On the Analysis of the Optical Rotatory Dispersion of Proteins*

M. E. Magar†

ABSTRACT: A method is described to analyze the optical rotatory dispersion of proteins and polypeptides in terms of linear combination of reference configurations. This is done by computing an approximation in the L_2 norm.

In this manner the percentage of various reference forms is found by a single computation instead of trial and error. While we have analyzed data for lysozyme and myoglobin in terms of the α helix, random coil, and β form, the method is capable in principle of analyzing data in terms of any number of reference configurations. If contributions from aromatic groups and disulfide bonds are precisely assessed, then they may be readily incorporated in the analysis.

A recent paper by Greenfield *et al.* (1967) analyzes the optical rotatory dispersion of the proteins myoglobin and lysozyme and poly L-lysine in terms of a linear combination of three reference configurations (α helix, β form, and random coil). These authors do this by generating linear combinations of the three reference configurations. Next, by comparing the generated curve to the experimental curve they select the best fit as the one having the lowest variance. The authors stated that their program did not converge to the best fit because it limited its comparisons to generated curves only. We also note that Sarkar and Doty (1966) used the same method of generating curves to analyze their data.

In a previous abstract we have outlined an analytical method to do the precise job the above authors are doing by trial and error. In view of the apparent utility of this method we reproduce it below.

Theory and Results

The problem can be stated in the most general terms as follows. Given the optical rotatory dispersion of a protein or polypeptide $f(\lambda)$ it is required to analyze it in terms of a linear combination of $f_1(\lambda)$, $f_2(\lambda)$, $\ldots f_i(\lambda)$, and $\ldots f_n(\lambda)$, where the $f_i(\lambda)$ terms are reference conformations, *e.g.*, α helix, random coil, β form, etc. We require the determination of the best values of α_i in the equation

$$f(\lambda) \simeq \sum_{i=1}^{n} \alpha_i f_i(\lambda) \qquad (1)$$

We suppose the existence of an approximating function $\Phi(\alpha,\lambda)$ which is a linear combination of the $f_i(\lambda)$ such that

$$\Phi(\alpha,\lambda) = \sum_{i=1}^{n} \alpha_i f_i(\lambda) \qquad (2)$$

For a given $f(\lambda)$ we wish to find the α_i's so that

$$\left\{\int_a^b [f(\lambda) - \Phi(\alpha,\lambda)]^2 d\lambda\right\} \qquad (3)$$

is a minimum. This criteria is called an approximation in the L_2 norm (Rice, 1964).[1] This minimum will occur for all α_i when

$$\frac{\partial\left\{\int_a^b [f(\lambda) - \Phi(\alpha,\lambda)]^2 d\lambda\right\}}{\partial \alpha_i} = 0 \qquad i = 1,2\ldots n \qquad (4)$$

Carrying out the differentiation we obtain

$$\int_a^b \Phi(\alpha,\lambda) f_i(\lambda) d\lambda = \int_a^b f(\lambda) f_i(\lambda) d\lambda \qquad i = 1,2\ldots n \qquad (5)$$

or

$$\sum_{j=1}^{n} \alpha_i \int_a^b f_i(\lambda) f_j(\lambda) d\lambda = \int_a^b f(\lambda) f_i(\lambda) d\lambda \qquad i = 1,2\ldots n \qquad (6)$$

* From the Department of Chemistry, University of California, San Diego at La Jolla, California. *Received August 9, 1967.* This work was supported by a U. S. Public Health Service Grant GM-11916 to Professor B. H. Zimm. Presented at the Pacific Slope Biochemical Conference held at the University of Oregon, Eugene, Ore., Aug 1966.

† Present address: Department of Chemistry, University of Montana, Missoula, Mont. 59801.

[1] The exact origin of the method of computing an approximation in the L_2 norm has been disputed in the mathematical literature since 1805. The rival claimants are the two celebrated mathematicians Johann Carl Friedrich Gauss (1777–1855) and Adrien Marie Legendre (1752–1833). The consensus is that they discovered it independently but Legendre published first.

The previous n equations are linear in the undetermined coefficients (α_i) and they can be solved by the usual techniques. In matrix notation they can be written as follows

$$\begin{bmatrix} \int_a^b f_1(\lambda)f_1(\lambda)d\lambda & \int_a^b f_1(\lambda)f_{n2}(\lambda)d\lambda \ldots & \int_a^b f_1(\lambda)f_n(\lambda)d\lambda \\ \int_a^b f_2(\lambda)f_1(\lambda)d\lambda & \int_a^b f_2{}^2(\lambda)d\lambda \ldots & \\ & \cdots\cdots\cdots\cdots & \\ \int_a^b f_n(\lambda)f_1(\lambda)d\lambda & \ldots & \int_a^b f_n{}^2(\lambda)d\lambda \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \cdot \\ \cdot \\ \alpha_n \end{bmatrix} = \begin{bmatrix} \int_a^b f(\lambda)f_1(\lambda)d\lambda \\ \int_a^b f(\lambda)f_2(\lambda)d\lambda \\ \cdots\cdots \\ \int_a^b f(\lambda)f_n(\lambda)d\lambda \end{bmatrix} \tag{7}$$

To carry out the actual computation we used only three reference configurations (α helix, random coil, and β form). The data for the α helix and random coil were taken from the graphs of Carver *et al.* (1966). The β form has been recently characterized by Sarkar and Doty (1966), Davidson *et al.* (1966), and Iizuka and Yang (1966). In our computation our values were taken from the graphs of Sarkar and Doty (1966). The interval of integration was set with $a = 190$ mμ and $b = 250$ mμ. The actual integration and matrix inversion were subroutines of the University of California San Diego Computer Center and a CDC 3600 computer was used to carry out the computations.

With the above reference conformations and with the above interval of integration we analyzed the data of Harrison and Blout (1965) on myoglobin and the data of Tomimatsu and Gaffield (1965) on lysozyme. Our conclusions were that the best fit for lysozyme corrected to the nearest integer and normalized were necessary so that

$$\sum_{i=1}^{n} \alpha_i = 100\% \tag{8}$$

was 21% α helix, 33% β form, and 46% random coil and the best fit for myoglobin was 55% α helix, 35% β form, and 10% random coil. Those values coincide closely with the results of Greenfield *et al.* (1967).

Discussion

Inasmuch as our results are close to Greenfield *et al.* (1967) and inasmuch as they have already commented on their results and the possible causes of the lack of agreement with X-ray diffraction data, we shall refrain from commenting on the above results except to agree that one of the paramount reasons for disagreement is the contribution of aromatic amino acids and disulfide linkages in this region (Fasman *et al.*, 1964, 1965; Beychok, 1965).

Further progress in this field appears to depend on the assessment of the contribution of the aromatic groups and disulfide linkages in the wavelength range for which the analysis is being conducted. This in turn will require a great deal of further experimentation. In the event of successful assessment of these contributions they can be incorporated quite naturally in the above analysis. Suppose $\psi(\lambda)$ represents the contribution of the aromatic groups and disulfide linkages. The effect of this function will be to change the column vector on the right-hand side of the matrix (eq 7) thus

$$\begin{bmatrix} \int_a^b [f(\lambda) - \psi(\lambda)]f_1(\lambda)d\lambda \\ \int_a^b [f(\lambda) - \psi(\lambda)]f_2(\lambda)d\lambda \\ \vdots \\ \int_a^b [f(\lambda) - \psi(\lambda)]f_n(\lambda)d\lambda \end{bmatrix} \tag{9}$$

The nature of $\psi(\lambda)$ will necessarily be complicated. For example, it may be written as

$$\psi(\lambda) = f_{arm}(\lambda) + f_{disul}(\lambda) \tag{10}$$

where $f_{arm}(\lambda)$ is the contribution due to aromatic groups and $f_{disul}(\lambda)$ is the contribution of the disulfide bonds. It might be found desirable to split the aromatic contribution, in which case

$$\psi(\lambda) = f_{Tyr}(\lambda) + f_{Trp}(\lambda) + f_{disul}(\lambda) \tag{11}$$

where $f_{Tyr}(\lambda)$ is the contribution due to tyrosine and $f_{Trp}(\lambda)$ is the contribution due to tryptophan. One must also bear in mind that the form of $\psi(\lambda)$ will depend to a large extent on the secondary structure. Whatever the final form of $\psi(\lambda)$ and whatever its dependence on secondary structure, it is clear that those factors have to be elucidated before further progress can be made. It is also clear that the task of elucidating $\psi(\lambda)$ falls squarely on the shoulders of the empirical scientists working in this field.

With respect to the method itself, we note that our criteria of the best fit are the same as those of Greenfield *et al.* (1967). The advantage of our method is that it does not depend on trial and error but that the coefficients determined with this method are the best coefficients if the criteria of best fit are the minimization of the variance; with the trial and error method it is possible that a linear combination of reference forms

which has not been generated may fit the data better. Further, in addition to being the best coefficients, the theory of the approximation of functions in the L_2 norm states that the α_i's are *unique* provided the $f_i(\lambda)$'s are linearly independent (Rice, 1964). This is the case for the three reference functions used above. This method also has the advantage that it can be extended to any number of reference conformations. If more reference forms are discovered, the trial and error method will be extremely cumbersome if not impossible.

We must point out however that approximation in the L_2 norm is not the only method that can be used for fitting the data. It is, however, the simplest to compute. In general the theory of the approximation of functions generally seeks to minimize the quantity

$$\left[\int_a^b f(\lambda) - \Phi(\alpha,\lambda)^p d\lambda\right]^{1/p} \qquad (12)$$

where p is generally (but need not necessarily be) an integer. The value of p denotes the kind of norm. An approximation in the L_2 norm means that $p = 2$. With the advent of the high-speed digital computers, a norm used quite frequently is the L_∞ norm or the Tchebycheff norm. Approximation in this norm seeks to minimize

$$\lambda^{\max}\epsilon[a,b]\, f(\lambda) - \Phi(\alpha,\lambda)| \qquad (13)$$

The above expression means the maximum value of the absolute magnitude of the quantity in brackets in the interval $[a,b]$. We have used this norm to solve the above problem but owing to our use of a slowly converging computational process, we did not obtain "better" results than the L_2 norm. Our use of the L_2 norm does not imply that an investigator might not find an approximation in a different norm (specifically the L_1 and L_∞ norms) more suitable to his purposes. We have used the L_2 norm because it is the most convenient from a computational standpoint.

The question of which norm to use is a rather involved question which one cannot go deeply into here. In general, it depends on the situation. Our particular situation is best clarified by an illustration. From the figures of Greenfield *et al.*, we note the kind of fit obtained for myoglobin and lysozyme. This is the type of curve one would get for an approximation in the L_2 norm. For an approximation in the L_∞ norm, the real curve and the approximating curve are shown in Figure 1. In this figure we note that the approximating curve alternates with the real curve and the maximum deviation (d) will occur one more time than there are parameters.

In the final analysis it really becomes a matter for the investigator what norm best suits his purpose, *i.e.*, whether he wants his curve fitted the way Greenfield *et al.* have theirs fitted (L_2 norm), or whether he wants his approximating curve to alternate with his real curve

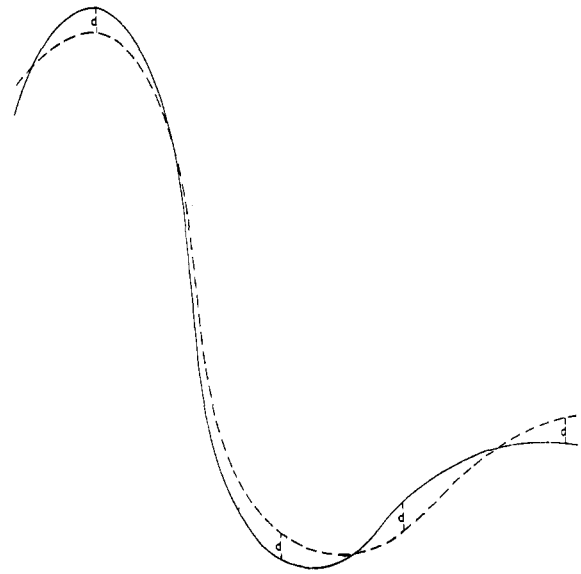

FIGURE 1: Idealized diagram of an approximation in the L_∞ norm.

(L_∞ norm), or whether he wants to minimize the area between the two curves in which case he will want an approximation in the L_1 norm (this is particularly difficult to do analytically but may be done by trial and error in complete analogy to Greenfield *et al.* (1967) except that instead of minimizing the variance we minimize the area).

Our analysis of the data of myoglobin and lysozyme was done without the use of weight functions. Weight functions are inserted to give more emphasis to data in one particular region over another. For instance, it is clear that data in the 220–250-mμ range are more accurate than data in the 190–220-mμ range. When considering weight functions the quantity that is minimized is

$$\int_a^b f(\lambda) - \Phi(\alpha,\lambda)^p w(\lambda) d\lambda]^{1/p} \qquad (14)$$

where $w(\lambda)$ is a weight function. If the standard deviation of the measurements at each wavelength is known then there will be no difficulty in constructing a weight function. The use of weight functions will greatly enhance the accuracy of the analysis and will overcome the type of difficulty encountered by Greenfield *et al.* (1967) in fitting curves such as curve 2 of their eighth figure.

In conclusion we note that this method is not restricted to the analysis of optical rotatory dispersion curves but can be used to determine the quantity of each component of a mixture whose qualitative composition is known. Needless to say that in each instance of application the spectrum of each component must be well documented.

Acknowledgment

The author gratefully acknowledges a number of helpful discussions with Professor B. H. Zimm.

Appendix

To carry out the differentiation indicated in eq 4 we expand the quantity in square brackets. Proceeding thus eq 4 becomes

$$\frac{\partial\left\{\int_a^b [f^2(\lambda) - 2f(\lambda)\Phi(\alpha,\lambda) + \Phi^2(\alpha,\lambda)]d\lambda\right\}}{\partial\alpha_i} = 0$$

$$i = 1, 2\cdots n$$

Since the integrals in the above expression really denote finite sums it is legitimate to differentiate before integrating. For any i differentiating the first term in the square brackets we obtain

$$\frac{\partial\left(\int_a^b f^2(\lambda)d\lambda\right)}{\partial\alpha_i} = 0$$

The second term yields

$$\frac{\partial\left(\int_a^b 2f(\lambda)\Phi(\alpha,\lambda)d\lambda\right)}{\partial\alpha_i} = 2\int_a^b f(\lambda)f_i(\lambda)d\lambda$$

and the last term yields

$$\frac{\partial\left(\int_a^b \Phi^2(\alpha,\lambda)d\lambda\right)}{\partial\alpha_i} = 2\sum_{j=1}^{n}\alpha_j\int_a^b f_i(\lambda)f_j(\lambda)d\lambda$$

Making use of these expressions we obtain eq 5 and 6.

References

Beychok, S. (1965), *Proc. Nat. Acad. Sci. U. S. 53*, 999.

Carver, J. P., Schechter, E., and Blout, E. R. (1966), *J. Am. Chem. Soc. 88*, 2550.

Davidson, B., Tooney, N., and Fasman, G. D. (1966), *Biochem. Biophys. Res. Commun. 23*, 156.

Fasman, G. D., Bodenheimer, E., and Lindblow, C. (1964), *Biochemistry 3*, 1665.

Fasman, G. C., Landsberg, M., and Buchwald, M. (1965), *Can. J. Chem. 43*, 1588.

Greenfield, N., Davidson, B., and Fasman, G. D. (1967), *Biochemistry 6*, 1630.

Harrison, S. C., and Blout, E. R. (1965), *J. Biol. Chem. 240*, 299.

Iizuka, E., and Yang, J. T. (1966), *Proc. Nat. Acad. Sci. U. S. 55*, 1175.

Rice, J. R. (1964), The Approximation of Functions, Reading, Mass., Addison-Wesley.

Sarkar, P. K., and Doty, P. (1966), *Proc. Nat. Acad. Sci. U. S. 55*, 981.

Tomimatsu, Y., and Gaffield, W. (1965), *Biopolymers 3*, 509.