mathrm{calc}}$ in the least squares sense. The second uses the reciprocal function method of Baker and Isenberg [3]. In this method they define functions ϕ_j such that:

$$\phi_j = \sum_{i=1}^{3} \xi_{ji}\,[\theta]_i \ , \quad j = 1,2,3, \tag{2}$$

with the ξ_{ji} determined by the condition:

$$\int_{\lambda_1}^{\lambda_2} \phi_j\,[\theta]_i\,\mathrm{d}\lambda = \delta_{ji} \ . \tag{3}$$

The coefficients f_i are determined by evaluating the integrals:

$$f_i = \int_{\lambda_1}^{\lambda_2} \phi_i\,[\Theta]\,\mathrm{d}\lambda \ .$$

Proof of the equivalence of these two methods is given in the Appendix.

4. Results

Table 1 summarizes the results obtained with two sets of synthetic data, and the three sets of experimental data at our disposal. The synthetic data bear no relationship to the circular dichroic data. In the first row the data has no error, in the second the data has been subjected to a 20% random error. The coefficients determined by the linear least squares fitting procedure and by the use of the reciprocal functions are identical. They remain equal as the noise is introduced, i.e. neither one has any advantage over the other insofar as being less subject to random error. The difference seen in coefficient A_3 at the 20% noise level between the two methods of analysis was determined to be due to round-off error. When the same calculation was carried out with double precision arithmetic, the coefficients became equal to 0.6889885. In the case of the experimental data, the coefficients A_1, A_2, and A_3 correspond to f_α, f_β, and f_R, respectively. We note that the same coefficients are obtained when the data is analyzed by either of the two methods.

Table 1

Data set	Coefficients (linear least square)					Coefficients (reciprocal functions)				
	A_1	A_2	A_3	SSQ [a)]	Sum [b)]	A_1	A_2	A_3	SSQ [a)]	Sum [b)]
synthetic	4.7143	3.7143	3.0000	1.4286	11.4286	4.7143	3.7143	3.0000	1.4286	11.4286
synthetic –20% error	4.0192	6.6460	0.68899	7.3610	17.5551	4.0192	6.6460	0.68902	7.3611	17.5554
E. coli L_7	0.52233	0.22267	0.72882	0.45445×10^8	1.47382	0.52233	0.22267	0.72882	0.45445×10^8	1.47382
HX L_{20}	0.43688	0.15488	1.0817	0.64465×10^8	2.26522	0.43688	0.15488	1.0817	0.64465×10^8	2.26522
tropon.yosin	1.2870	0.36245	0.9871	0.84018×10^8	2.63656	1.2870	0.36245	0.98710	0.84018×10^8	2.63655

[a)] Sum of squared residuals. [b)] Sum of coefficients.

The actual experimental spectra used for the tests of the methods are shown in figs. 1,2 and 3. These are *representative* spectra for E. coli L_7 and H'X' L_{20} both of which are the acidic proteins from the 50S ribosomal subunits of these bacteria. These proteins have a considerable sequency homology (A.T. Matheson, personal communication) and are predicted to have similar secondary structure homology by the empirical method of Chou and Fasman [5]. However, it is obvious from their CD spectra that there is some difference in their secondary structure. The visual fit is obviously much worse for that of H'X' L_{20} (fig. 2) than that for E. coli L_7 (fig. 1). This is reflected in the much higher sum of squared residuals for the former over that of the latter protein (table 1). There is a relatively close equivalence of the calculated fraction of α-helix, β-sheet, and random coil, despite the spectral difference.

A CD spectra for tropomyosin is shown in fig. 3. This protein has a known primary sequence and the secondary structure can be predicted from this sequence and other physical chemical data [6]. Tropomyosin is predicted to be 100% in the α-helical

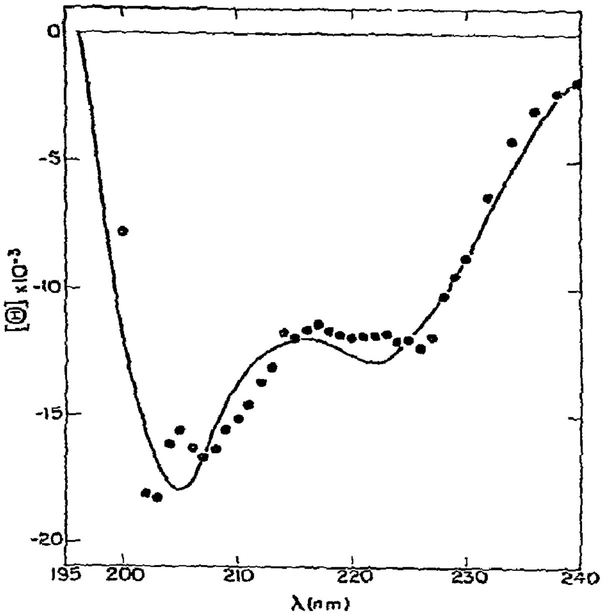

Fig. 2. CD spectra of H'X' L_{20} in 0.1 M potassium chloride, 0.001 M potassium phosphate, pH 7.4.

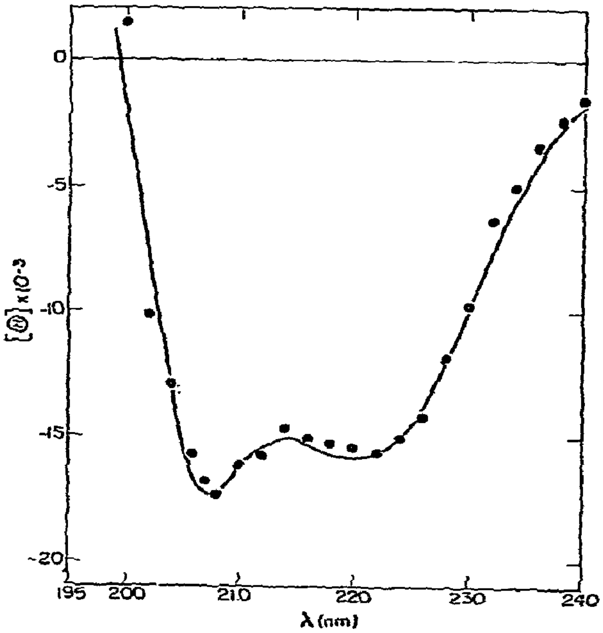

Fig. 1. CD spectra of E coli L_7 in 0.1 M potassium chloride, 0.001 M potassium phosphate, pH 7.4.

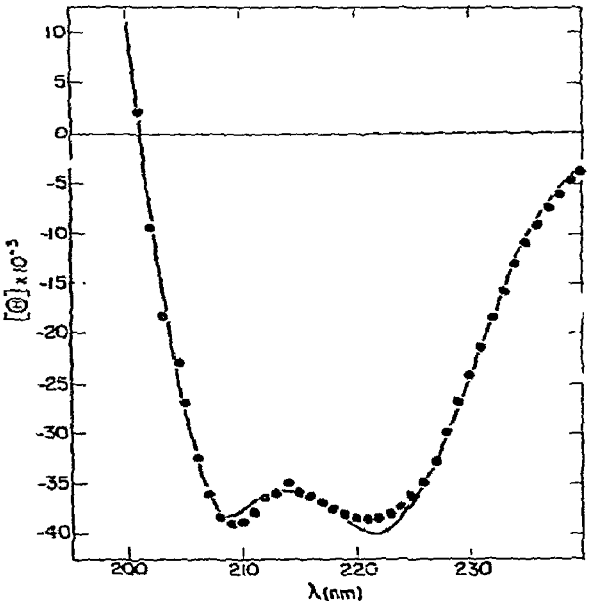

Fig. 3. CD spectra of tropomyosin in 0.6 M potassium chloride, 0.05 M potassium phosphate, pH 7.0.

conformation. The fraction of α-helix is calculated to be 1.287 with additional values for f_β and f_R which added together give a sum of coefficient equal to 2.63656 (table 1). Visually, the calculated values fit the curve quite well. The sum of the coefficients is 2.64 and reflects the fact that the spectral intensity exceeds that of the pure α-helix reference spectra.

5. Discussion

The principal and most important point is that a linear least squares fitting procedure and the integrative method using reciprocal functions proposed by Baker and Isenberg [3] lead to numerically equivalent results if the same reference spectra are used for both calculations. The method proposed by Baker and Isenberg [3] may or may not be computationally easier, but it definitely leads to no further insight into the data or choice of reference spectra.

A further question that should be considered is whether the deviation of the calculated coefficients from summing to unity says anything about the choice of reference spectra. The reference spectra used in these calculations have in a sense been normalized since the values for $[\theta]_R$ have been obtained from the valueof $1 - [\theta]_\alpha - [\theta]_\beta$ [2]. If the spectra were a correct set, *both* in a physical and mathematical sense, then we would expect a perfect fit to the experimental curve and the coefficients would sum to unity. However, we can easily imagine spectra which would give perfect fits in a mathematical sense, but which would suggest the reference spectra are obviously physically incorrect, e.g. a case of a perfect helical spectrum for which $f_\alpha > 1$. Such a case is that of the tropomyosin data (fig. 3). The curve is fitted quite well by using a value of f_α equal to 1.29, plus additional contributions from the β and R spectra (table 1). We happen to know that tropomyosin is a coiled pair of α-helices, and therefore it would normally be surmised that this is the reason for its intense α-helical type of spectrum. In other words, it would appear the α-helix reference spectrum is not an appropriate choice for this molecule. If we knew nothing else about the molecule other than its CD spectrum we would know that the α-helical content was high and that something was wrong in our assumptions, since the value for $f_\alpha > 1$ and the sum of the coefficients is much greater than one. However, the latter facts may say nothing about where the error lies.

Whenever a curve is well-fitted, in the sense that the sum of squares is approximately zero, the coefficients are mathematically correct regardless of whether they sum to one or not. On the other hand, if the sum of the coefficients is equal to one, we are not assured that the fit to the data is good. In a physical sense, the fact that coefficients fail to sum to one simply indicates that there is an error in the choice of one or more of the reference spectra, or the assumption inherent in the use of eq. (1), or of both.

Acknowledgement

The authors thank Dr. Donald Miller for helpful discussions and critically reviewing the manuscript, Mrs. Joanne Ridgeway for computer programming, and Dr. Cyril Kay for the tropomyosin CD spectrum.

Appendix

Proof of equivalence of reciprocal vectors method and least squares curve fitting

Given vectors ϕ, ϕ_1, ϕ_2, and ϕ_3 of size n, we wish to determine three numbers α_1, α_2, and α_3 so that the sum of squares

$$\sum_{j=1}^{n} [\Phi_j - \sum_{i=1}^{3} \alpha_i \phi_{ij}]^2$$

is minimized. This is a standard least squares problem and is best expressed in matrix notation. Hence, let us define the following matrices:

$$X \equiv \begin{pmatrix} \phi_1' \\ \phi_2' \\ \phi_3' \end{pmatrix}_{3\times n} ; \quad A \equiv \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix}_{3\times 1} ; (\Phi) \equiv (\Phi)_{n\times 1} .$$

It is assumed that ϕ_1, ϕ_2, and ϕ_3 are independent, so that X has rank 3. The model that corresponds to this problem is:

$$\Phi' = A'X + \epsilon' , \qquad (1)$$

where $\epsilon' = (\epsilon_1,...,\epsilon_n)$ is normally distributed random error, and Φ', A', and ϵ' denote the transposes of Φ, A, and ϵ. The least squares estimate of A is given by:

$$A = (XX')^{-1}\, X\Phi \tag{2}$$

where XX' is non-singular because X has rank 3. The corresponding estimated value of Φ, $\hat{\Phi}$, is given by:

$$\hat{\Phi}' = \Phi' X' (XX')^{-1}\, X \tag{3}$$

The reciprocal vectors method of determining $\hat{\Phi}'$ proceeds as follows. Let ψ_1, ψ_2, and ψ_3 be three vectors such that

$$\psi_i = \sum_{j=1}^{3} e_{ij} \phi_j \tag{4}$$

$$\text{and } \psi_i \cdot \phi_j = \delta_{ij} \ . \tag{5}$$

To translate into matrix notation, let

$$E \equiv (e_{ij})_{3\times 3} \text{ and } Y \equiv \begin{pmatrix} \psi'_1 \\ \psi'_2 \\ \psi'_3 \end{pmatrix}_{3\times n} \ .$$

Then (4) and (5) imply that:

$$Y = EX \tag{6}$$

$$\text{and } YX' = I \tag{7}$$

We now attempt to represent Φ by least-squares using the vectors ψ_i, $i = 1,2,3$. The corresponding model is

$$\Phi' = B'Y + \epsilon' \ , \tag{8}$$

where $B \equiv \begin{pmatrix} \beta_3 \\ \beta_2 \\ \beta_3 \end{pmatrix}$

We claim that B is given by $B = X\Phi$ and that $\hat{\Phi}' = B'Y$ is the same estimate of Φ as that given by eq. (3).

In order to prove this, it suffices to show that $B = X\Phi$ and that

$$B'Y = \Phi' X' (XX')^{-1} X \ .$$

Using eqs. (6) and (7) we get:

$$EXX' = I$$

and therefore, $E = E' = (XX')^{-1}$.

Using eq. (2) with X replaced by Y, we obtain:

$$B = (YY')^{-1}\, Y\Phi$$

Using eq. (6) $B = (EXX'E)^{-1} EX\Phi = (EE^{-1}E)^{-1} EX\Phi = X\Phi$.

Therefore, $B'Y = \Phi' X' EX = \Phi' X' (XX')^{-1} X$.

References

[1] N. Greenfield and G.D. Fasman, Biochemistry 8 (1969) 4108.

[2] Y. Chen, J.T. Yang and K.H. Chan, Biochemistry 13 (1974) 3350.

[3] C.C. Baker and I. Isenberg, Biochemistry 15 (1976) 629.

[4] C. Terhorst, W. Möller, R. Laursen and B. Wittman-Liebold, Eur. J. Biochem. 34 (1973) 138.

[5] D.Y. Chou and G.D. Fasman, Biochemistry 13 (1974) 222.

[6] D. Stone, J. Sodak, P. Johnson and L.B. Smillie, Fed. Eur. Biochem. Soc. Meet. (Proc.) 31 (1974) 125.