ESTIMATION OF PARAMETERS OF ROTATORY DISPERSION CURVES OF PROTEINS

DON RIDGEWAY

From the Gates and Crellin Laboratories of Chemistry and the Norman W. Church Laboratory of Chemical Biology, California Institute of Technology, Pasadena. Dr. Ridgeway's present address is Department of Biophysics and Biometry, Medical College of Virginia, Richmond

ABSTRACT Estimation of the constants a, b, and λ_0 by means of the standard Moffitt-Yang plot is evaluated. It is found that the method is very insensitive as an estimation procedure and that large errors in b may be expected. Expressions for the maximum-likelihood estimates of the constants are derived.

1. INTRODUCTION

Analysis of the optical rotatory dispersion curves of a great many proteins and synthetic polypeptides has shown that in general such curves may be fitted by a relation of the form:

$$[m'] = a\left(\frac{\lambda_0^2}{\lambda^2 - \lambda_0^2}\right) + b\left(\frac{\lambda_0^2}{\lambda^2 - \lambda_0^2}\right)^2 \tag{1}$$

Here $[m'] = [3M_R/(n^2 + 2)][\alpha]/100$, n being the solution refractive index, M_R the mean molecular weight of the amino acid residues comprising the protein, and $[\alpha]$ the specific optical rotation; λ is the wavelength of the light; and a, b, and λ_0 are constants. Equation (1) was first fitted to dispersion curves of polypeptides by Moffitt and Yang (1). The values of a and b are believed to be related to the proportion of the length of the protein molecule which is in the configuration of the α-helix, and λ_0 is a weighted average of the wavelengths of the electronic transitions contributing to the optical rotation.

Although a and b have received most of the attention of protein chemists, it is evidently necessary, in determining their values, to estimate λ_0 simultaneously. The method of evaluation of rotatory dispersion data suggested by Moffitt and Yang (1) is now usually employed. One plots $[m'](\lambda^2 - \lambda_0^2)$ against $1/(\lambda^2 - \lambda_0^2)$ for a number of trial values of λ_0 and selects that λ_0 which provides the best fit to a straight line with the data. Then a and b are obtained from the slope and intercept, respectively, of this line. Typical values for λ_0 obtained by this procedure which have been reported are 212 mμ for poly-γ-benzyl-L-glutamate (1), 218 mμ for myosin (2), 208 mμ for poly-γ-benzyl-L-glutamate (3), and 216 mμ for myoglobin (4).

Data are now frequently being described in the literature which would not permit reasonably precise estimation of λ_0 because of the small number of wavelengths observed, and in these cases the value of 212 mμ originally given by Moffitt and Yang is usually being assumed. Downie *et al.* (5) indicate that their data provide equally good fit to a straight line using a value of λ_0 of either 212 or 200 mμ. In substituting the one value for the other they find a large shift to be produced in b with their data. Moffitt and Yang (1) give $\pm$ 5 mμ for the precision of their λ_0-value, and Cohen and Szent-Gyorgyi (2) give $\pm$ 10 mμ. Although values for a and b for several proteins under a wide variety of conditions have now been published, only one direct study of λ_0 (Ridgeway (6), on bovine serum albumin) has appeared so far, and in this case λ_0 was shown to be pH-dependent. In the present investigation, statistical techniques are presented for estimation of a, b, and λ_0. The expressions for a and b turn out to be simple, whereas that for λ_0 requires considerable computation. Since λ_0 itself has not until recently been of particular interest to many protein chemists in any case, an investigation is made in the first section to demonstrate that the additional computational effort is warranted. The two questions involved, the sensitivity of a and b to the selection of λ_0 and the efficiency of graphical methods such as that of Moffitt and Yang, are discussed in turn, the one in terms of general expressions relating a and b to the error in λ_0, the other in terms of an example.

2. DEPENDENCE OF a AND b ON λ_0

We wish to determine the errors in estimates of a and b from rotatory dispersion data associated with an error in λ_0. We shall study the case in which it is assumed that the data themselves are free of error. In this case, the problem becomes one of approximation of one curve of the form of equation (1) by another of the same form with a different value of λ_0.

Assume the correct functional dependence of $[m']$ on λ_0 to be

$$[m'] = \alpha\left[\frac{(\lambda_0 + \delta)^2}{\lambda^2 - (\lambda_0 + \delta)^2}\right] + \beta\left[\frac{(\lambda_0 + \delta)^2}{\lambda^2 - (\lambda_0 + \delta)^2}\right]^2 \qquad (2)$$

It is desired to calculate those values of a and b, in terms of δ and of parameters α, β, and the extreme wavelengths at which measurements are made, which will provide the best approximation to equation (2) with equation (1). Let

$$Q_1 = \int_{z(\lambda_1)}^{z(\lambda_2)} [(az + bz^2) - (\alpha\zeta + \beta\zeta^2)]^2\, dz \qquad (3)$$

where λ_1 and λ_2 are the extreme wavelengths ($\lambda_1 > \lambda_2$) and where $z = \lambda_0^2/(\lambda^2 = \lambda_0^2)$ and $\zeta = (\lambda_0 + \delta)^2/[\lambda^2 - (\lambda_0 + \delta)^2]$. We can expand ζ as a power series in z as follows:

$$\zeta = \frac{(\lambda_0 + \delta)^2}{\lambda^2 - (\lambda_0 + \delta)^2} = z\frac{(1 + \epsilon)^2}{1 - z(2\epsilon + \epsilon^2)}$$

$$= (1 + \epsilon)^2 \sum_{n=1}^{\infty} (2\epsilon + \epsilon^2)^{n-1} z^n \tag{4}$$

where $\epsilon = \delta/\lambda_0$

For convenience, we shall select as a criterion of best approximation the conditions that a and b be those values which minimize Q_1 (*i.e.* the least-squares condition):

$$\frac{\partial Q_1}{\partial a} = \frac{\partial Q_1}{\partial b} = 0 \tag{5}$$

Substituting equation (4) into (3), differentiating and equating to zero according to (5), and solving the simultaneous equations in the usual fashion we obtain:

$$a = \alpha\left[1 + 2\epsilon + \left(1 - \frac{8}{5}\Delta^2\right)\epsilon^2 - \left(\frac{24}{5}\Delta^2 + \frac{32}{7}\Delta^3\right)\epsilon^3\right]$$

$$- \beta\left[\frac{8}{5}\Delta^2\epsilon + \left(\frac{36}{5}\Delta^2 + \frac{48}{7}\Delta^3\right)\epsilon^2 + \left(\frac{64}{5}\Delta^2 + \frac{240}{7}\Delta^3 + \frac{144}{7}\Delta^4\right)\epsilon^3\right],$$

$$b = \beta\left[1 + \left(4 + \frac{16}{3}\Delta\right)\epsilon + \left(6 + 24\Delta + \frac{120}{7}\Delta^2\right)\epsilon^2 + \left(4 + \frac{128}{3}\Delta\right.\right.$$

$$\left.\left. + \frac{600}{7}\Delta^2 + \frac{320}{7}\Delta^3\right)\epsilon^3\right] + \alpha\left[2\epsilon + \left(5 + \frac{16}{3}\Delta\right)\epsilon^2 + \left(4 + 16\Delta + \frac{80}{7}\Delta^2\right)\epsilon^3\right], \tag{6}$$

where $\Delta^n = [z^n(\lambda_2) - z^n(\lambda_1)]$. In equations (6), terms containing powers of ϵ higher than 3 have been neglected.[1]

For reasons discussed in the following section, the variable $[m']$ is preferable to $\lambda_0^2[m']/z$ from a statistical standpoint. However, since almost all data in the literature are based on the Moffitt-Yang procedure, in which the variable $\lambda_0^2[m']/z$ appears, it is worthwhile to present the calculation for this variable as well. We define

$$Q_2 = \int_{z(\lambda_1)/\lambda_0^2}^{z(\lambda_2)/\lambda_0^2} \left[\lambda_0^2(a + bz) - \frac{\lambda_0^2}{z}(\alpha\zeta + \beta\zeta^2)\right]^2 d\left(\frac{z}{\lambda_0^2}\right)$$

$$= \frac{1}{\lambda_0^2}\int_{z(\lambda_1)}^{z(\lambda_2)} \left[(a + bz) - \frac{1}{z}(\alpha\zeta + \beta\zeta^2)\right]^2 dz$$

Proceeding exactly as before, we obtain

$$a = \alpha\left[1 + 2\epsilon + \left(1 - \frac{2}{3}\Delta^2\right)\epsilon^2 - \left(2\Delta^2 - \frac{8}{5}\Delta^3\right)\epsilon^3\right]$$

$$- \beta\left[\frac{2}{3}\Delta^2\epsilon + \left(3\Delta^2 + \frac{12}{5}\Delta^3\right)\epsilon^2 - \left(\frac{16}{3}\Delta^2 + 12\Delta^3 + \frac{32}{5}\Delta^4\right)\epsilon^3\right]$$

[1] It is found from sample calculations that this approximation is sufficient if the extreme short wavelength is greater than about 300 mμ and if the error in λ_0 is less than about 5 per cent.

$$b = \beta\left[1 + (4 + 4\,\Delta)\epsilon + \left(6 + 18\,\Delta + \frac{54}{5}\,\Delta^2\right)\epsilon^2 + \left(4 + 32\,\Delta + 54\,\Delta^2 + \frac{128}{5}\,\Delta^3\right)\epsilon^3\right] + \alpha\left[2\epsilon + (5 + 4\,\Delta)\epsilon^2 + \left(4 + 12\,\Delta + \frac{36}{5}\,\Delta^2\right)\epsilon^3\right].$$

The values calculated for a and b from equations (6) for errors of *ca.* 1, 2, and 5 per cent in λ_0 are presented in Table I. In these computations, α and β have both

TABLE I

LEAST-SQUARES VALUES OF a AND b CORRESPONDING TO INCORRECT λ_0

λ_0	$-a$	$-b$
201	306.5	502.4
208	302.7	365.8
210	301.6	330.7
212*	300.0	300.0
214	299.2	270.8
216	297.4	248.3
223	292.9	162.9

* In the curve being approximated, λ_0 = 212

been taken to equal −300, which is in the range commonly encountered in proteins, and the actual value of the dispersion constant, *i.e.* ($\lambda_0 + \delta$), has been taken to be 212 mμ. The extreme wavelengths were taken at 300 and 600 mμ, which is about the region being studied in most laboratories (the values of a and b are in any case rather insensitive to the extreme long wavelength in this calculation). In Table II, the values for $[m']$ at various wavelengths are given which are computed by substitution of the values of a and b corresponding to various values of λ_0 from Table I into equation (1).

In equation (6), a is shown to contain a correction term (to the equality $a = \alpha$) dependent on β of the same magnitude as that dependent on α, and the same is true

TABLE II

COMPUTED VALUES OF OPTICAL ROTATION USING INCORRECT VALUES OF a AND b FROM TABLE I

λ	λ_0 = 201	208	210	212*	214	216	223
300	583.0	592.8	594.9	597.8	600.6	607.6	610.1
350	272.5	274.0	274.2	274.6	275.1	276.8	276.2
400	160.9	162.3	162.7	163.0	163.5	164.5	165.2
450	107.6	109.2	109.6	110.0	110.5	111.2	112.6
500	77.7	79.3	79.7	80.1	80.7	81.3	82.8
550	59.2	60.7	61.1	61.5	62.0	62.4	63.9
600	46.8	48.2	48.5	48.9	49.4	49.8	51.1

* In the curve being approximated, λ_0 = 212

of the correction terms in b. Whereas they are of opposite sign and tend to cancel in a, they are additive in the case of b. Consequently, one would expect Table I to show, as it does, that although a is rather insensitive to the value of λ_0, large variations in b are produced even by small errors in λ_0. In the example, the per cent error in b is approximately ten times that in λ_0. Moreover, it is evident from Table II that remarkably good agreement with data may be obtained in spite of large errors in a and b; and conversely, that graphical methods for estimation of λ_0 require highly precise data if subsequent estimation of b is desired. In stating this conclusion, one should be aware of its dependence on the value of the extreme short wavelength employed in computation. Whereas the graphical determination becomes somewhat worse if the shortest wavelength is greater than 300 mμ, it is markedly improved if data extend down as low as 250 mμ.

3. ESTIMATION OF a, b, AND λ_0

We shall now derive expressions for the maximum-likelihood estimates of a, b, and λ_0 in equation (1). It can be demonstrated (7) that the estimates obtained by the principle of least squares for a series of paired observations are the maximum-likelihood estimates if the independent variable is error-free and if the errors of the dependent variable are normally distributed and of constant variance over the range of the independent variable. In polarimetric dispersion data, one can take the wavelength determination to be free of error and can probably assume that the errors in the measured rotation (or in $[m']$) are normally distributed. However, since the variance of $[m']$ is usually wavelength-dependent, being commonly greater at the extreme wavelengths than in the center of the wavelength band, $[m']$ cannot itself be used in the least-squares method. The variable $[m_i']/s_i$, where s_i is the standard deviation of $[m_i']$, evidently fulfills the remaining condition (the subscript i referring to measurements at λ_i). It has the objection that s_i is rarely determined experimentally. However, since it is reasonable to suppose that the ratios of the variances at the various wavelengths of the dispersion curve are a property of the polarimeter itself rather than of the sample, particularly in the case of symmetrical-angle or null-detection photoelectric instruments, no serious error will be introduced if one applies a single set of variance data once determined for the polarimeter in all calculations. For these reasons, the variable $[m_i']/s_i$ is employed in the following treatment.

It is appropriate to mention the Moffitt-Yang procedure in this connection. In their case, the variable is $[m_i'](\lambda_i^2 - \lambda_0^2)$. $(\lambda_i^2 - \lambda_0^2)$ varies from about 5 at short wavelengths to 30 at long wavelengths. Hence, in applying the least-squares principle to this variable, one effectively weights the squares of deviations from long wavelengths up to 36 times as heavily as those from short wavelengths and thereby magnifies errors in the longer wavelengths by this amount in their effect on the estimates of a and b.

We could now proceed in the standard fashion to minimize the sum of squares of deviations of observed data from equation (1) and obtain the desired expressions for the estimates of a, b, and λ_0. However, because of the form in which λ_0 appears in the sum of squares, it would be very difficult to carry out this calculation. We can take advantage of our foreknowledge about λ_0, *i.e.*, that it is likely to be within a few per cent of 212 and thereby simplify the problem considerably. We shall replace λ_0 with $(\lambda_0' + \delta)$ and derive the least-squares expressions for a, b, and ϵ $(= \delta/\lambda_0')$, using approximations which are valid if ϵ is small. The final equations will then be of such a form that, in applying them, one will substitute $\lambda_0' = 212$ and solve for ϵ. If ϵ turns out to be larger than expected in a particular case, the procedure may be iterated (using the new estimate of λ_0' instead of 212) until the error becomes sufficiently small.

Let

$$Q = \sum_i \frac{1}{s_i^2} ([m_i'] - a\zeta_i' - b\zeta_i'^2)^2 \tag{7}$$

where $\zeta_i' = (\lambda_0' + \delta)^2/[\lambda_i^2 - (\lambda_0' + \delta)^2]$. Here and elsewhere the sum extends over all observed wavelengths λ_i. The least-squares criteria for a and b are, as before, $\partial Q/\partial a = \partial Q/\partial b = 0$. Differentiating equation (7), solving simultaneously for a and b, and introducing the expansions

$$\sum \zeta_i'^j = \sum z_i'^j + 2j\epsilon \sum z_i'^j(1 + z_i') \tag{8}$$

(where $z_i' = \lambda_0'^2/(\lambda^2 - \lambda_0'^2)$), one obtains

$$\begin{aligned} a &= a'd' + \epsilon(a'd'' + a''d') \\ &= a_1 + \epsilon a_2, \quad \text{say}, \\ b &= b'd' + \epsilon(b'd'' + b''d') \\ &= b_1 + \epsilon b_2. \end{aligned} \tag{9}$$

with

$$a' = \sum \frac{[m_i']z_i'}{s_i^2} \sum \frac{z_i'^4}{s_i^2} - \sum \frac{[m_i']z_i'^2}{s_i^2} \sum \frac{z_i'^3}{s_i^2}$$

$$a'' = \sum \frac{[m_i']z_i'}{s_i^2}\left(10 \sum \frac{z_i'^4}{s_i^2} + 8 \sum \frac{z_i'^5}{s_i^2}\right) - \sum \frac{[m_i']z_i'^2}{s_i^2}\left(10 \sum \frac{z_i'^3}{s_i^2} + 4 \sum \frac{z_i'^4}{s_i^2}\right) - 4 \sum \frac{[m_i']z_i'^3}{s_i^2} \sum \frac{z_i'^3}{s_i^2}$$

$$b' = \sum \frac{[m_i']z_i'}{s_i^2} \sum \frac{z_i'^2}{s_i^2} - \sum \frac{[m_i']z_i'}{s_i^2} \sum \frac{z_i'^3}{s_i^2}$$

$$b'' = -\sum \frac{[m_i']z_i'}{s_i^2}\left(8 \sum \frac{z_i'^3}{s_i^2} + 6 \sum \frac{z_i'^4}{s_i^2}\right) + \sum \frac{[m_i']z_i'^2}{s_i^2}\left(8 \sum \frac{z_i'^2}{s_i^2} + 2 \sum \frac{z_i'^3}{s_i^2}\right) + 4 \sum \frac{[m_i']z_i'^3}{s_i^2} \sum \frac{z_i'^2}{s_i^2} \tag{9a}$$

$$d' = \frac{1}{\sum \frac{z_i'^2}{s_i^2} \sum \frac{z_i'^4}{s_i^2} - \left(\sum \frac{z_i'^3}{s_i^2} \right)^2}$$

$$d'' = \frac{\sum \frac{z_i'^2}{s_i^2} \left(3 \sum \frac{z_i'^4}{s_i^2} + 2 \sum \frac{z_i'^5}{s_i^2} \right) - \sum \frac{z_i'^2}{s_i^2} \left(3 \sum \frac{z_i'^2}{s_i^2} + 2 \sum \frac{z_i'^4}{s_i^2} \right)}{\sum \frac{z_i'^2}{s_i^2} \sum \frac{z_i'^4}{s_i^2} - \left(\sum \frac{z_i'^3}{s_i^2} \right)^2}$$

Setting $\partial Q / \partial \epsilon = 0$, substituting for a and b from equations (9), and solving for $\delta \; (= \lambda_0' \epsilon)$, one finds

$$\delta = \lambda_0' \frac{a_1 \sum \frac{[m_i']\lambda_i^2 z_i'^2}{s_i^2} + 2b_1 \sum \frac{[m_i']\lambda_i^2 z_i'^3}{s_i^2} - a_1^2 \sum \frac{\lambda_i^2 z_i'^4}{s_i^2} - 2b_1^2 \sum \frac{\lambda_i^2 z_i'^5}{s_i^2}}{\begin{array}{l} \Big\{ -a_2 \sum \frac{[m_i']\lambda_i^2 z_i'^2}{s_i^2} - 2b_2 \sum \frac{[m_i']\lambda_i^2 z_i'^3}{s_i^2} + 2a_1 a_2 \sum \frac{\lambda_i^2 z_i'^3}{s_i^2} \\ + 3(a_1 b_2 + a_2 b_1) \sum \frac{\lambda_i^2 z_i'^4}{s_i^2} + 4b_1 b_2 \sum \frac{\lambda_i^2 z_i'^5}{s_i^2} \\ - 4a_1 \sum \frac{[m_i']\lambda_i^2}{s_i^2} (z_i'^2 + z_i'^3) - 12b_1 \sum \frac{[m_i']\lambda_i^2}{s_i^2} (z_i'^3 + z_i'^4) \\ + 6a_1^2 \sum \frac{\lambda_i^2}{s_i^2} (z_i'^3 + z_i'^4) + 24a_1 b_1 \sum \frac{\lambda_i^2}{s_i^2} (z_i'^4 + z_i'^5) \\ \qquad + 20b_1^2 \sum \frac{\lambda_i^2}{s_i^2} (z_i'^5 + z_i'^6) \Big\}. \end{array}} \tag{10}$$

Equations (9) and (10) are the desired expressions for the maximum-likelihood estimates of a, b, and δ.

One should be aware, in applying the above equations that iterative procedures of this type are commonly observed to oscillate in their approach to the correct answer. If this should be the case in a particular set of data, one can either turn to a damping procedure, in which some fraction of δ, instead of δ itself, is added to the preceding λ_0' for the next iteration, or to an alternate procedure in which the minimization with respect to λ_0 is carried out graphically. Values of a and b are obtained from equations (9) for each of a number of trial values of λ_0. The values of Q obtained by substituting each pair into equation (7) are then plotted against the corresponding values of λ_0 and the minimum is determined by inspection. Whereas the iterative procedure is preferable in work with high speed computers, the alternate procedure is probably more convenient with smaller computers.

The writer is grateful to Dr. Linus Pauling for discussions with him on this subject.

This paper is Contribution No. 2868 from the Gates and Crellin Laboratories of Chemistry. The work was presented in part at the 6th annual meeting of the Biophysical Society in Washington, D. C., February 15, 1962.

Dr. Ridgeway was a United States Public Health Service Postdoctoral Fellow and Research Fellow in Chemistry of the California Institute of Technology.

Received for publication, June 17, 1962.

REFERENCES

1. MOFFITT, W., and YANG, J. T., *Proc. Nat. Acad. Sc.*, 1956, **42**, 596.
2. COHEN, C., and SZENT-GYORGYI, A., *J. Am. Chem. Soc.*, 1957, **79**, 248.
3. Doty, P., *in* Proceedings of the IVth International Congress of Biochemistry (Vienna), London, Pergamon Press, 1959, **8**, 8.
4. URNES, P. J., IMAHORI, K., and DOTY, P., *Proc. Nat. Acad. Sc.*, 1961, **47**, 1635.
5. DOWNIE, A. R., ELLIOTT, A., HANBY, W. E., and MALCOLM, B. R., *Proc. Roy. Soc., London, Series A*, 1957, **242**, 325.
6. RIDGEWAY, D., *Proc. Nat. Acad. Sc.*, 1962, **48**, 1482.
7. See *e.g.* BENNETT, C. A., and FRANKLIN, N. L., Statistical Analysis in Chemistry and the Chemical Industry, New York, John Wiley and Sons, 1954, 214f.