Bolton, A. E., & Hunter, W. M. (1973) *Biochem. J. 133*, 529-539.
Bowie, E. J. W., Thompson, J. H., Jr., Didisheim, P., & Owen, C.A., Jr. (1971) *Mayo Clinic Laboratory Manual of Hemostasis*, W. B. Saunders, Philadelphia, PA.
Canfield, W., Nesheim, M. E., Kisiel, W., & Mann, K. G. (1978) *Circulation 58*, 4.
Chulkova, T., & Hernandez, G. (1975) *Clin. Chim. Acta 62*, 21-28.
Colman, R. W. (1969) *Biochemistry 8*, 1438-1445.
Day, W. C. (1975) *Biochim. Biophys. Acta 386*, 352-361.
Dombrose, F. A., & Seegers, W. H. (1974) *Thromb. Diath. Haemorrh., Suppl. 57*, 241-254.
Edelhoch, H., & deChrombrugghe, B. (1966) *J. Biol. Chem. 241*, 4357-4365.
Esmon, C. G. (1979) *J. Biol. Chem. 254*, 964-973.
Esnouf, M. P., & Jobin, F. (1967) *Biochem. J. 102*, 660-665.
Grabar, P., & Williams, C. A. (1953) *Biochim. Biophys. Acta 10*, 193-194.
Hibbard, L. S., & Mann, K. G. (1980) *J. Biol. Chem. 255*, 638-645.
Ittyerah, T. R., Rawala, R., & Colman, R. W. (1980) in *The Regulation of Coagulation* (Mann, K. G., & Taylor, F. B., Jr., Eds.) pp 161-172, Elsevier/North-Holland Publishing Co., Amsterdam.
Kawahara, K., & Tanford, C. (1966) *Biochemistry 5*, 1578-1584.
Mann, K. G., & Fish, W. W. (1972) *Methods Enzymol. 26*, 28-42.
Mann, K. G., Fass, D. N., & Fish, W. W. (1973) *Adv. Chem. Ser., No. 125*, 310-326.
Murphy, R. C., & Seegers, W. H. (1948) *Am. J. Physiol. 154*, 134-139.
Nesheim, M. E., & Mann, K. G. (1979) *J. Biol. Chem. 254*, 1326-1334.
Nesheim, M. E., Myrmel, K. H., Hibbard, L. S., & Mann, K. G. (1978) *Fed. Proc., Fed. Am. Soc. Exp. Biol. 37* (6), 1750.
Nesheim, M. E., Taswell, J. B., & Mann, K. G. (1979a) *J. Biol. Chem. 254*, 10952-10983.
Nesheim, M. E., Myrmel, K. H., Hibbard, L., & Mann, K. G. (1979b) *J. Biol. Chem. 254*, 508-517.
Nesheim, M. E., Prendergast, F. G., & Mann, K. G. (1979c) *Biochemistry 18*, 996-1003.
Owren, P. A. (1947a) *Lancet No. 1*, 446-448.
Owren, P. A. (1947b) *Acta Med. Scand., Suppl. No. 194*, 1-316.
Saraswathi, S., Rawala, R., & Colman, R. W. (1978) *J. Biol. Chem. 253*, 1024-1029.
Smith, C. M., & Hanahan, D. J. (1976) *Biochemistry 15*, 1830-1838.
Sumner, J. B., Gralen, N., & Eriksson-Quensel, I. (1938) *J. Biol. Chem. 125*, 37-44.
Tanford, C. (1960) *Physical Chemistry of Macromolecules*, p 365, Wiley, New York.
Tanford, C., & Lovrien, R. (1962) *J. Am. Chem. Soc. 84*, 1892-1896.
Tracy, P. B., Peterson, J., Nesheim, M. E., McDuffie, F. C., & Mann, K. G. (1979) *J. Biol. Chem. 254*, 10354-10361.
Ware, A. G., & Seegers, W. H. (1948) *Am. J. Physiol. 152*, 699-705.
Yphantis, D. A. (1964) *Biochemistry 3*, 297-317.

Estimation of Globular Protein Secondary Structure from Circular Dichroism†

Stephen W. Provencher* and Jürgen Glöckner

ABSTRACT: A new method is developed in which a circular dichroism (CD) spectrum is analyzed directly as a linear combination of the CD spectra (from 190 to 240 nm) of 16 proteins whose secondary structures are known from X-ray crystallography. This avoids the dilemma encountered in previous methods of trying to define single reference CD spectra that were supposed to characterize such broad and variable classes as helix, β sheet, β turn, and "remainder". It also permits a more accurate and flexible analysis. The usual instability in using so many parameters is automatically controlled by a simple constrained statistical regularization procedure (similar to ridge regression). Sixteen tests were made by removing 1 spectrum at a time from the set of 16 and analyzing it in terms of the other 15. The product moment correlation coefficients between the computed fractions of helix, β sheet, β turn, and remainder and the fractions from the X-ray data were 0.96, 0.94, 0.31, and 0.49, respectively. Thus, the helix and β-sheet accuracy is very good. (The corresponding values calculated by a previous method with four reference spectra were 0.85, 0.25, −0.31, and 0.46.)

CD[1] is a convenient and widely used method for studying the conformations and conformational changes of globular proteins in solution. The usual procedure for estimating secondary structure composition is to approximate $y(\lambda)$, the mean residue ellipticity of the protein at wavelength λ, simply by a linear superposition of a small set of N_f reference spectra, $r_i(\lambda)$, each of which is supposed to be characteristic of a particular conformational class:

$$y(\lambda) = \sum_{i=1}^{N_f} f_i r_i(\lambda) \quad (1)$$

where f_i is the fraction of residues in class i, and the $r_i(\lambda)$ values are previously determined from CD spectra of model polypeptides (Greenfield et al., 1967) or globular proteins (Saxena & Wetlaufer, 1971; Chen et al., 1972) whose secondary

† From the European Molecular Biology Laboratory, Postfach 10.2209, D-6900 Heidelberg, Federal Republic of Germany. *Received June 3, 1980.*

[1] Abbreviation used: CD, circular dichroism.

0006-2960/81/0420-0033$01.00/0

structures are known. The classes helix, β sheet, and remainder (where remainder means everything not belonging to the other classes) have been most often used. Improved accuracy has been obtained by introducing an empirical chain length dependent factor for the helix spectrum (Chen et al., 1974) by adding one (Chang et al., 1978) or three (Brahms & Brahms, 1979, 1980) β-turn classes and by splitting β sheet into parallel and antiparallel classes (Bolotina et al., 1979).

It is well-known, however, that no single reference spectrum can accurately represent all members of any of these broad and somewhat vaguely defined classes; conformations in globular proteins deviate significantly from the ideal model forms. In addition, the mean residue ellipticity depends on chain length and number of strands in the β sheet, and the contributions of nonpeptide chromophores have been neglected. It is therefore not surprising that the estimates of the f values are sometimes very unreliable [see Chang et al. (1978) for a detailed discussion of the above problems]. On the other hand, it is clear that attempting to take all these factors into account would involve so many parameters that the usual least-squares estimates of the f values would be hopelessly unstable to experimental error. Defining all the reference spectra would also be problematic.

We thus have the dilemma, common in experimental studies of complex biochemical systems, of choosing between an inadequate (but stable) model and an unstable (but more realistic) one. We get around this by using a simple constrained regularization procedure that yields solutions stable to experimental error even when there are many parameters. This allows us to analyze a CD spectrum directly as a linear combination of 16 CD spectra of proteins of known secondary structure. We thus completely avoid the problem of defining reference spectra and have a more flexible and accurate analysis in terms of actual protein spectra. We test this method and compare it with a method using reference spectra.

Methods

We analyze the CD spectrum of a protein as a sum of N_γ CD spectra, $R_j(\lambda)$, of proteins whose structures are known from X-ray crystallography:

$$y(\lambda) = \sum_{j=1}^{N_\gamma} \gamma_j R_j(\lambda) \tag{2}$$

If we can determine the γ values, then we immediately have the f values:

$$f_i = \sum_{j=1}^{N_\gamma} \gamma_j F_{ji} \qquad i = 1, \ldots, N_f \tag{3}$$

where F_{ji} is the fraction of residues of protein j in conformational class i. The F values are obtained from the X-ray structures, and we are therefore assuming that the $R(\lambda)$ values [but not necessarily $y(\lambda)$] were measured under conditions where the solution and crystallographic secondary structures were the same.

The γ values are determined from the N_y measured mean residue ellipticities, $y_{obsd}(\lambda_k)$, by the criterion

$$\sum_{k=1}^{N_y} [y(\lambda_k) - y_{obsd}(\lambda_k)]^2 + \alpha \sum_{j=1}^{N_\gamma} \left(\gamma_j - \frac{1}{N_\gamma} \right)^2 = \text{minimum} \tag{4}$$

with the obvious constraints

$$\sum_{i=1}^{N_f} f_i = 1 \tag{5}$$

$$f_i \geq 0 \qquad i = 1, \ldots, N_f \tag{6}$$

For a given α, eq 2–6 are solved by using a quadratic programming procedure (Lawson & Hanson, 1974; Provencher, 1979). When $\alpha = 0$, we get the ordinary constrained least-squares solution, which is likely to be poor in our case where N_γ is too large for the data to accurately determine so many γ values. When $\alpha > 0$, the second term on the left of eq 4, the regularizor, tends to stabilize the solution by keeping each γ_j small (i.e., near $1/N_\gamma$) unless the corresponding $R_j(\gamma)$ happens to have components that fit $y(\lambda)$ well and therefore significantly reduce the first term on the left. In this way, we have a very flexible but stable model that can adapt itself to the data by selecting from the large number of possible CD components the ones that match the data best. This is discussed more precisely under Appendix.

A general Fortran regularization package has been developed and also applied to the analysis of polydispersity with quasi-elastic light scattering (Provencher et al., 1978) and chemical relaxation spectra (Provencher & Dovi, 1979). A user-oriented version will be available on request. The method is completely automatic and objective in that the only input required is $y(\lambda)$; all other decisions (e.g., the choice of α) are made by the program.

For the $R(\lambda)$ values, we used the CD spectra, available from Chang et al. (1978), of 16 proteins (1–16 in Table I) in aqueous solutions (pH 7.0) in 1-nm intervals from 190 to 240 nm. Subtilisin BPN′ and thermolysin (17 and 18 in Table I) were not used for reasons given in the next section.

Results

We made 16 tests of the method by removing one CD spectrum at a time from the set of 16 $R(\lambda)$ spectra and analyzing it as the $y(\lambda)$ in eq 2 in terms of the other 15 (i.e., with $N_\gamma = 15$ and $N_y = 51$). The resultant f values are shown as method 1 in Table I. Method 2 is that of Chang et al. (1978) with four reference spectra and the constraints in eq 5 and 6. Our results are better than theirs because we did not use spectra 1, 3, 6, 17, and 18 to compute the reference spectra for the β-sheet, β-turn, and remainder classes whereas they did not use spectra 1, 3, and 6. However, method 2 does use the spectrum and the (ordinarily unknown) X-ray F values of the protein being analyzed to compute the reference spectra. Thus (except for proteins 1, 3, 6, 17, and 18), method 2 is not a fair test because it biases the computed f values toward the X-ray values. (If we did this with method 1, we would obtain exact agreement with the X-ray values.) Therefore, we can only compare method 1 with method 2a, which is method 2 except that the spectrum being analyzed is not used to compute the reference spectra.

Table II summarizes the results of the 16 tests. The product moment correlation coefficients [see, e.g., Chang et al. (1978)] can range between 1.0 and −1.0, with 1.0 meaning perfect correlation between the CD and the X-ray f values and 0.0 meaning no correlation. Thus, method 1 yields very good helix and β-sheet accuracy.

The flexibility of method 1 is especially clear in the analysis of the CD spectrum of concanavalin A, which is about 51% β sheet and only 2% helix. Chang et al. (1978), as well as Bolotina et al. (1979), obtained their worst results for this protein (see Table I) and were not able to fit the spectrum (see Figure 1). We obtained good f values and a good fit because α in eq 4 automatically adapted to the apparent difficulty of fitting this spectrum and became especially small (see Appendix).

Many combinations of the following modifications were tried. (1) The empirical corrections of Chang et al. (1978) for the dependence of the mean residue ellipticity on $\bar{n}$, the

Table I: Comparison of Secondary Structure Compositions Estimated from CD with X-ray Values

no.	protein[b]	method[c]	100 × f[a] H	β	t	R	no.	protein	method	100 × f H	β	t	R
1	adenylate kinase	X-ray	54	12	19[d]	15[d]	10	myoglobin	X-ray	79	0	5	16
		1	44	15	20	21			1	86	0	0	14
		2	45	32	5	18			2	80	0	3	17
		2a	45	32	5	18			2a	80	1	2	17
2	carboxypeptidase	X-ray	37	15	26	22	11	nuclease	X-ray	24	15	18	43
		1	43	15	25	16			1	32	25	14	29
		2	40	19	28	13			2	29	24	12	36
		2a	32	44	17	7			2a	29	26	10	34
3	α-chymotrypsin	X-ray	9	34	34	23	12	papain	X-ray	28	14	17	41
		1	9	29	22	40			1	27	5	31	36
		2	6	50	3	42			2	28	0	17	56
		2a	6	50	3	42			2a	24	0	13	63
4	concanavalin A	X-ray	2	51	9	38	13	parvalbumin	X-ray	62	5	17	16
		1	8	41	15	36			1	58	0	0	42
		2	20	67	8	6			2	51	0	27	22
		2a	35	0	39	26			2a	46	0	29	25
5	cytochrome c	X-ray	39	0	24	37	14	ribonuclease A	X-ray	23	40	13	24
		1	33	9	17	41			1	26	44	11	19
		2	42	7	26	25			2	23	36	12	30
		2a	41	15	22	21			2a	25	29	14	32
6	elastase	X-ray	7	52	26	15	15	ribonuclease S	X-ray	26	44	13	17
		1	4	49	14	32			1	25	37	16	23
		2	0	45	8	47			2	25	33	14	28
		2a	0	45	8	47			2a	27	24	18	32
7	insulin	X-ray	51	24	12	13	16	trypsin inhibitor	X-ray	28	33	3	36
		1	49	23	27	0			1	21	28	23	29
		2	45	30	15	10			2	9	23	6	62
		2a	45	27	17	11			2a	14	0	24	62
8	lactate dehydrogenase	X-ray	45	24	6	25	17	subtilisin BPN'	X-ray	31	10	22	37
		1	40	22	13	26			1	15	48	18	18
		2	42	27	9	22			2	15	62	1	22
		2a	43	26	10	22			2a	15	62	1	22
9	lysozyme	X-ray	41	16	23	20	18	thermolysin	X-ray	36	22	18	24
		1	45	21	26	8			1	28	48	21	3
		2	32	30	7	31			2	27	60	1	12
		2a	30	35	6	30			2a	27	60	1	12

[a] Abbreviations used: f, fraction of residues in a given conformational class; H, helix; β, β sheet; t, β turn; R, remainder. [b] See Chang et al. (1978) for sources of the proteins, references, and the determination of the f values from the X-ray results. [c] Methods used: (1) this paper; (2) constrained method of Chang et al. (1978); (2a) method 2 but not using the spectrum being analyzed to determine the reference spectra. See text for details. [d] For adenylate kinase, the correction of Chang et al. (1978) for the net contribution of β turns was not available.

Table II: Correlation Coefficients and Root-Mean-Square Deviations between CD Estimates and X-ray Values of the f Values for 16 Proteins

method	correlation coefficient H	β	t	R	100 × rms[a] H	β	t	R
1	0.96	0.94	0.31	0.49	5	6	10	11
2	0.92	0.83	0.23	0.37	8	11	11	16
2a	0.85	0.25	−0.31	0.46	11	21	15	15

[a] rms, root-mean-square deviations of the f values from X-ray values. See also footnotes to Table I.

mean number of residues per helical segment, were tried. When $\bar{n}$ was made a free parameter, the solution became unstable, with $\bar{n}$ usually becoming unrealistically large or small. When $\bar{n}$ was fixed at 10, the results became worse than with no $\bar{n}$. With the (somewhat artificial) constraint $8 \leq \bar{n} \leq 14$, the improvements were minor. (2) Both functions of Chang et al. (1978) were used as a fixed helix reference spectrum, with slightly worse results. (3) For $\lambda \leq 200$ nm or $\lambda \leq 210$ nm, the data were given lower statistical weights in eq 4 or completely discarded. The results were quite insensitive to this. (4) The X-ray F values from the automatic algorithm of Levitt & Greer (1977) were used for the helix and β sheet. The improvements were small. (5) To test the sensitivity of the

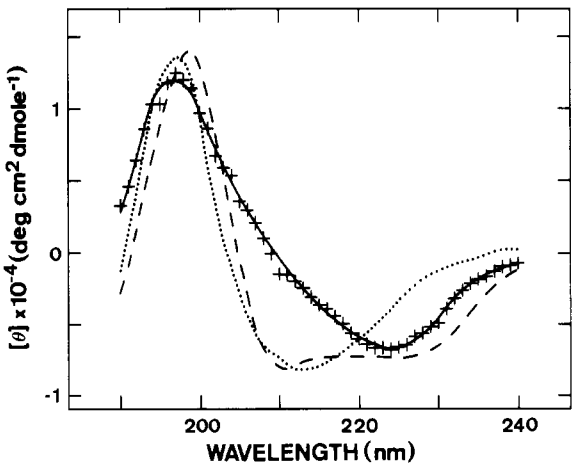

FIGURE 1: Fit to the mean residue ellipticity data (—) on concanavalin A with method 1 of this paper (+) and with reference spectra methods 2 (···) and 2a (– – –). See the text for details.

solution to the biasing effect of the second term in eq 4, $1/N_\gamma$ was replaced with zero; i.e., the γ values were biased toward zero rather than $1/N_\gamma$ (see Appendix). The first two significant figures of the γ and f values did not change. None of these modifications were adopted because any improvements

were small, and the extreme simplicity of method 1 seemed an advantage.

Two modifications were adopted, however. First, two rather extreme outliers were arbitrarily changed to smooth the $R(\lambda)$ values, $R_{10}(202)$ from 4170 to 1170 deg cm^2 dmol^{-1} and $R_{15}(231)$ from −3854 to −2854 deg cm^2 dmol^{-1}. The improvements in the results were minor, but we intend eventually to smooth all the $R(\lambda)$ values anyway, to make interpolation to other wavelengths easier.

The second modification made was that the CD spectra of subtilisin BPN′ and thermolysin were not used as $R(\lambda)$ spectra. Initially, when we used all 18 spectra, the correlation coefficients were only 0.86, 0.07, 0.42, and 0.13 for helix, β sheet, β turn, and remainder, respectively. We then systematically removed nearly 100 combinations of CD spectra from the set of $R(\lambda)$ spectra. Usually, the results became slightly worse. Only the removal of thermolysin and particularly of subtilisin BPN′ led to dramatic improvements. When the β-sheet X-ray F of 0.17 (and hence a remainder F of 0.30), as quoted by Chou & Fasman (1974), was used for subtilisin BPN′, instead of the F values in Table I, the results were only slightly improved. Another possibility is that the CD spectrum itself is wrong; it differs significantly from the subtilisin Novo CD spectra of Brahms & Brahms (1979, 1980), although the two X-ray structures are very similar (Drenth et al., 1972) and have identical amino acid sequences. [Unfortunately, it appears that the wrong figure has been used in one of the papers by Brahms & Brahms (1979, 1980) since their two spectra are quite different.] We do not have a second CD spectrum of thermolysin for comparison. Colman et al. (1972) note that many of the helices in thermolysin are quite distorted, but this is true to a certain extent in nearly all globular proteins, and it is difficult to quantify this effect. In view of the reassuring stability of the method to all other modifications, a clear explanation of the anomalous sensitivity of the results to the CD spectra or F values of these two proteins would be especially desirable.

Discussion

The greatly improved results (see Tables I and II) over previous methods are due to two main advantages, the flexibility of the model with the automatically adjusted α in eq 4 and the freedom from having to define reference spectra characteristic of conformational classes. The latter means that the accuracy of the f value for one class can be quite independent of the (e.g., lower) accuracy of the f values of other classes. We analyze a spectrum directly as a sum of protein spectra in eq 2, independent of any classification system (except that the X-ray F values appear indirectly in the constraints in eq 6, but these are not often active). The classification in terms of the F values is only imposed afterward in eq 3, and here the f value of a class is simply a sum over the F values of the *same* class. This explains why the excellent helix accuracy in Tables I and II is not affected by the errors in the F values for the other forms, which are more difficult to classify unambiguously from X-ray data.

This also means that there is no problem if one class has members with very different CD spectra. We could just as well have combined the F_{f3} (β turn) and F_{f4} (remainder) and analyzed for helix, β sheet, and remainder. Similarly, combining parallel and antiparallel β sheets or the 15 types of β turns causes no difficulty whereas it obviously does with the previous methods that use reference spectra. The only difference is the number of constraints in eq 6. Since these constraints help stabilize the analysis, it is reasonable to use any classes for which the X-ray F values can be unambiguously and accurately assigned, even if they are so weakly represented that they are best combined with other classes after the analysis.

In order for the models in eq 1 or 2 to be adequate (i.e., complete), the $r(\lambda)$ or $R(\lambda)$ spectra would have to contain all the types of CD components that the measured spectrum, $y(\lambda)$, does. The estimation of a large set of reference spectra is difficult and error prone. For example, the β-turn reference spectrum of Chang et al. (1978) bears little resemblance to any of the three of Brahms & Brahms (1980), who had to arbitrarily multiply one of their reference spectra by 0.5 to account for the higher regularity of their synthetic polypeptide compared to globular proteins. It therefore seems natural to avoid the estimation of reference spectra altogether and analyze a protein CD spectrum directly as a sum of other protein CD spectra. It is much easier to find a collection of $R(\lambda)$ spectra of proteins of known secondary structure that is varied enough to make eq 2 nearly complete.

What our method is not immune to are nonlinear effects like the dependence of the mean residue ellipticity on the helix chain length or the β-sheet length or width (Woody, 1969; Madison & Schellman, 1972). The fact that including an empirical chain-length dependence for the helix had little effect probably means that there are other just as serious effects being ignored. Furthermore, the number of linear effects that can be analyzed is determined by the effective rank of the noisy matrix of the $R(\lambda)$ values, the extent to which the constraints in eq 5 and 6 can stabilize the analysis, and the accuracy of the $y(\lambda)$ data. Thus, although all the types of nonpeptide contributions can in principle be combined in the remainder class, the ultimate accuracy attainable with any CD method will be limited by these contributions, as well as by the nonlinear effects and the wide range of distorted conformations in globular proteins. Success will partly hinge on the "averaging out" of these extra effects over the large number of residues so that they are not very different from those in some of the other reference proteins. This may not be the case for proteins (particularly small polypeptides) with very peculiar amino acid compositions and with structures very different from all the reference proteins.

The susceptibility to this last problem should decrease as the number, and especially as the variety, of the reference proteins increases (i.e., as N_γ increases). We observed a gradual but clear trend toward decreased accuracy as N_γ was decreased, and we expect the converse to hold. Variety should be especially important. For example, when ribonuclease A was analyzed, 50% of its $y(\lambda)$ in eq 2 was the $R(\lambda)$ of ribonuclease S, whose $\gamma = 0.84$ was 5 times larger than any of the other γ values; the converse was also true. Thus, a reference protein with a structure similar to the one being analyzed tends to be automatically selected with a large γ. The ability of the method to perform a stable analysis with a large and varied set of $R(\lambda)$ spectra is therefore very important. A possible new $R(\lambda)$ can easily be screened with the same procedure that we used for Tables I and II. If the new $R(\lambda)$ leads to worse accuracy (as thermolysin and subtilisin BPN′ did), it should not be used. The accuracy should also improve by increasing the accuracy of the CD spectra and by extending the spectra down to about 165 nm (Brahms & Brahms, 1979, 1980).

A consistent classification system for computing the F values from the X-ray atomic coordinates is essential. We also obtained good results for the helix and β sheet by using the F values of Levitt & Greer (1977) for those classes, although these F values for β sheet were often much larger than those

of Chang et al. (1978). This means that internal consistency is of prime importance. A version of the algorithm of Levitt & Greer (1977) tuned to the needs of CD could be very useful.

Finally, it should be noted that we had the advantage of another internal consistency: the CD spectra of the protein being analyzed and of the reference proteins came from the same laboratory. Obviously, careful calibration of the instrument and perhaps a test spectrum of one of the reference proteins will be important.

Acknowledgments

We thank Drs. Peter M. Bayley, Thomas A. Reed, and Patrice Vachette for helpful discussions.

Appendix

Justification and Use of Eq 4. Ordinary least squares (eq 4 with $\alpha = 0$) is widely used because the Gauss–Markov theorem (Silvey, 1975) shows that, of all linear *unbiased* methods, it is the best in the sense that it yields the γ values with the minimum expected variance from the true values. Nevertheless, a large N_γ or nearly linearly dependent $R(\lambda)$ spectra in eq 2 can make the least-squares estimates hopelessly unstable to experimental error. In view of the optimal properties of least squares, it is widely believed that the only way of reducing this instability is to reduce N_γ, even at the expense of having a totally inadequate model. However, Hoerl & Kennard (1970) have shown that there always exists a linear estimator (analogous to $\alpha > 0$ in eq 4) that yields γ values with smaller variance than least squares. These estimates are *biased* (i.e., the expectation values of the γ values are not the true values) and therefore are not covered by the Gauss–Markov therorem.

We can get a clearer justification of eq 4 with some additional assumptions that are reasonable for CD data. Generally, we have no a priori knowledge of the composition of $y(\lambda)$ in terms of the $R(\lambda)$ values in eq 2. It is therefore reasonable to assign the same (unknown) a priori probability distribution to each of the γ values. (If we did not, then we would be giving an unjustified preference to some of the γ values.) This is the principle of exchangeability (Lindley & Smith, 1972). We take the mean of this distribution to be $\bar{\gamma} = 1/N_\gamma$ since, from eq 3 and 5, the γ values must sum to 1. If we further assume that the γ values and the $y_{obsd}(\lambda)$ values are normally distributed, then the Bayesian estimator (Lindley & Smith, 1972) for the γ values is just eq 4 with $\alpha = \sigma_y^2/\sigma_\gamma^2$, where σ_λ^2 and σ_y^2 are the variances of the γ and $y_{obsd}(\lambda)$ values, respectively. Even without assuming normality, eq 4 is formally just ordinary least squares with our prior assumptions about the mean and variance of the γ values added as "data".

In practice, we do not know $\sigma_y^2/\sigma_\gamma^2$ a priori, and there have been many procedures proposed for automatically choosing the optimum α during the analysis (Draper & Van Nostrand, 1979). When $\alpha = 0$, the variance of the fit to the data (the first term in eq 4) is minimal, but the γ values are generally unreliable because their instability to the experimental errors is maximal. As α increases, the variance of the fit monotonically increases, and the effective number of degrees of freedom (Provencher, 1979), which is analogous to the number of free parameters in an ordinary least squares, monotonically decreases. Usually, the variance of the fit increases slowly at first and then rapidly as α becomes so large and the degrees of freedom so few that the model is no longer flexible enough and the γ values are too strongly biased toward $\bar{\gamma} = 1/N_\gamma$. The aim is to choose the α that represents the optimum compromise between flexibility and stability, thereby yielding the best γ values. We use a rough hypothesis test on the relative increase of the variance of the fit as α increases (Provencher, 1979) to choose α. The automatic choice of α is very important; α tends to adapt itself to the CD spectrum being analyzed. For example, if the data are especially precise or the spectrum requires more degrees of freedom to be adequately represented by eq 2, then a smaller α is automatically chosen. The method also produces rough confidence regions for the f values (Provencher, 1979).

In this appendix, we have ignored some complications that were not essential to the discussion or conclusions, e.g., the constraints in eq 5 and 6 and the experimental errors in the $R(\lambda)$ values. Also, as required by the Gauss–Markov theorem, a weighting matrix can be introduced into eq 4 to account for correlations or unequal variances in the experimental errors. Similarly, we can incorporate extra a priori information by weighting the individual terms in the second sum in eq 4. For example, if we knew that the protein being analyzed (e.g., ribonuclease A) was structurally similar to some of the reference proteins (e.g., ribonuclease S), then we could give these γ values more freedom by weighting the terms more lightly in the second sum, or, if there were only one such protein, we could perhaps increase its $\bar{\gamma}$. However, this would be usually difficult to do objectively and quantitatively, and we have not done it here. We have found that the γ values of proteins structurally similar to the one being analyzed automatically tend to be large anyway, without any added information.

References

Bolotina, I. A., Chekhov, V. O., & Lugauskas, V. Y. (1979) *Int. J. Quantum Chem. 16*, 819.
Brahms, S., & Brahms, J. G. (1979) *J. Chim. Phys. 76*, 841.
Brahms, S., & Brahms, J. (1980) *J. Mol. Biol. 138*, 149.
Chang, C. T., Wu, C. S. C., & Yang, J. T. (1978) *Anal. Biochem. 91*, 13.
Chen, Y. H., Yang, J. T., & Martinez, H. M. (1972) *Biochemistry 11*, 4120.
Chen, Y. H., Yang, J. T., & Chau, K. H. (1974) *Biochemistry 13*, 3350.
Chou, P. Y., & Fasman, G. D. (1974) *Biochemistry 13*, 222.
Colman, P. M., Jansonius, J. N., & Matthews, B. W. (1972) *J. Mol. Biol. 70*, 701.
Draper, N. R., & Van Nostrand, R. C. (1979) *Technomet. 21*, 451.
Drenth, J., Hol, W. G. J., Jansonius, J. N., & Koekoek, R. (1972) *Eur. J. Biochem. 26*, 177.
Greenfield, N., Davidson, B., & Fasman, G. D. (1967) *Biochemistry 6*, 1630.
Hoerl, A. E., & Kennard, R. W. (1970) *Technomet. 12*, 55.
Lawson, C. L., & Hanson, R. J. (1974) *Solving Least Squares Problems*, Prentice-Hall, Englewood Cliffs, NJ.
Levitt, M., & Greer, J. (1977) *J. Mol. Biol. 114*, 181.
Lindley, D. V., & Smith, A. F. M. (1972) *J. R. Statist. Soc., Ser. B 34*, 1.
Madison, V., & Schellman, J. (1972) *Biopolymers 11*, 1041.
Provencher, S. W. (1979) *Makromol. Chem. 180*, 201.
Provencher, S. W., & Dovi, V. G. (1979) *J. Biochem. Biophys. Methods 1*, 313.
Provencher, S. W., Hendrix, J., De Maeyer, L., & Paulussen, N. (1978) *J. Chem. Phys. 69*, 4273.
Saxena, V. P., & Wetlaufer, D. B. (1971) *Proc. Natl. Acad. Sci. U.S.A. 68*, 969.
Silvey, S. D. (1975) *Statistical Inference*, Chapman and Hall, London.
Woody, R. W. (1969) *Biopolymers 8*, 669.