ANALYTICAL BIOCHEMISTRY 191, 110–118 (1990)

Estimation of Protein Secondary Structure and Error Analysis from Circular Dichroism Spectra

Ivo H. M. van Stokkum,[1] Hans J. W. Spoelder, Michael Bloemendal,*
Rienk van Grondelle, and Frans C. A. Groen
*Faculty of Physics and Astronomy, and *Faculty of Chemistry, Free University, Amsterdam, The Netherlands*

Received June 4, 1990

The estimation of protein secondary structure from circular dichroism spectra is described by a multivariate linear model with noise (Gauss–Markoff model). With this formalism the adequacy of the linear model is investigated, paying special attention to the estimation of the error in the secondary structure estimates. It is shown that the linear model is only adequate for the α-helix class. Since the failure of the linear model is most likely due to nonlinear effects, a locally linearized model is introduced. This model is combined with the selection of the estimate whose fractions of secondary structure summate to approximately one. Comparing the estimation from the CD spectra with the X-ray data (by using the data set of W. C. Johnson Jr., 1988, *Annu. Rev. Biophys. Chem.* 17, 145–166) the root mean square residuals are 0.09 (α-helix), 0.12 (anti-parallel β-sheet), 0.08 (parallel β-sheet), 0.07 (β-turn), and 0.09 (other). These residuals are somewhat larger than the errors estimated from the locally linearized model. In addition to α-helix, in this model the β-turn and "other" class are estimated adequately. But the estimation of the antiparallel and parallel β-sheet class remains unsatisfactory. We compared the linear model and the locally linearized model with two other methods (S. W. Provencher and J. Glöckner, 1981, *Biochemistry* 20, 1085–1094; P. Manavalan and W. C. Johnson Jr., 1988, *Anal. Biochem.* 167, 76–85). The locally linearized model and the Provencher and Glöckner method provided the smallest residuals. However, an advantage of the locally linearized model is the estimation of the error in the secondary structure estimates.

[1] To whom correspondence should be addressed: Faculty of Physics and Astronomy, Free University, De Boelelaan 1081, NL-1081 HV Amsterdam, The Netherlands.

Structural information on proteins is necessary to understand their function. Although high-resolution methods, like X-ray diffraction or NMR spectroscopy, are available, in most cases their application is cumbersome or even impossible. Therefore other techniques have to be used that yield less detailed, but nevertheless important information. For the further development of such methods a quantification of the resulting information and estimates of possible errors are essential.

Proteins show characteristic uv circular dichroism (CD) spectra that are related to the presence of secondary structure (1, 2). In contrast to high-resolution methods the CD spectrum represents an average over the entire protein. There are two ways in which one can investigate the relation between the secondary structure of a protein and its CD spectrum. First one can consider the *forward problem:* given a model for the secondary structure of a protein, what will be its CD spectrum? This leads to complicated quantum mechanical calculations (1, 3–5). Second, one can regard the *inverse problem:* given a measured CD spectrum, how can one estimate the corresponding secondary structure? During the past 25 years this second question has been investigated extensively (1–15). These analyses are based upon secondary structure classifications obtained from X-ray diffraction (16). However, the comparison of classifications between several investigators reveals discrepancies (cf. Fig. 6 from Yang *et al.* (15)). Since even the X-ray specialists disagree about the criteria for classifying protein secondary structure there is no golden standard available (cf. (16)). Still one can stick to a certain classification method (9) and investigate whether the fractions of secondary structure classes can be estimated adequately from protein CD spectra. This estimation needs a model (8, 9, 13, 14). In this paper a multivariate linear model with noise (Gauss–Markoff model (17)) is applied. An important extension, in com-

0003-2697/90 $3.00

parison with the aforementioned studies, is that this model allows estimation of the error in the secondary structure estimates. We will first explore the basic assumption of the linear model. Then we apply the theory of the Gauss–Markoff model to our problem and evaluate the linear model. To incorporate also nonlinear effects a locally linearized model will be introduced. Finally, both models are compared with two other well-known methods (13, 14).

EXPLORATIONS

The basic assumption of the different methods to solve the inverse problem is the following: a CD spectrum $c(\lambda)$[2] is a superposition of the contributions of the different secondary structure classes. In formula, it can be written as

$$c(\lambda) = \sum_{k=1}^{N_{cl}} f_k b_k(\lambda), \qquad [1]$$

where $b_k(\lambda)$ denotes the CD spectrum belonging to the kth secondary structure class and f_k denotes the fraction of class k. N_{cl} is the number of secondary structure classes. We adopt the classification of Hennessey and Johnson (9) who distinguished five classes: α-helix (H), antiparallel β-sheet (A), parallel β-sheet (P), β-turn (T), and other (O). By definition fractions are non-negative, and summate to one:

$$f_k \geq 0; \quad \sum_{k=1}^{N_{cl}} f_k = 1. \qquad [2]$$

At first glance these two equations pose problems. What if a $b_k(\lambda)$ contributes negligibly to the CD spectrum? Or what if $b_k(\lambda) = -b_{k'}(\lambda)$, like the contributions of left- and right-hand β-turns, which coexist in some proteins, cf. Brahms and Brahms (Ref. (6), p. 173). It will in general be impossible to recover fractions with negligible contributions $b_k(\lambda)$. In addition taking into account the effect of experimental errors upon the CD analysis (10), the application of the normalization constraint $\sum_{k=1}^{N_{cl}} \hat{f}_k = 1$ is questionable. Some authors use this constraint (6, 7, 14) in the estimation of the $\hat{f}_k$, whereas others (8–13) simply consider an analysis with $\sum \hat{f}_k$ far from one as unsuccessful. With the second constraint, $\hat{f}_k \geq 0$, the same dichotomy appears.

Regarding Eq. [1] a natural question to ask is: do similar structures possess similar CD spectra, and vice versa? This brings up the next question: how do we measure similarity? We investigated several measures and chose the root mean square (rms) difference δ, which is a symmetric distance measure:

$$\delta_{x,12} = \left(\frac{1}{N} \sum_{i=1}^{N} (x_{i1} - x_{i2})^2 \right)^{1/2}. \qquad [3]$$

The summation in Eq. [3] extends over N_{cl} when two secondary structure classifications ($x = f$) are compared, and over N_λ when two CD spectra ($x = c$) are involved. An alternative measure is the maximum of the cross-correlation function between two CD spectra, which takes into account the correlations between successive values. We found that this measure produced equivalent results.

To compare distances between CD spectra with distances between the accompanying secondary structure classifications we will use the Pearson product moment correlation coefficient r:

$$r_{xy} = \frac{N \sum_i x_i y_i - \sum_{i,j} x_i y_j}{((N \sum_i x_i^2 - (\sum_i x_i)^2)(N \sum_j y_j^2 - (\sum_j y_j)^2))^{1/2}} = \frac{\frac{1}{N} \sum_i (x_i - \bar{x})(y_i - \bar{y})}{\hat{\sigma}_x \hat{\sigma}_y}. \qquad [4]$$

The summations in Eq. [4] extend over all pair combinations of reference proteins, and we substitute $x = \delta_c$ and $y = \delta_f$. r varies between -1 and 1, where an r of 1 indicates perfect correlation, -1 indicates anticorrelation, and 0 indicates no correlation at all.

We will explore the basic assumption using the reference data of Johnson and co-workers (8–13). For illustrative purposes we have reordered the proteins according to increasing α-helix fraction. A first look at the CD spectra in Fig. 1a is very instructive. Consider protein 22, which is the pure α-helix compound polyglutamic acid, possessing the largest CD spectrum. The similarities of the CD spectra 21, 20, and 19 compared with spectrum 22 are evident, and the corresponding f (Fig. 1b) all indicate a large α-helix content of 70–80%. But for proteins with small α-helix content a connection between CD spectrum and secondary structure classification is hardly discernible, compare the lower halves of Figs. 1a and 1b. It is not surprising that the scatter diagram in Fig. 1c, where we compare the distance between the CD spectra δ_c with the difference in α-helix fraction δ_{f_1}, shows a significant correlation $r_{\delta_c,\delta_{f_1}}$ (H in Fig. 1e). In contrast, the correlations between δ_c and δ_{f_2}, δ_{f_3}, δ_{f_4} are much weaker (A, P, T, in Fig. 1e), and even insignificant in the case of δ_{f_3}. Regarding the correlation with δ_{f_5} (O in Fig. 1e) one needs to be careful, because the fraction f_5 is not classified according to the X-ray analysis but derived directly from the first four classes: $f_5 = 1 - \sum_{k=1}^{4} f_k$. Finally, the correlation between δ_f averaged over the classes and δ_c in Fig. 1d (cross-hatched bar in Fig. 1e) seems to result from the correlation in Fig. 1c, where the largest δ_f's were found. The correlation coefficients between the different classes show a significant negative

[2] See the Appendix for a glossary of terms used in this paper.

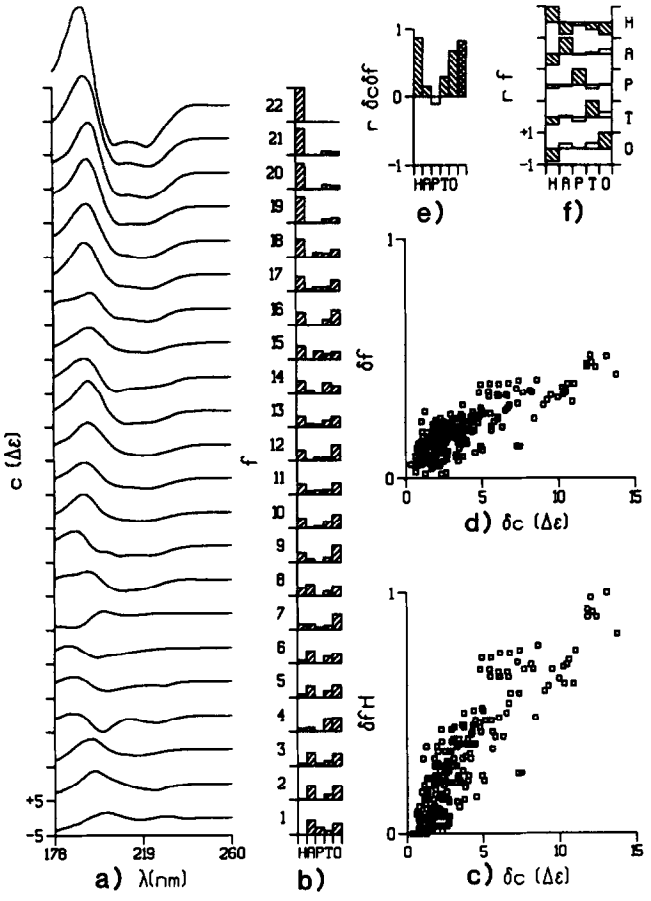

FIG. 1. Comparison of CD spectra (a) and classifications of accompanying secondary structures (b) of 22 reference proteins (data from Johnson and co-workers (9, 11)). The scatter diagrams represent the relation between the CD spectra distance and the accompanying secondary structure distance of all pair combinations of reference proteins. In (c) δ_c (abscissa) is compared to δ_{f_1} (ordinate), whereas in (d) it is compared with δ_f averaged over the five classes. The correlation coefficients belonging to (c) and (d) constitute the outer bars in (e). In between are shown the correlation coefficients for the classes A, P, T, and O. (f) shows the correlation coefficients between the different f_k. The hatching with the negative slope (e and f) indicates correlation coefficients which differ significantly from zero at the 5% level. The reference proteins, ordered according to increasing α-helix fraction, are:

1 Bence-Jones protein
2 concanavalin A
3 prealbumin
4 rubredoxin
5 α-chymotrypsin
6 elastase
7 carboxypeptidase A
8 ribonuclease A
9 papain
10 subtilisin BPN'
11 glyceraldehyde-3-phosphate dehydrogenase
12 subtilisin novo
13 thermolysin
14 lysozyme
15 flavodoxin
16 cytochrome *c*
17 lactate dehydrogenase
18 triose phosphate isomerase
19 hemerythrin
20 hemoglobin
21 myoglobin
22 polyglutamic acid

correlation between on the one hand f_1 and on the other hand f_2, f_4, and f_5 (H and A, T, and O in Fig. 1f). All other correlations between different classes are insignificant.

Summarizing this data impression, the CD spectrum belonging to the α-helix (spectrum 22) is striking and dominates the spectra of compounds with α-helix fractions above 30% (i.e., proteins 10–22 in Fig. 1a). We also note that there are no obvious correlations between the CD spectra and the secondary structure classes A, P, T, and O.

THE LINEAR MODEL

One approach to solve the inverse problem is to determine first CD spectra $b_k(\lambda)$ for the pure secondary structures, and then fit a CD spectrum of an unknown protein with these estimates $\hat{b}_k(\lambda)$ (2, 6, 7). Since these $\hat{b}_k(\lambda)$ are nonorthogonal it is better to start from a set of CD spectra of reference proteins and estimate the parameters of the inverse of Eq. [1] (8, 9). For this estimation we formulate the inverse problem as a multivariate linear model with noise, a so-called Gauss–Markoff model (17). The terminology of the model is largely adopted from Compton and Johnson (8). c is a digitized CD spectrum measured at N_λ wavelengths. $\hat{f}_k$ is the estimate for the fraction of secondary structure class k, belonging to c. C is a matrix which contains the CD spectra that are used as references in its columns, whereas F is the matrix of the accompanying fractions of secondary structure. For each class k the model is formulated as

$$F_k = X_k C + \nu_k. \qquad [5]$$

The coupling row vector X_k, which has to be estimated from the reference data F_k and C, can be considered as the inverse of the $b_k(\lambda)$ in Eq. [1]. It is assumed that the noise ν_k is $N(0, \sigma^2_{\nu,k})$, where $\sigma^2_{\nu,k}$ still needs to be estimated. When we have estimated X_k from the set of reference proteins, we can use this estimate, $\hat{X}_k$, to estimate the fraction f_k from the CD spectrum c of an unknown protein:

$$\hat{f}_k = \hat{X}_k c. \qquad [6]$$

For each class k the solution of the least-squares problem of Eq. [5] is given by (e.g., (17))

$$\hat{X}_k = F_k C^T (CC^T)^\dagger = F_k C^T C^{T\dagger} C^\dagger = F_k C^\dagger. \qquad [7]$$

Here $C^\dagger$ denotes the (Moore–Penrose) generalized inverse of C. Since C is rank deficient we perform a singular value decomposition (18, 19) to find $C^\dagger$:

$$C = USV^T \quad C^\dagger = VS^\dagger U^T. \qquad [8]$$

Here S is a diagonal matrix which contains the singular values in decreasing order. The problem is to determine how many of these singular values are significant. In our further analysis we treat this number as a variable N_s.

The matrix $S^\dagger$ contains as non-zero elements the reciprocals of the significant singular values. It is thus important to determine N_s carefully, since too high an N_s contributes noise to $S^\dagger$. The matrices U and V are orthogonal: $U^{-1} = U^T$. The first N_s columns of the matrix U contain the basis vectors which are used for the decomposition of the CD spectra. The first N_s columns of the matrix V contain the least-squares coefficients which fit C to US (Eq. [8]). Combining Eqs. [7] and [8] we arrive at

$$\hat{X}_k = F_k V S^\dagger U^T. \quad [9]$$

The main difference between this and other models (8, 9) lies in the inclusion of a noise term in Eq. [5], which allows estimation of errors. The covariance of $\hat{X}_k$, $D(\hat{X}_k)$, depends linearly on the covariance of F_k (Eq. [5]):

$$\hat{D}(\hat{X}_k) = C^{\dagger T}\hat{D}(F_k)C^\dagger = \hat{\sigma}^2_{\nu,k} U S^{\dagger T} S^\dagger U^T, \quad [10]$$

where we have used $\hat{D}(F_k) = \hat{\sigma}^2_{\nu,k} I$. The covariance of F_k, $\sigma^2_{\nu,k}$, is estimated by

$$\hat{\sigma}^2_{\nu,k} = \frac{\sum_{i=1}^{N_{ref}} (F_{ki} - \sum_{l=1}^{N_\lambda} \hat{X}_{kl} C_{li})^2}{N_{ref} - N_s}. \quad [11]$$

Finally, the covariance of the estimator $\hat{f}_k = \hat{X}_k c$ is estimated by

$$\hat{D}(\hat{f}_k) = c^T \hat{D}(\hat{X}_k) c = \hat{\sigma}^2_{\nu,k} ||S^\dagger U^T c||^2, \quad [12]$$

where we have assumed that the error in c is uncorrelated for the different λ. Then the product $U^T c$ will be insensitive for noise, and thus the main contribution to $D(\hat{f}_k)$ arises from the covariance of $\hat{X}_k$, which reflects the goodness of fit of the linear model. Thus the error in the secondary structure estimate, $\sqrt{D(\hat{f}_k)}$, depends upon the covariance of the coupling row vector X_k, which in turn depends upon the covariance $\sigma^2_{\nu,k}$.

It is also possible to estimate the CD spectra belonging to the different secondary structures, i.e., the $b_k(\lambda)$ of Eq. [1]. The method is analogous to that outlined above for the estimation of X_k. Now the model reads, for each wavelength λ,

$$C_\lambda = B_\lambda F + \nu_\lambda \quad \nu_\lambda \sim N(0, \sigma^2_{\nu,\lambda}), \quad [13]$$

with as solution

$$\hat{B}_\lambda = C_\lambda F^\dagger = C_\lambda V_F S_F^\dagger U_F^T \quad [14]$$

and covariance estimate

$$\hat{D}(\hat{B}_\lambda) = F^{\dagger T}\hat{D}(C_\lambda)F^\dagger = \hat{\sigma}^2_{\nu,\lambda} U_F S_F^{\dagger T} S_F^\dagger U_F^T, \quad [15]$$

where

$$\hat{\sigma}^2_{\nu,\lambda} = \frac{\sum_{i=1}^{N_{ref}} (C_{\lambda i} - \sum_{k=1}^{N_{cl}} \hat{B}_{\lambda k} F_{ki})^2}{N_{ref} - N_{s_F}}. \quad [16]$$

Note that the estimate of $b_k(\lambda)$ consists of N_λ estimates $\hat{B}_{\lambda k}$.

ADEQUACY OF THE MODEL

An often used indicator for the quality of the estimate $\hat{f}$ is the difference $\delta_{c,\hat{c}}$ between the original CD spectrum and its reconstruction $\hat{c}$:

$$\hat{c} = CVS^\dagger U^T c = USS^\dagger U^T c. \quad [17]$$

This $\delta_{c,\hat{c}}$ measures to what extent the original CD spectrum is reconstructed using the first N_s orthogonal basis vectors of U. Thus it is a monotonically declining function of N_s. To test the model we use one member from the database as a test protein, estimate the model from the remaining 21 reference proteins, and estimate the f of the test protein. This is done for all 22 members of the database.

The adequacy of the model is determined in three ways:

First, the rms difference between the estimation from the CD spectrum and the X-ray data, $\delta_{f,\hat{f}}$, is estimated, together with its standard deviation $\hat{\sigma}(\delta_{f,\hat{f}})$, for each test protein pooled over its five structure classes and for each structure class pooled over the 22 possible test proteins. In the following $\delta_{f,\hat{f}}$ will be termed *rms residual.*

Second, a significant correlation coefficient $r_{f_k,\hat{f}_l}$ (Eq. [4]) indicates the presence of a linear relation between f_k and $\hat{f}_l$. These $r_{f_k,\hat{f}_l}$ are calculated for all pair combinations of secondary structure classes, pooled over the 22 possible test proteins.

Third, since we want an unbiased estimate of f_k, the hypothesis that the relation between f_k and $\hat{f}_k$ is described by

$$\hat{f}_k = \beta f_k + \mu \quad \mu \sim N(0, \sigma^2_\mu) \quad [18]$$

with $\beta = 1$ is tested. This is done, for each secondary structure class k, by means of Student's t test with

$$t = \frac{|\hat{\beta} - 1|}{\hat{\sigma}_{\hat{\beta}}} = |\hat{\beta} - 1| \sqrt{\frac{(N-1)\sum_i f_i^2}{\sum_i (f_i^2 - \hat{\beta} \hat{f}_i f_i)}}$$

$$\hat{\beta} = \frac{\sum_i f_i \hat{f}_i}{\sum_i f_i^2} \quad [19]$$

where the summations extend over the f_k of all the 22 test proteins. With this test one can only conclude whether or not the hypothesis has to be rejected. Rejection implies that the estimate is biased.

RESULTS

Our first concern is to determine the number of significant singular values, N_s. The results of the linear model as a function of N_s are shown in Fig. 2 for three test proteins. The rms error between reconstructed and original CD spectra, $\delta_{c,\hat{c}}$ in Fig. 2a, shows a steady decrease with increasing N_s. The usual noise level of experimental CD spectra, about $0.3\Delta\epsilon$ (9), is reached with $N_s = 5$. Thus the fit of the CD spectrum over this wavelength range needs about five singular values (9). However, the rms residual $\delta_{f,\hat{f}}$ in Fig. 2b shows quite a different picture. Minimal $\delta_{f,\hat{f}}$ is reached with different N_s for these three different proteins. With hemoglobine (protein 20, squares, 75% α-helix) the smallest $\delta_{f,\hat{f}}$ is 0.08, which is reached with $N_s = 1$. Addition of two more singular values deteriorates the prediction until $\delta_{f,\hat{f}} = 0.24$. The other two proteins in Fig. 2 reach minimal $\delta_{f,\hat{f}}$ at $N_s = 7$ (triangles) and $N_s = 3$ (circles). Thus although the addition of more singular values improves the fit of the CD spectrum, it sometimes deteriorates the secondary structure prediction.

The CD spectra of Fig. 1a are almost noise-free due to smoothing. In Fig. 2c we simulated a more realistic situation by adding (Gaussian white) noise to the test CD spectrum. This allows us to estimate an upper bound for the number of significant singular values N_s. Figure 2c shows that $\delta_{f,\hat{f}}$ starts to rise after $N_s = 7$. We conclude that singular values above seven represent noise and can be considered insignificant.

Following Johnson and co-workers (8, 9, 13) we will evaluate the linear model with $N_s = 5$, the number needed to fit the CD spectrum. The overview of the linear model in Fig. 3 shows that the rms residuals are still quite large. For the test proteins $\delta_{f,\hat{f}}$ varies between 0.05 and 0.23 (Fig. 3c). When the residuals $(\hat{f}_k - f_k)$ are related to the values of f_k (Fig. 3a), $\delta_{f,\hat{f}}$ for the structure classes (Fig. 3f) is only acceptably small for class H (α-helix). For instance, with the P class the rms residual (Fig. 3f) is large compared to the f_3 values (triangles in Fig. 3a). There appears to be a clear underestimation of f_2 and f_3 (note the negative residuals with A and P in Fig. 3a). There is no significant correlation between $\delta_{c,\hat{c}}$ (Fig. 3d) and $\delta_{f,\hat{f}}$ (Fig. 3c) ($r = 0.40$, df = 20, $P > 0.05$). Note that, as to be expected, the test proteins whose $\sum_{k=1}^{N_{el}} \hat{f}_k$ is far away from 1 (Fig. 3b) also show a large $\delta_{f,\hat{f}}$. There is a significant correlation between $|1 - \sum \hat{f}_k|$ and $\delta_{f,\hat{f}}$ ($r = 0.73$, df = 20, $P < 0.001$). The correlation coefficients between f_k and $\hat{f}_l$ (Fig. 3e) show on the diagonal only a significant correlation $r_{f_l,\hat{f}_l}$. The upper row is approximately equal to the upper row in Fig. 1f, which confirms that class H is estimated well. Quite astonishing $\hat{A}$ shows the best correlation with T. Comparing rows two to five in Fig. 1f and Fig. 3e we see the inadequacy of the estimates $\hat{A}$, $\hat{P}$, $\hat{T}$, and $\hat{O}$ corroborated. Figure 4 shows estimates (solid) of the coupling matrices X and B, together with their errors (dotted). Note that only for $\hat{X}_1$ and $\hat{b}_1(\lambda)$ is the error relatively small, whereas with the other components the error is about as large as the absolute values of the $\hat{X}_k$ and $\hat{b}_k(\lambda)$. The dotted lines in Fig. 4b show the same shape, which follows from Eq. [15], where it is seen that for all classes the covariance is proportional to $\hat{\sigma}^2_{r,\lambda}$. The dotted lines in Fig. 4a also possess the same shape, here the covariances are proportional to $US^{\dagger T}S^{\dagger}U^T$ (Eq. [10]). Figure 4c shows the product of X and B, which in the ideal case should result in the identity matrix. It is clear that only the (H, H) element of XB suffices. All other diagonal elements are less than 0.5, and there are large off-diagonal elements

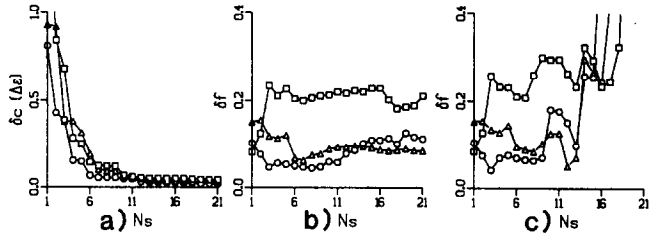

FIG. 2. Overview of the model prediction for proteins 20 (squares), 17 (circles), and 14 (triangles) as a function of N_s, the number of singular values. (a) $\delta_{c,\hat{c}}$; (b) $\delta_{f,\hat{f}}$; (c) $\delta_{f,\hat{f}}$ with a Gaussian white noise added to the CD spectra of the test proteins, $\sigma_\nu = 1.0\Delta\epsilon$.

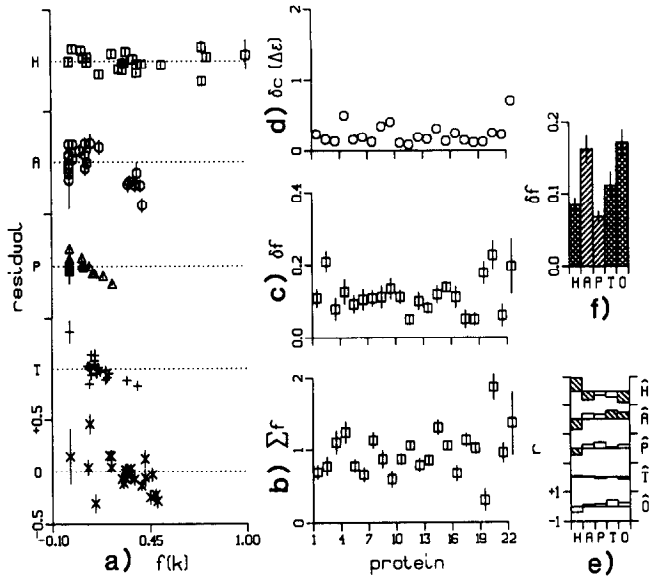

FIG. 3. Overview of the results of the linear model with $N_s = 5$. (a) $\hat{f}_k - f_k$ (ordinate) as a function of f_k (abscissa) for the 22 different test proteins. The residuals are shown for the five different classes (indicated by squares, circles, triangles, plusses, and crosses, respectively). The vertical lines indicate plus or minus one standard deviation ($\hat{\sigma}_{\hat{f}_k}$). The dotted lines indicate $\hat{f}_k - f_k = 0$. (b) $\sum_{k=1}^{N_{el}} \hat{f}_k$ (ordinate) for the 22 different test proteins (abscissa). The vertical lines indicate plus or minus one standard deviation ($\hat{\sigma}(\sum_{k=1}^{N_{el}} \hat{f}_k)$). (c) $\delta_{f,\hat{f}}$ for the 22 test proteins. (d) $\delta_{c,\hat{c}}$ for the 22 test proteins. (e) $r_{f,\hat{f}}$, the hatching indicates correlation coefficients significantly different from zero, at the 5% level. (f) $\delta_{f,\hat{f}}$ per secondary structure class. The cross-hatching indicates that the hypothesis $\hat{f}_k = f_k$ was not rejected at the 5% level.

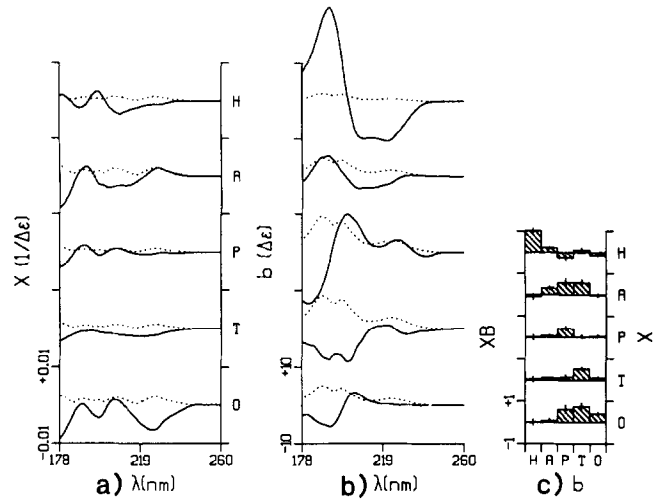

FIG. 4. Overview of the coupling matrices X and B of the linear model. Both have been estimated from the complete reference set ($N_{ref} = 22$) with $N_s = 5$. (a) From top to bottom $\hat{X}_1, \ldots, \hat{X}_5$, the dotted lines indicate the standard error belonging to each estimate. (b) From top to bottom $\hat{b}_1$ (λ), . . . , $\hat{b}_5$ (λ). (c) The product of $\hat{X}_k$ (ordinate) and $\hat{b}_l$ (abscissa). Vertical lines indicate standard deviation.

(XB(A, P), (A, T), (O, P), (O, T)). These deviations from the identity matrix indicate that the linear model is far from ideal.

In summary, the linear model predicts only the α-helix fraction accurately.

THE LOCALLY LINEARIZED MODEL

There are two complementary ways to improve the linear model: one can remove from the reference set those proteins which add conflicting information (13), or one can synthesize an appropriate reference set. According to the basic assumption the N_{cl}-dimensional f space of secondary structure classification and the N_λ-dimensional c space of CD spectra are related through Eq. [1]. We hypothesize that this assumption applies only to *regions* of f space which are related to *regions* of c space. Thus for the proteins 19–22 with large f_1 a different model is needed as for the proteins with small α-helix content. In this way nonlinear effects, like the chain length dependence of the CD spectrum of α-helix (11), can be incorporated. Looking back at Figs. 1c and 1d we note that in general a small δ_c correlates with a small δ_f. To synthesize a locally linearized model we adopt the hypothesis that those reference proteins with a small δ_c relative to the test protein are more likely to contribute valuable structural information. We reordered the reference proteins according to δ_c, and now repeated the analysis as a function of the number of reference proteins N_{ref} and of N_s. Thus a multitude of estimates is generated. To choose from this multitude we adopted the following selection criteria: $-0.05 \leq \hat{f}_k \leq 1.05$ (note the circles and triangles in the lower left corners in Fig. 3a which indicate negative $\hat{f}_k$). Furthermore we chose the estimate whose $\sum \hat{f}_k$ was nearest to 1. As a refinement the estimate with the smallest $\hat{\sigma}(\sum_{k=1}^{N_{cl}} \hat{f}_k)$ can be chosen among the 10 estimates with $\sum \hat{f}_k$ nearest to 1. This selection procedure thus resulted in models fine-tuned per test protein, with varying N_{ref} and N_s.

The selection procedure is illustrated in Fig. 5. We note first that the three references with smallest δ_c also show the smallest δ_f (circles and squares, respectively, in Fig. 5a). It should be stressed that for an unknown protein the $\delta_{f,\hat{f}}$ of Fig. 5b are not available. For an unknown protein we want to select a (N_{ref}, N_s) pair with a small $\delta_{f,\hat{f}}$. In this case of a protein with a dominant contribution of α-helix, small $\delta_{f,\hat{f}}$ are reached with $N_s = 1$ or 2. A requirement for our selection procedure is that those N_s values are selected. The $\sum \hat{f}_k$ in Fig. 5c shows only a few estimates near 1, and the estimates with (N_{ref}, N_s) = (21, 1) and (3, 1) are nearest to 1. The latter possesses the smallest $\hat{\sigma}(\sum \hat{f}_k)$. Thus with the refined selection we find the minimum $\delta_{f,\hat{f}}$ of Fig. 5b, whereas with the selection of the estimate with $\sum \hat{f}_k$ nearest to 1 we get the fifth best estimate. These two estimates have $\delta_{f,\hat{f}}$ of 0.04 and 0.08, which are appreciable improvements compared to Fig. 3 ($N_{ref} = 21$, $N_s = 5$, $\delta_{f,\hat{f}} = 0.23$).

An overview of the locally linearized model with the selection of the estimate with $\sum \hat{f}_k$ nearest to 1 is shown in Fig. 6. With all but two of the test proteins a smaller

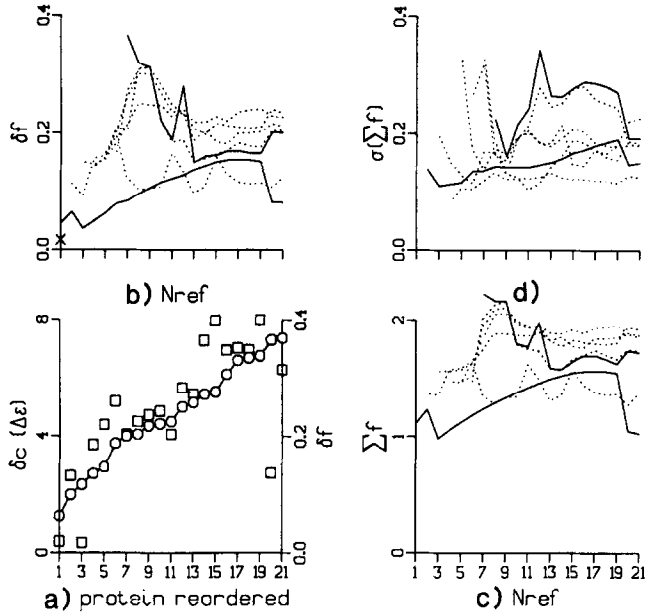

FIG. 5. Overview of the locally linearized model estimates for protein 20 (hemoglobin). The reference set is reordered according to distance δ_c relative to the test protein. (a) δ_c (circles) and δ_f (squares, right ordinate) of reordered reference proteins. (b) $\delta_{f,\hat{f}}$ for the different models as a function of N_{ref} (abscissa). The solid lines indicate N_s = 1 (starting at the left) and N_s = 7 (starting near the middle), whereas the dotted lines represent the intermediate N_s values. (c) $\sum \hat{f}_k$ and (d) $\hat{\sigma}(\sum \hat{f}_k)$, both according to the format of (b).

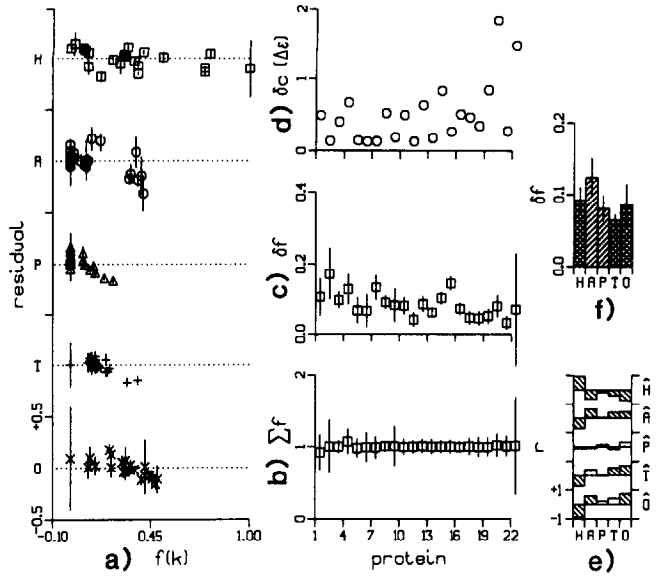

FIG. 6. Overview of the results of the locally linearized model with the selection of $\sum_{k=1}^{N_{sl}} \hat{f}_k$ nearest to one (see text). Format as in Fig. 3.

$\delta_{f,\hat{f}}$ is found, compare Figs. 3c and 6c. The largest improvements are found with proteins 20, 22, and 19, which were proteins with large CD spectra because of their α-helix fraction being larger than 70%. The selection procedure does not take into account the quality of the fit of the CD spectrum. The differences in $\delta_{c,\hat{c}}$ between Figs. 3d and 6d indicate that in the majority of the cases an estimate with $N_s < 5$ is selected. Although this deteriorates the fit of the CD spectrum, it reduces the noise in $S^\dagger$ because less singular values are taken into account. The correlation coefficients $r_{f_k,\hat{f}_k}$ are larger for f_2, f_4, and f_5 (A, P, and O)—compare the diagonals of Figs. 3e and 6e. But there are still differences in rows two to five when we compare Figs. 1f and 6e. The rms residual in Fig. 6f is also appreciably smaller for classes A, T, and O (compare Fig. 3f). Like before the $\hat{f}_k = f_k$ hypothesis is rejected for classes A and P (cf. the circles and triangles below the dotted lines in Fig. 6a).

COMPARISON OF DIFFERENT METHODS

In this section we compare the residuals of the models presented in this paper with those of two methods well known from the literature. All methods are applied to the data set of Fig. 1. Manavalan and Johnson (13) extended the generalized inverse method (i.e., the linear model without noise) with a variable selection procedure. They deleted triplets of reference proteins from the reference set and selected the estimates which fulfilled the following criteria: $0.9 \leqslant \sum_k \hat{f}_k \leqslant 1.1$, $\hat{f}_k \geqslant -0.05$, and $\delta_{c,\hat{c}} \leqslant 0.22\Delta\epsilon$ (the measurement error). They applied brute force, testing removal of up to three reference proteins (1562 combinations when $N_{ref} = 21$). The final estimate is the average of all estimates that fulfill the criteria. Provencher and Glöckner (14) applied a damped least-squares method (also called ridge regression (19)) in which they directly fitted the CD spectrum with the spectra of the reference set. With zero damping, and $N_s = 5$, their method is equivalent to the linear model without noise but with the constraints of Eq. [2]. With damping the method is biased toward reference proteins whose CD spectra resemble the test spectrum, and thus the method resembles more the locally linearized model.

The first row of Table 1 represents the linear model (cf. Fig. 3f). The second and third row summarize the improvements which can be achieved with the locally linearized model and the two different selection procedures. The greatest gain in accuracy is found with classes O, A, and T. The variable selection method of Manavalan and Johnson (13) failed to produce estimates that fulfill the criteria with proteins 6, 19, 20, and 22. When we retained for these proteins the estimate of the linear model (row one), this method showed only a minor improvement (row four). Apart from these four proteins the method is about equal to that of row two. The damped least-squares method of Provencher and Glöckner (14) produces a series of solutions which depend on the damping parameter. Selection according to their criteria results in row five, which is not better than the linear model. As pointed out by Manavalan and Johnson (13) the best results with this method are reached when the estimate with five degrees of freedom is chosen (row six). Then the rms residuals are between those of rows two and three.

With all methods, except for row five, the linear hypothesis $\hat{f}_k = f_k$ had to be rejected for classes A and P. With the improved damped least-squares method the hypothesis was also rejected for class H, because it pro-

TABLE 1

Comparison of Root Mean Square Residuals $\delta_{f,\hat{f}}$ (×100) of the Five Classes for Different Analysis Methods

Method	Secondary structure class				
	H	A	P	T	O
Linear model[a]	9	16	7	11	17
Locally linearized model					
I[b]	9	12	8	7	9
II[c]	7	12	7	7	8
Variable selection (13)	9	14	7	11	14
Damped least squares (14)					
I[d]	12	21	6	11	8
II[e]	9	13	6	7	8

[a] $N_{ref} = 21$, $N_s = 5$.
[b] Selection of estimate with $\sum \hat{f}_k$ nearest to 1.
[c] Selection of estimate with $\sum \hat{f}_k$ near 1 and smallest $\hat{\sigma}$ ($\sum \hat{f}_k$).
[d] Selection of estimate according to criteria of (14).
[e] Selection of estimate with five degrees of freedom (13).

duced a slight but significant underestimation. Regarding the amount of computation time we found that the linear model required 1 s per protein on a SUN 4/280 minicomputer. The locally linearized model and the damped least-squares method consumed 10 s, whereas the brute force variable selection method required 1000 s.

DISCUSSION

The basis for the estimation of protein secondary structure from the CD spectrum is expressed in the linear relations of Eqs. [1] and [5]. In Figs. 1c and 1d we noted that in general, similar structures produce similar CD spectra. Still there is a large scatter in these figures. Part of this scatter is due to nonlinear effects, like the chain length dependence of the CD spectra of α-helix and β-sheet (Figs. 39, 40, and 43 in Ref. (11)) and the contribution of aromatic side chains to the CD spectrum (4–6, 12). These reasons for the scatter contribute to the inadequacy of the linear model. The local linearization circumvents this problem, by restricting the set of reference proteins to those with small δ_c, which provided an appreciable improvement, especially with the proteins with more than 70% α-helix. We expect that the locally linearized model will benefit from a larger set of reference proteins, thus providing a sophisticated interpolation method which leaves out inappropriate information.

Still there remains the problem which criteria should be used for the selection of a solution. The *fit* of the CD spectrum indicated by small $\delta_{c,\hat{c}}$ is not a good predictor for small $\delta_{f,\hat{f}}$ (cf. Figs. 2a, 2b, 3c, 3d, 6c, and 6d). The criteria of Manavalan and Johnson (13) provide an alternative. We found that next to the selection of the estimate with $\sum \hat{f}_k$ near 1, the refined selection according to small $\hat{\sigma}(\sum \hat{f}_k)$ produced the best results in Table 1. But we observe in Figs. 3a and 6a that the error bars often do not cross the line $\hat{f}_k - f_k = 0$. On the one hand the refined selection produced the smallest rms residuals, by selecting an estimate with small $\hat{\sigma}(\sum \hat{f}_k)$. On the other hand, the residuals are larger than the estimated errors, which pleads against selecting an estimate with a small $\hat{\sigma}(\hat{f}_k)$. Since with both selection methods the residuals are larger than the $\hat{\sigma}(\hat{f}_k)$ we conclude that the error estimate $\hat{\sigma}(\hat{f}_k)$ is only a lower bound.

With the selection according to $\sum \hat{f}_k$ near 1 the independence of the estimates of the different f_k disappears. It was discussed already under Explorations that the constraint is sometimes questionable. Furthermore, when the CD spectrum contains experimental errors, the use of the constraint can lead to failure (10). To deal with errors in the protein concentration we suggest to select those estimates whose $\sum \hat{f}_k$ is within 1 plus or minus the error bounds of the concentration estimate. Since all improvements upon the linear model use the constraint $\sum \hat{f}_k$ equal (14) or near to 1 (Ref. (13) and the locally linearized model) accurate determination of the concentration is of paramount importance.

The model measure that is used most in the literature, $r_{f_k,\hat{f}_k}$ pooled over all test proteins (the diagonals in Figs. 3e and 6e), seems to us an inappropriate measure. The aim of the model is not to produce a linear relation between $\hat{f}_k$ and f_k, but to find an unbiased estimate equal to f_k (Eq. [18]). With class A (antiparallel β-sheet) a significant correlation coefficient $r_{f_2,\hat{f}_2}$ is found (Fig. 6e) but the hypothesis $\hat{f}_2 = f_2$ had to be rejected. It is clear from the circles in Fig. 6a that $\hat{f}_2 - f_2$ is negatively correlated with f_2, especially with the larger f_2 values. From the estimates of the basis CD spectra $b_k(\lambda)$ one might infer the reasons for the inaccuracy of $\hat{f}_2$. $b_1(\lambda)$ and $b_2(\lambda)$ show a large resemblance, with $b_1(\lambda)$ being about three times as large as $b_2(\lambda)$ (Fig. 4b). Since f_1 and f_2 are negatively correlated (Fig. 1f), one expects that: (i) with large f_1 and small f_2 the CD spectrum will be large and a small f_2 is easily overestimated (cf. the left group of circles above the dotted line in Fig. 6a); (ii) with small f_1 and large f_2 the CD spectrum will be small and a small f_1 may account for a large part of the CD spectrum, thereby causing underestimation of f_2 (cf. the right group of circles below the dotted line in Fig. 6a).

Thus the resemblance of $b_1(\lambda)$ and $b_2(\lambda)$ together with their chain length dependence is responsible for the inadequacy of the antiparallel β-sheet estimate.

Regarding the underestimation of f_3 one must keep in mind that only a small range of f_3 is represented in the reference set (cf. the distribution of the triangles in Fig. 6a) including many f_3 equal to zero. Thus it is not surprising that with so little information about f_3 available its estimate is inaccurate.

From the comparison of the different methods we conclude that with the standard linear model only the α-helix class can be estimated accurately. The locally linearized model estimates also the β-turn and the other class adequately. But the estimation of the antiparallel and parallel β-sheet class remains unsatisfactory. Furthermore the residuals of about 0.08 (Table 1) warn against overinterpretation of the estimates. The method of Manavalan and Johnson (13) failed with 4 of the 22 proteins. With the other proteins the residuals were comparable to the locally linearized model. The method of Provencher and Glöckner (14) with the selection of the estimate with five degrees of freedom (13) was as good as the locally linearized model. However, the advantage of the latter is the estimation of the standard deviations $\hat{\sigma}(\hat{f}_k)$. This error estimate is useful for the appreciation of the secondary structure estimate of an unknown protein. Furthermore it facilitates the quantitative interpretation of differences between protein CD spectra which result from temperature, pH, or solvent variation.

APPENDIX: GLOSSARY

$b_k(\lambda)$	CD spectrum belonging to secondary structure class k
B_λ	Row vector which describes the coupling between C_λ and F
B	Matrix consisting of rows B_λ and columns $b_k(\lambda)$
$c(\lambda)$, c	CD spectrum
C	Matrix whose columns contain the CD spectra of the reference proteins
C_λ	Row of matrix C
D	Covariance matrix
df	Degrees of freedom
δ	Root mean square difference (Eq. [3])
f_k	Fraction of secondary structure class k
F_k	Row vector consisting of the f_k of the reference proteins
F	Matrix whose columns contain secondary structure classification of the reference proteins
N_{cl}	Number of secondary structure classes
N_{ref}	Number of proteins in the reference set
N_λ	Number of wavelengths of CD spectrum
N_s	Number of significant singular values
ν	Gaussian white noise with zero mean and standard deviation σ_ν, $N(0, \sigma_\nu^2)$
r	Pearson product–moment correlation coefficient (Eq. [4])
σ	Standard deviation
USV^T	Singular value decomposition (Eq. [8])
X_k	Row vector which describes the coupling between F_k and C
X	Matrix consisting of rows X_k
$Z^\dagger$	(Moore–Penrose) generalized inverse of Z
$\hat{z}$	Estimate of z
Z^T	Transpose of Z

Secondary Structure Classes

H, $k = 1$	α-helix
A, $k = 2$	antiparallel β-sheet
P, $k = 3$	parallel β-sheet
T, $k = 4$	β-turn
O, $k = 5$	other

ACKNOWLEDGMENTS

We thank Drs. Johnson and Provencher for providing their analysis programs and protein data. Drs. Bloemendal and Van Grondelle are financially supported by the Netherlands Organization for Scientific Research (NWO).

REFERENCES

1. Cantor, C. R., and Schimmel, P. R. (1980) Biophysical Chemistry. Part II: Techniques for the study of biological structure and function, Freeman, New York.
2. Greenfield, N., and Fasman, G. D. (1969) *Biochemistry* **8,** 4108–4116.
3. Bayley, P. M. (1973) *Prog. Biophys. Mol. Biol.* **27,** 3–76.
4. Sears, D. W., and Beychok, S. (1973) *in* Physical Principles and Techniques of Protein Chemistry (Leach, S. J., Ed.), Part C, pp. 445–593, Academic Press, New York.
5. Woody, R. W. (1985) *in* The Peptides, (Hruby, V., Ed.), Vol. 7, pp. 15–114, Academic Press, New York.
6. Brahms, S., and Brahms, J. (1980) *J. Mol. Biol.* **138,** 149–178.
7. Chang, C. T., Wu, C.-S. C., and Yang, J. T. (1978) *Anal. Biochem.* **91,** 13–31.
8. Compton, L. A., and Johnson, W. C., Jr. (1986) *Anal. Biochem.* **155,** 155–167.
9. Hennessey, J. P., Jr., and Johnson, W. C., Jr. (1981) *Biochemistry* **20,** 1085–1094.
10. Hennessey, J. P., Jr., and Johnson, W. C., Jr. (1982) *Anal. Biochem.* **125,** 177–188.
11. Johnson, W. C., Jr. (1985) *Methods Biochem. Anal.* **31,** 61–163.
12. Johnson, W. C., Jr. (1988) *Annu. Rev. Biophys. Chem.* **17,** 145–166.
13. Manavalan, P., and Johnson, W. C., Jr. (1987) *Anal. Biochem.* **167,** 76–85.
14. Provencher, S. W., and Glöckner, J. (1981) *Biochemistry* **20,** 33–37.
15. Yang, J. T., Wu, C.-S. C., and Martinez, H. M. (1986) *in* Methods in Enzymology (Hirs, C. H. W., and Timasheff, S. N., Eds.), Vol. 130, pp. 208–269, Academic Press, San Diego, CA.
16. Levitt, M., and Greer, J. (1977) *J. Mol. Biol.* **114,** 181–239.
17. Koch, K.-R. (1988) Parameter Estimation and Hypothesis Testing in Linear Models, Springer, Berlin.
18. Forsythe, G. E., Malcolm, M. A., and Moler, C. B. (1977) Computer Methods for Mathematical Computations, Prentice Hall, Englewood Cliffs, NJ.
19. Lawson, C. L., and Hanson, R. J. (1974) Solving Least Squares Problems, Prentice Hall, Englewood Cliffs, NJ.