J. Appl. Cryst. (1996). **29**, 547–554

The Use of Small-Angle Scattering and the Maximum-Entropy Method for Shape-Model Determination from Distance-Distribution Functions

JÜRGEN J. MÜLLER,[a] STEEN HANSEN[b] AND HANS-VOLKER PÜRSCHEL[a]

[a]*Max-Delbrück-Center for Molecular Medicine, Robert-Rössle-Strasse 10, D-13122 Berlin, Germany, and* [b]*Department of Mathematics and Physics, Royal Veterinary and Agricultural University, Thorvaldsensvej 40, 1871 FRB C, Denmark*

(*Received 11 October 1995; accepted 12 April 1996*)

Abstract

The maximum-entropy method is well established for the analysis of scattering data [Bricogne (1993). *Acta Cryst.* D**49**, 37–60]. For this method, prior structure knowledge can be included in the structure determination. This prior estimate is an essential element for a successful application of the maximum-entropy method. The most likely prior estimate can be found by maximization of the entropy. With the assumption *a priori* of a special type of structure model, the unknown parameters can be calculated from real-space functions. For practical use, analytical expressions for the Fourier transform of model scattering curves, the distance-distribution function of the models, are of interest. Formulas are presented for rotational ellipsoids, Gaussian chains and two-phase spheres, and a parameter estimation by the program *MAXENT* is demonstrated for the ellipsoidal shape of cytochrome *c* using theoretical X-ray scattering curves calculated from atomic coordinates. The calculated dimensions of prolate and oblate ellipsoids agree within the error limits with the direct structure-related inertia-equivalent ellipsoid of the molecule. Furthermore, error limits have been determined from the *a posteriori* probability or 'evidence' function for the model parameters. To avoid overinterpretation of the scattering data, the real number of degrees of freedom is calculated for noisy data. This measure of information content is almost independent of the collimation distortion but strongly influenced by the statistical noise in the scattering data. The numerical value is smaller than the ideal number of degrees of freedom provided by the information theory.

1. Introduction

Small-angle X-ray and neutron scattering contain information about shape and inner electron or scattering-length density inhomogeneities of a scatterer. Usually, shape-parameters of a scatterer are determined by a trial-and-error process that includes cycles of modelling and comparison of model and experimental data in either reciprocal or real space (for details see the textbooks: Glatter & Kratky, 1982; Feigin & Svergun, 1987). In special cases, for a relatively crude straightforward estimation of shape parameters with a resolution of about 2.0 nm, automatic curve-fitting procedures of scattering curves of physical models to the innermost angular region of experimental uncorrected slit-smeared (Sjöberg, 1977) or corrected desmeared (Müller, Damaschun & Hübner, 1979) scattering curves have been used. Stuhrmann (1970) developed a method to represent a scatterer by a set of mathematical model functions – by a series of spherical harmonics. All these methods for a direct parameter estimation need some additional *a priori* information from other methods or assumptions, *e.g.* about the type or symmetry class of the particles under investigation. Equivalent results would be obtained by a curve fitting of the corresponding model distance-distribution function $p(x)$ and its experimental counterpart, because of equal information content in the two types of curves. Preconditions of shape-parameter estimation from experimental data in real space are analytical or semianalytical expressions for the distance-distribution function of a physical or mathematical model. Analytical expressions for $p(x)$ are known for spheres (Guinier & Fournet, 1955), for a random distribution of holes and solid (Debye, Anderson & Brumberger, 1957), for cubes (Goodisman, 1980), for aggregates of spheres (Glatter, 1980) and for cones (Gille & Handschug, 1995). Glatter (1981) gave an overview for direct parameter estimation in real space for some simple model bodies such as platelets, infinitely long cylinders and spherical shells. We present here the analytical formula for rotational ellipsoids and Gaussian chains, to add to the collection mentioned above. A formula for a two-phase sphere is also provided here for use in maximum-entropy algorithms, although it is implicitly contained in the general formula for spherical symmetric multishell systems presented by Glatter (1981).

If a special model type or symmetry is assumed to be adequate for describing the sample structure, the unknown parameters of such a prior estimate can be found by the maximum-entropy method maximizing the 'evidence', and error limits of the estimated model parameters can be given. Furthermore, by the maximum-entropy method the real maximum number of degrees of freedom (free parameters) can be determined for an experimental scattering curve instead of the idealized number, formally estimated for noiseless data by using the rules of information theory (Goldman, 1954; Damaschun, Müller & Pürschel, 1968). Consequently, the experimentalist is less likely to overinterpret the experimental data. This will be discussed for a theoretical noisy scattering curve of cytochrome *c*.

2. Theory

2.1. *Small-angle scattering*

In small-angle scattering, the scattered intensity $I(s)$ is linked to the spherically averaged autocorrelation function $C(x)$ of the excess electrons (for X-rays) or scattering lengths (for neutrons) in a particle by the Fourier transformation

$$I(s) = 4\pi \int_0^L xC(x)[\sin(sx)/s]\,\mathrm{d}x \qquad (1)$$

and for the autocorrelation function the following then holds:

$$C(x) = (1/2\pi^2) \int_0^\infty sI(s)[\sin(sx)/x]\,\mathrm{d}s. \qquad (2)$$

s is the length of the scattering vector $(4\pi/\lambda)\sin\theta$, 2θ is the scattering angle, λ is the wavelength of the radiation and L is the largest diameter of the particles or the correlation range.

Frequently, the more obviously direct structure-related electron (for X-rays) or scattering-length (for neutrons) *distance-distribution function*

$$p(x) = 4\pi x^2 C(x) \qquad (3)$$

is used for data correction and modelling. A normalization can be done by

$$p^*(x) = 4\pi x^2 C(x)/V_c, \qquad (4)$$

where

$$V_c = 4\pi \int_0^L x^2 C(x)\,\mathrm{d}x \qquad (5)$$

is the so-called correlation volume, being identical with the geometrical volume of particles with constant electron and scattering-length density, respectively.

In the maximum-entropy method, the experimental distance distribution $p(x)$ is approximated by $\mathbf{p} = (p_1, \ldots, p_N)$ and the intensity is expressed as

$$I(s_i) = \sum_{j=1}^{N} A_{ij}p_j + \varepsilon_i, \qquad (6)$$

where ε_i is the statistical noise at point s_i and the matrix elements are defined by

$$A_{ij} = \Delta x \sin(s_i x_j)/s_i x_j, \quad \Delta x = x_j - x_{j-1}. \qquad (7)$$

2.2. *The maximum-entropy method*

The maximum-entropy method estimates the experimental distance distribution by solving the equation

$$\nabla(\alpha S + \chi^2/2) = 0, \qquad (8)$$

so that the gain of information corresponding to a *prior* estimate is minimized in real space and the least-squares sum from transformed and experimental scattering curves will be minimized also. In (8), the entropy of the distance distribution p_j is defined by (see *e.g.* Skilling, 1988):

$$S(p,m) = \sum_{j=1}^{N} -p_j \ln(p_j/m_j) + p_j - m_j, \qquad (9)$$

where m_j is a *prior estimate* of p_j. Different types of prior estimates were introduced recently (Müller & Hansen, 1994). A *prior estimate* was named *intrinsic prior* if no model was used but the transform of the low-resolution part of the scattering curve itself. A two-step procedure renders possible the slit-distortion corrections or the transformation of the complete small- and wide-angle scattering curves of any resolution without structural preknowledge. If prior information about the structure is used, a *model prior*, meaning the distance distribution of a model, can be chosen.

The quality of the fit is determined by the usual χ^2:

$$\chi^2 = \sum_{i=1}^{M} [I(s_i) - I_{\mathrm{fit}}(s_i)]^2/\sigma_i^2, \qquad (10)$$

where $I(s_i)$ is the scattered intensity measured at point s_i, M is the number of data points, $I_{\mathrm{fit}}(s_i)$ is the fit of the data point and σ_i is the standard deviation of the Gaussian noise at point s_i. In (8), α is a Lagrange multiplier, the value of which determines the relative weighting of the *prior estimate* (given by the value for the entropy) and the constraint from the measured data (given by the value for the χ^2). To a given value of α corresponds a given value for the χ^2. In the absence of constraints from measured data, maximizing the entropy will give the *prior estimate* as the result. For the application of the maximum-entropy method (and similar methods for inverse problems) the major problems are the determination of the *prior estimate* (or model) and the α (or the χ^2) as the noise level is usually not known exactly. By the use of the recent developments of Bayesian methods for data analysis (Gull, 1989), both of these can be determined by maximizing of the posterior probability for the data. This posterior probability is usually termed the *evidence*. It can be shown that

$$\textit{evidence} \propto \exp(\alpha S - \chi\chi/2)\det[1/(I + B/\alpha)^{1/2}], \quad (11)$$

where I is the unity matrix and B is equal to one-half times the Hessian calculated in the entropy metric (see *e.g.* Gull, 1989; MacKay, 1992, and references therein). For a given model, the *evidence* will often be determined by the entropy, especially if a *prior* close to the final estimate is used.

It can be shown (Gull, 1989) that the *evidence* for the Lagrange multiplier α has a maximum when

$$-2\alpha S = N_g. \quad (12)$$

This position of the *evidence* maximum provides a value α_1. N_g is the number of parameters that can be determined from a given set of experimental data and from (12) follows then $N_g = N_{g,1}$ for α_1. This *number of good parameters* N_g (synonyms used are *real sampling points* and *degrees of freedom*) can be derived also from B using the size of the eigenvalues λ_j according to

$$N_g = \sum_j \lambda_j/(\lambda_j + \alpha). \quad (13)$$

By combination of (12) and (13), α_2 and $N_{g,2}$ can be calculated. For consistency, $\alpha_1 \simeq \alpha_2$ and $N_{g,1} \simeq N_{g,2}$ should hold.

N_g can be taken as a measure of the information content of the experiment. In reality, this number will be smaller than the ideal number of degrees of freedom $s_{\max}L/\pi$ provided by information theory.

2.3. *Analytical model priors*

Analytical expressions for some *model priors* will be discussed shortly. Primarily, the autocorrelation function $C(x)$ of the model is calculated and for the distance distribution (3) then holds.

We present the formula for rotational ellipsoids, Gaussian chains, and an explicit expression for two-phase spheres.

2.3.1. *Rotational ellipsoids.* The correlation function of rotational ellipsoids has been calculated by the Fourier transformation (2) using the intensity formula for ellipsoids (Guinier & Fournet, 1955),

$$I_{\rm ell}(s) = (1/4\pi)\int_0^{2\pi} {\rm d}\varphi \int_0^{\pi} \Phi^2(s, R_{\rm eff})\sin\vartheta\,{\rm d}\vartheta, \quad (14)$$

where Φ is the scattering amplitude of a sphere with the effective radius $R_{\rm eff}$.

With the substitutions $t = \cos\vartheta$, $u = \varphi/2\pi$, one obtains

$$I_{\rm ell}(s) = \int_0^1 {\rm d}u \int_0^1 \Phi^2(s, R_{\rm eff})\,{\rm d}t \quad (15)$$

and with

$$R_{\rm eff}^2 = (a^2\cos^2 2\pi u + b^2\sin^2 2\pi u)(1 - t^2) + c^2t^2, \quad (16)$$

where a, b and c are half-axes of the ellipsoid, one obtains

$$C(x) = (2abc/3\pi)\int_0^1 {\rm d}u \int_0^1 {\rm d}t \int_0^{\infty} s\Phi^2(s, R_{\rm eff})[\sin(sx)/x]\,{\rm d}s. \quad (17)$$

The inner integral corresponds to the definition of the correlation function in (2) and can be replaced by the analytical expression for a sphere:

$$C(x) = \begin{cases} 1 - (3x/4R_{\rm eff}) + (x^3/16R_{\rm eff}^3) & x \le 2R_{\rm eff} \\ 0 & x > 2R_{\rm eff} \end{cases}. \quad (18)$$

Constraints can be derived from (16) and (18) for the angular regions permitted for φ and ϑ, because the condition $x \le 2R_{\rm eff}$ has to be fulfilled.

The result is, for prolate ellipsoids with the axes A, A and $Av (v \ge 1)$,

$$C(x) = \begin{cases} C_1(x) & 0 \le x \le A \\ C_1(x) - C_2(x) & A \le x \le vA, \\ 0 & x \ge vA \end{cases} \quad (19)$$

with

$$\begin{aligned} C_1(x) = {}& \tfrac{1}{2}\{1/4v^3 + 3/8v + 3v/8(|v^2 - 1|)^{1/2} \\ & \times {\rm ATA}\,[(|v^2 - 1|)^{1/2}]\}(x^3/A^3) \\ & - \tfrac{3}{2}\{1/2v + v/2(|v^2 - 1|)^{1/2} \\ & \times {\rm ATA}\,[(|v^2 - 1|)^{1/2}]\}(x/A) + 1 \end{aligned} \quad (20)$$

and

$$\begin{aligned} C_2(x) = {}& \tfrac{3}{16}[v/(|v^2 - 1|)^{1/2}]\{(|x^2/A^2 - 1|)^{1/2} \\ & \times (2A/x + x/A) + (x^3/A^3 - 4x/A) \\ & \times {\rm ATA}\,[(|x^2/A^2 - 1|)^{1/2}]\}. \end{aligned} \quad (21)$$

For prolate rotational ellipsoids, ATA equals arctan. For oblate ellipsoids with the axes A, A and $vA (v \le 1)$, ATA is the hyperbolic area tangent function and

$$C(x) = \begin{cases} C_1(x) & 0 \le x \le vA \\ C_2(x) & vA \le x \le A. \\ 0 & x \ge A \end{cases} \quad (22)$$

The volume is

$$V_c = (\pi/6)A^3 v. \tag{23}$$

2.3.2. *Gaussian chains.* The correlation function of a Gaussian chain can be calculated by insertion of the analytical expression for the scattered intensity (Debye, 1947)

$$I(s) = 2[\exp(-s^2\bar{R}_G^2) + s^2\bar{R}_G^2 - 1]/s^4\bar{R}_G^4 \tag{24}$$

in (2) to give

$$\begin{aligned} C(x) = {} & (1/4\pi x^2)(x/\bar{R}_G\pi^2)\{(\pi/2)(1 + x^2/2\bar{R}_G^2) \\ & \times [1 - \hat{\Phi}(x/2\bar{R}_G)] - (x\pi^{1/2}/2\bar{R}_G) \\ & \times \exp(-x^2)/(4\bar{R}_G^2)\}. \end{aligned} \tag{25}$$

$\hat{\Phi}(x/\bar{R}_G)$ is the error integral

$$\hat{\Phi}(x/\bar{R}_G) = (2/\pi^{1/2}) \int\limits_0^{x/\bar{R}_G} \exp(-t^2)\,\mathrm{d}t \tag{26}$$

and the radius of gyration $\bar{R}_G$ is the mean value $\langle R_G^2\rangle^{1/2}$ for the whole population of chain molecules.

2.3.3. *Two-phase spheres.* The correlation function of two spheres concentrically arranged with the inner radius $R_{\rm in}$ and the outer $R_{\rm out}$, with excess electron or scattering length densities $\rho_{\rm in}$ and $\rho_{\rm out}$, has been calculated by the method described by Guinier & Fournet (1955). Because two-phase spheres are suitable models for micellar systems in pharmaceutics and molecular biology, we provide the special three-parametric formula for use in maximum-entropy algorithms. The result is implicitly described also by the overlap integrals for spherical symmetry given by Glatter (1981). The resulting terms are

$$C(x) = \begin{cases} 0 & x \geq 2R \\ (1/y)\{[1 - (3x/4R) + (x^3/16R^3)] & 0 \leq x \leq 2R \\ +[(1-\rho)^2\{k^3 - (3k^2x/4R) + (x^3/16R^3)\}] & 0 \leq x \leq 2Rk \\ +[(1-\rho)(-1 - k^3 + \{3x(k^2-1)/4R\} + (3R/8x)(k^2-1)^2 - (x^3/8R^3))] & R - kR \leq x \leq R + kR \\ -[2k^3(1-\rho)]\} & 0 \leq x \leq R - kR \end{cases} \tag{27}$$

with $\rho = \rho_{\rm in}/\rho_{\rm out}$, $k = R_{\rm in}/R_{\rm out}$, $R = R_{\rm out}$ and $y = 1 + k^3(\rho^2 - 1)$.

The terms in square brackets have to be taken into account for the corresponding x regions. A hollow sphere and a sphere are special cases with $\rho = 0$ and $(k = 0, \rho = 1)$, respectively. The correlation volume is

$$V_c = (4\pi R^3/3)[1.0 + (\rho - 1)Rk^3]^2/[1.0 + Rk^3(\rho^2 - 1)]. \tag{28}$$

3. Results

Here, we discuss the shape modelling with real-space data for an ellipsoidal *model prior* only, because of equivalent results when using structure-adequate *model priors*. As a test example, we have chosen the molecule cytochrome *c*. The small- and wide-angle X-ray scattering curve and the inertia-equivalent ellipsoid (IEE) of the molecule have been calculated from the atomic coordinate set 1cyc (Tanaka, Yamane, Tsukihara, Ashida & Kakudo, 1975) stored in the Brookhaven Protein Data Bank (PDB; Bernstein *et al.*, 1977) by using the improved cube algorithm described recently (Müller, Gernat, Schulz, Müller, Vorwerg & Damaschun, 1994). A statistical noise of 5% has been added to the solution scattering to simulate experimental conditions (Fig. 1). The determination of an inertial ellipsoid and of the related inertia-equivalent ellipsoid of a body is done by methods of classical mechanics (Sommerfeld, 1962). The numeric expressions have been described recently (Müller & Schrauber, 1992). The IEE of the solvent-excluded body of the molecule used for checks of the *MAXENT* results contains 90% of the molecular volume and represents the shape of a globular protein very well. The high-resolution distance-distribution function $p(r)$ of the molecule (Fig. 2) has been calculated by a modified direct Fourier transformation of the theoretical noiseless small- and wide-angle scattering curve avoiding termination errors (Müller & Hansen, 1994). This theoretical $p(r)$ function is set to be the correct distance distribution for the molecule, because a direct calculation of $p(r)$ from the atomic coordinates is not possible for a macromolecule in solution at present. The only preknowledge of the molecular structure used during the modelling procedure was that cytochrome *c* is a globular compact molecule and, therefore, an ellipsoid could be an adequate homogeneous body for modelling its shape.

3.1. *Modelling by MAXENT*

The theoretical distance distribution of the cytochrome *c* molecule and the *MAXENT* result are drawn for comparison in Fig. 2. No significant systematic differences can be detected for distances larger than 0.4 nm, the recent resolution limit of the modified Fourier transformation method used (Müller & Hansen, 1994). That is a proof of the used *model prior* type being structure-adequate. The scattering curve obtained by *MAXENT* fits the input data within the statistical error level in the small-angle ($s \simeq 6$ nm^{-1}) as well as in the wide-angle [$6 \leq s$ (nm^{-1}) ≤ 30] region. During the maximum-entropy procedure for determination of the high-resolution distance-distribution function from the scattering curve, the optimum geometrical parameters of the *model prior* are calculated by maximization of the '*evidence*' in (11). By this, the most likely parameters for the *model prior* are determined while the data are fitted simultaneously.

For the calculation of the *evidence* for the dimensions of the ellipsoid, the Bayesian criteria (12) has been used for selection of α (Fig. 3). A value of 0.000095 is determined. Then, the axes of the ellipsoid can be calculated from their *evidences* (Fig. 4) without any subjective decision. For a prolate ellipsoid, the axes are $A = B = 2.98$ nm, $C = 3.79$ nm. The slight deviation of the dimensions from the positions of the maxima is due to a symmetrical Gaussian fit to the *evidence* curves and choice of the maximum positions of the Gaussians. The variations of N_g as a function of α calculated according to (13) and $-2\alpha S$ are shown in Fig. 5. From the cross point of both curves follows the most likely value for $\alpha = 0.000095$ and a number of good parameters $N_g = 32$. This value of α is too small in comparison with the value selected by the conventional 'classic' maximum-entropy method requiring that $\chi^2 = M$. By this criterion, the maximum-entropy method leads to a value of 0.0009 for α (Fig. 3). The reason for the discrepancy is that the derivation of α has been done subject to the assumption of *one* particular *model prior* only. Also, when smoothness (instead of entropy) is used as the regularizer, the Bayesian estimation of the Lagrange multiplier of the regularizing term tends to overfit the data. This has been shown recently by Archer & Titterington (1995). For the calculation of the *evidences* for the diameters of the prior, it is clear from Fig. 4 that variances for the *model prior* should be taken into account as well. The use of error estimates for the *prior*, as suggested *e.g.* by Hansen & Wilkins (1994) and depicted in Fig. 4, will increase the value for α and influence the estimate of $p(r)$. This is significant especially at large r values, where the variances of the axes for the ellipsoidal *model prior* will lead to large error estimates for m as shown in Fig. 2. Similarly, the calculation of the *evidence* for the axes of the *model prior* has also been done subject to one particular value of α, but a variation of α leads to only relatively small changes in the estimated $p(r)$, which is why the estimation of the axes by the *evidence* is only affected to a minor degree by a variation of α. To avoid overfitting of the data, for the further calculations the larger value $\alpha = 0.0009$ has been chosen; as a consequence, the number of free parameters is reduced to $N_g = 24$ in the scattering region $s < 30$ nm^{-1}. This is a considerably lower number than 38, determined for the noiseless data by the sampling-point theorem of information theory. To a small extent, the number of free parameters is

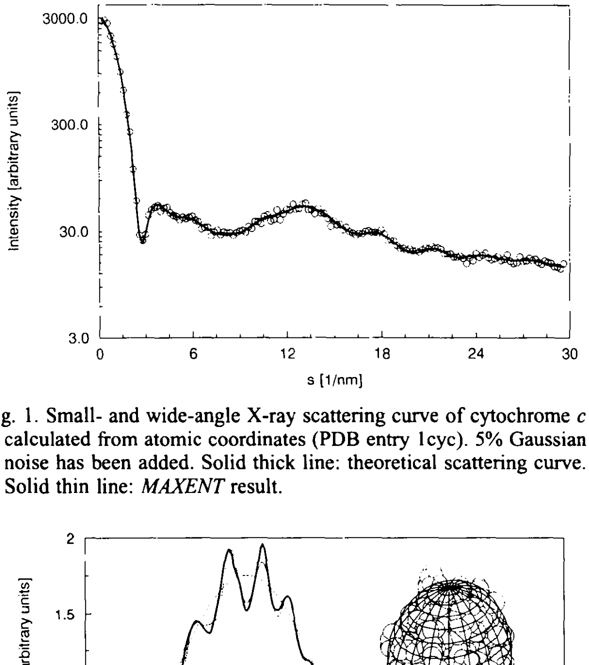

Fig. 1. Small- and wide-angle X-ray scattering curve of cytochrome *c* calculated from atomic coordinates (PDB entry 1cyc). 5% Gaussian noise has been added. Solid thick line: theoretical scattering curve. Solid thin line: *MAXENT* result.

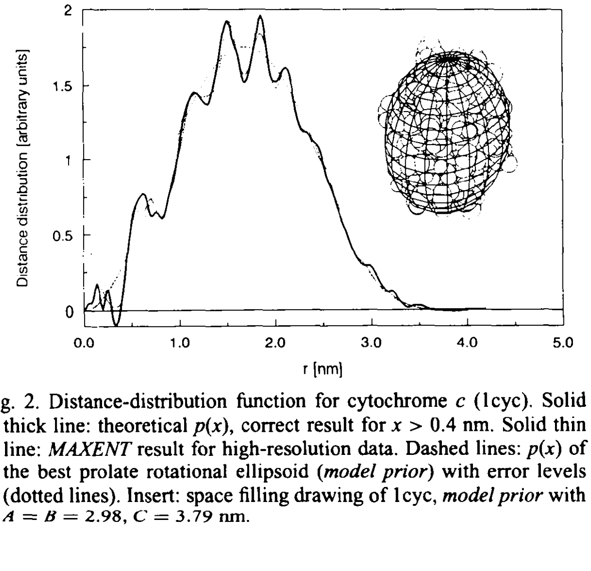

Fig. 2. Distance-distribution function for cytochrome *c* (1cyc). Solid thick line: theoretical $p(x)$, correct result for $x > 0.4$ nm. Solid thin line: *MAXENT* result for high-resolution data. Dashed lines: $p(x)$ of the best prolate rotational ellipsoid (*model prior*) with error levels (dotted lines). Insert: space filling drawing of 1cyc, *model prior* with $A = B = 2.98$, $C = 3.79$ nm.

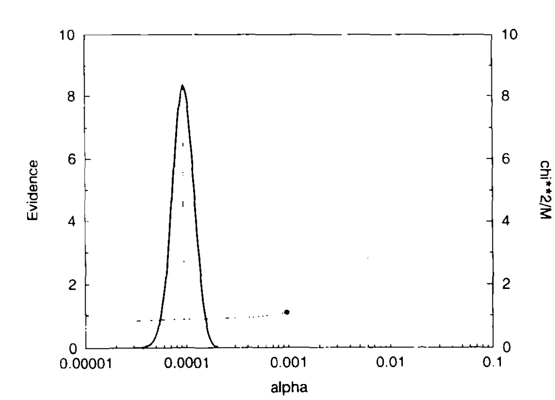

Fig. 3. '*Evidence*' for Lagrange parameter α and quality of the fit χ^2/M as function of α. Thick solid line: *evidence*. Thin solid line: χ^2/M. The dot marks the α value usually chosen 'by eye'.

dependent upon the model used through the value of the Lagrange multiplier α. However, this is consistent with the usual measure of information being taken relative to a *prior estimate*.

Little or no effect of the overlaid statistical noise percentage (1–10%) and of slit-length smearing [Gaussian slit-length profile $P(t) = \exp(-c^2t^2), c = 0.25$ nm; Müller & Hansen, 1994] has been detected (Table 1). All estimated parameters agree within the error limits.

The whole X-ray scattering curve for $s < 30\,\mathrm{nm}^{-1}$ has been included in the estimation of the *model-prior* parameters but, of course, no method for analysis can replace a contrast-variation experiment by which the shape scattering curve of the molecule can be separated from scattering contributions of inner electron or scattering-length fluctuations and scattering-interaction terms (*e.g.* Stuhrmann & Kirste, 1965). Similarly to the scattering curve, the distance-distribution function has contributions from the homogeneously filled solvent-excluded molecule body as well as from the fluctuation and interaction terms. These contributions have been calculated by the program *ICM* (Müller, 1983) for cytochrome *c* and are depicted in Fig. 6. When modelling the shape by ellipsoids or other homogeneous models, the long-periodical interaction term (shaded region in Fig. 6) will be included in the *prior estimate* and will falsify the dimensions of the axes. For cytochrome *c*, the $p(r)$ of the inertia-equivalent ellipsoid of its solvent-excluded volume body fits completely into the error band of the *model prior* (Fig. 7), but owing to the interaction term the *model-prior* electron distances are systematically shifted to somewhat larger values. The major axis of the IEE agrees with the axis of the prolate *model prior* and the minor axes $A = 2.62$ and $B = 3.26$ nm differ by about 10% from the corresponding *model-prior* axes. The major axis is a lower-limit estimate of the largest diameter of the molecule, which has been calculated to 4.04 nm from the atomic coordinates. An oblate ellipsoid cannot be excluded as *model prior* by the entropy criterion or from the *evidence* curves; on the contrary, the distance-distribution (Fig. 7) and the scattering curve (Fig. 8) agree better with the IEE data than those of the prolate ellipsoid. The minor axis $A = 2.65$ nm agrees with the minor axis of the IEE and the other differs by about 10% (Table 1). In this case, the major axis of the *model prior* is 12% smaller than the largest diameter of the molecule. Possibly, the similarity between the structure-related inertia-equivalent ellipsoid and *model prior* could be improved when an analytical expression for the $p(r)$ of triaxial ellipsoids is used that is unknown at present. On the other hand, the information content of the zero-angle maximum in the scattering curve that contains the shape information is very low. *MAXENT* predicts for cytochrome *c* $N_g = 2$ real sampling points for $s_{\max} < 2.4\,\mathrm{nm}^{-1}$ and 5% noise (instead of three ideal sampling points). That is the reason for the failure of Marquardt's (1963) curve-fitting procedure when used for fitting the scattering curves of triaxial ellipsoids to the data in reciprocal space.

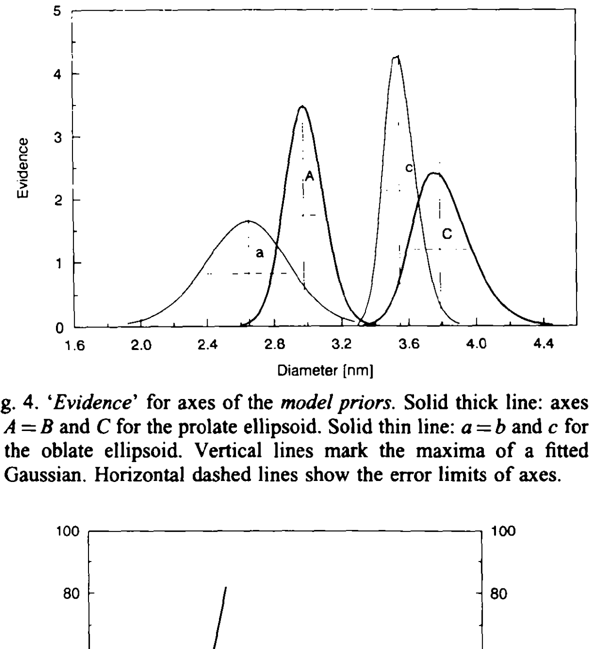

Fig. 4. '*Evidence*' for axes of the *model priors*. Solid thick line: axes $A = B$ and C for the prolate ellipsoid. Solid thin line: $a = b$ and c for the oblate ellipsoid. Vertical lines mark the maxima of a fitted Gaussian. Horizontal dashed lines show the error limits of axes.

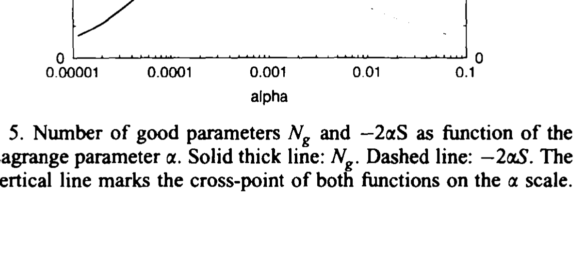

Fig. 5. Number of good parameters N_g and $-2\alpha S$ as function of the Lagrange parameter α. Solid thick line: N_g. Dashed line: $-2\alpha S$. The vertical line marks the cross-point of both functions on the α scale.

The number of real sampling points has been estimated for different experimental conditions also (Table 1). Whereas slit-smearing effects hardly influence the number, the noise reduces the information content drastically. Seemingly, the number of good parameters is increased by 8 when overfitting the curve by using $\alpha = 0.000095$. But, then the noise is partially included in the structure information.

3.2. *Comparison with curve-fitting results*

The parameter estimation from real-space data can be done also by any curve-fitting procedure separately from slit-correction or transformation procedures and without any additional regularization. As a precondition, the experimental $p(r)$ function, reliable error limits and an analytical or semianalytical model $p(r)$ have to be known. Here, a modification of Marquardt's nonlinear least-squares algorithm (1963) is used to check the

Table 1. *Axes of ellipsoidal shape models and number of good parameters determined for cytochrome c (Lagrange parameter $\alpha = 0.0009$)*

Type of ellipsoid	Collimation	Noise (%)	A (nm)	B (nm)	C (nm)	N_g
Prolate *model prior*	Pinhole	1	3.00 (10)	3.00 (10)	3.74 (11)	30.2
Prolate *model prior*	Pinhole	5	2.98 (10)	2.98 (10)	3.79 (16)	24.4
Prolate, $p(r)$ fit*	Pinhole		3.02 (6)	3.02 (6)	3.82 (12)	
Prolate *model prior*	Pinhole	10	2.99 (13)	2.99 (13)	3.82 (21)	20.9
Prolate *model prior*	Slit-length smeared	5	2.98 (13)	2.98 (13)	3.81 (18)	23.6
Prolate *model prior*	Slit-length smeared	10	2.97 (16)	2.97 (16)	3.87 (28)	20.0
Oblate *model prior*	Pinhole	5	2.65 (24)	3.55 (9)	3.55 (9)	
Oblate, $p(r)$ fit*	Pinhole		2.84 (12)	3.50 (6)	3.50 (6)	
IEE†			2.62	3.26	3.82	

* Calculated using Marquardt's (1963) curve-fitting program in real space. † Inertia-equivalent ellipsoid of the solvent-excluded molecular body (Müller & Schrauber, 1992).

MAXENT results. The best-fitting prolate and oblate rotational ellipsoids have been calculated from the noise-free highly resolved $p(r)$. With exclusion of the innermost erroneous region for $r < 0.4$ nm (Müller, Damaschun & Schrauber, 1990) and the outer region $r > 3.3$ nm, and the choice of an arbitrary value of 1% relative error over the whole distance distribution, the axes of the ellipsoids have been calculated (Table 1). Oblate and prolate ellipsoids cannot be discriminated (insert in Fig. 7). This is also understandable from the small differences between both scattering curves in the region $s < 5\ \text{nm}^{-1}$ (not shown here). The dimensions received by *MAXENT* from noisy scattering data agree very well with the values determined by curve fitting of the noiseless $p(r)$ function. The deviations are due to the noise and the regularization by the maximum-entropy condition, especially for the oblate ellipsoid.

4. Concluding remarks

The resolution of a shape model determined from small-angle X-ray scattering is in general restricted to about 2 nm when homogeneous model bodies and no contrast variation are used. The results of the shape modelling discussed above have to be appreciated with this fact taken into account. The maximum-entropy method renders possible the determination of a real number of parameters to be estimated from experimental data and can determine the parameters and their error limits if an analytical or semianalytical expression for the *model prior* are given. For rotational ellipsoids, Gaussian chains and two-phase spheres, the analytical model priors are presented to enlarge the family of available models. The relevance of the structure information extracted from the estimated parameters of the *model prior* depends on the chosen type of model, which means that some knowledge of the structure should be available from other methods. For the globular molecule cytochrome *c*, the automatically determined ellipsoids of revolution, both oblate and prolate, agree with the direct

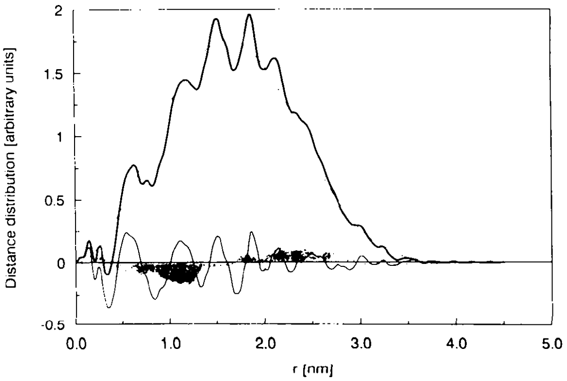

Fig. 6. Theoretical distance distribution of cytochrome *c*. Solid thick line: complete molecule. Dashed line: homogeneous body (solvent-excluded volume). Solid thin line: electron inhomogeneities. Shaded area: interaction vectors between homogeneous body and inhomogeneities. The data are calculated by the improved cube method (Müller, 1983).

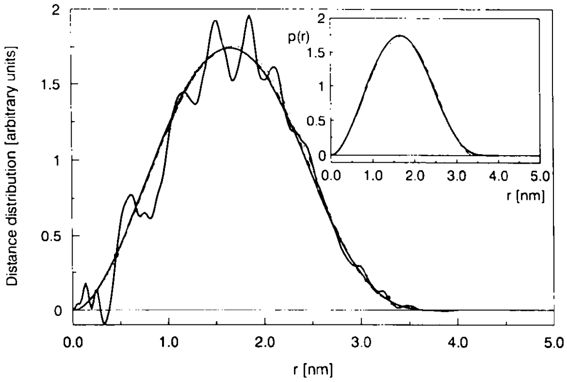

Fig. 7. Distance-distribution function of cytochrome *c* and of ellipsoidal shape models. Solid thick line: cytochrome *c*. Solid smooth thick line: inertia-equivalent ellipsoid with the axes $A = 2.62$, $B = 3.26$, $C = 3.82$ nm. Dashed line: prolate *model prior* $A = B = 2.98$, $C = 3.79$ nm with error levels (dotted lines). Dashed-dotted line: oblate *model prior* $A = B = 3.55$, $C = 2.65$ nm. Dots mark the fitting region used for Marquard's (1963) routine. The inset shows the result of the curve fitting in real space: Solid line: inertia-equivalent ellipsoid. Dashed line: prolate ellipsoid $A = B = 3.02$, $C = 3.82$ nm. Dashed-dotted line: oblate ellipsoid $A = B = 3.50$, $C = 2.84$ nm.

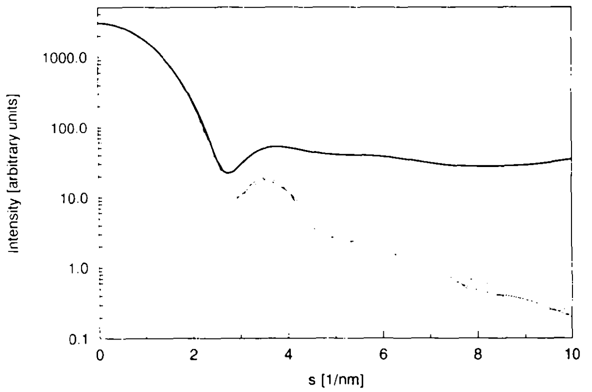

Fig. 8. X-ray scattering curves of cytochrome *c* and of ellipsoidal shape models. Solid thick line: cytochrome *c*. Solid thin line: inertia-equivalent ellipsoid with the axes $A = 2.62$, $B = 3.26$, $C = 3.82$ nm. Dashed line: prolate *model prior* $A = B = 2.98$, $C = 3.79$ nm. Dashed-dotted line: oblate *model prior* $A = B = 3.55$, $C = 2.65$ nm.

structure-related inertia-equivalent ellipsoid within the expected goodness reachable by models with two degrees of freedom. The estimation of a lower limit of the largest molecular diameter could be of interest for other indirect data-handling methods that need this value as a starting parameter. Further studies will be necessary concerning the discrepancy between the Lagrange multiplier α when estimated from *ad hoc* methods using the classical χ^2 or from the Bayesian criteria, but this is a problem that is independent of the regularization method used.

A Fortran77 version of the program *MAXENT* is available from SH.

The authors thank Udo Heinemann for helpful discussions and kind support.

References

Archer, G. & Titterington, D. M. (1995). *IEEE Trans. Image Processing*, **4**, 989–995.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.

Bricogne, G. (1993). *Acta Cryst.* D**49**, 37–60.

Damaschun, G., Müller, J. J. & Pürschel, H.-V. (1968). *Monatsh. Chem.* **99**, 2343–2348.

Debye, P. (1947). *J. Phys. Colloid. Chem.* **51**, 18–32.

Debye, P., Anderson, H. R. & Brumberger, H. (1957). *J. Appl. Phys.* **28**, 679–683.

Feigin, L. A. (1971). *Kristallografiya*, **16**, 711–714.

Feigin, L. A. & Svergun, D. I. (1987). *Structure Analysis by Small-Angle X-ray and Neutron Scattering.* New York: Plenum.

Gille, W. & Handschug, H. (1995). Computer Center, Universität Halle. Unpublished results.

Glatter, O. (1980). *Acta Phys. Austriaca*, **52**, 243–256.

Glatter, O. (1981). *J. Appl. Cryst.* **14**, 101–108.

Glatter, O. & Kratky, O. (1982). *Small-Angle X-ray Scattering.* London: Academic Press.

Goldman, S. (1954). *Information Theory.* New York: Prentice Hall.

Goodisman, J. (1980). *J. Appl. Cryst.* **13**, 132–134.

Guinier, A. & Fournet, G. (1955). *Small-Angle Scattering of X-rays.* New York: Wiley.

Gull, S. F. (1989). *Maximum-Entropy and Bayesian Methods*, edited by J. Skilling, pp. 53–71, Dordrecht: Kluwer Academic Publishers.

Hansen, S. & Wilkins, S. W. (1994). *Acta Cryst.* A**50**, 547–550.

MacKay, D. J. C. (1992). *Maximum-Entropy and Bayesian Methods*, edited by C. R. Smith, G. J. Erickson & P. O. Neudorfer, pp. 39–66. Dordrecht: Kluwer Academic Publishers.

Marquardt, D. W. (1963). *J. Soc. Ind. Appl. Math.* **2**, 431–469.

Müller, J. J. (1983). *J. Appl. Cryst.* **16**, 74–82.

Müller, J. J., Damaschun, G. & Hübner, G. (1979). *Acta Biol. Med. Ger.* **38**, 1–10.

Müller, J. J., Damaschun, G. & Schmidt, P. W. (1985). *J. Appl. Cryst.* **18**, 241–247.

Müller, J. J., Damaschun, G. & Schrauber, H. (1990). *J. Appl. Cryst.* **23**, 26–34.

Müller, J. J., Gernat, Ch., Schulz, W., Müller, E.-Ch., Vorwerg, W., & Damaschun, G. (1994). *Biopolymers*, **35**, 271–288.

Müller, J. J. & Hansen, S. (1994). *J. Appl. Cryst.* **27**, 257–270.

Müller, J. J. & Schrauber, H. (1992). *J. Appl. Cryst.* **25**, 181–191.

Sjöberg, B. (1977). *Eur. J. Biochem.* **81**, 277–283.

Skilling, J. (1988). *Maximum-Entropy and Bayesian Methods in Science and Engineering*, Vol. 1, edited by G. J. Erickson & C. Ray Smith, pp. 173–187. Dordrecht: Kluwer Academic Publishers.

Sommerfeld, A. (1962). *Mechanik.* Leipzig: Akademische Verlagsgesellschaft Geest & Portig K.-G.

Stuhrmann, H. B. (1970). *Z. Phys. Chem.* **72**, 177–184.

Stuhrmann, H. B. & Kirste, R. G. (1965). *Z. Phys. Chem.* **46**, 247–250.

Tanaka, N., Yamane, T., Tsukihara, T., Ashida, T. & Kakudo, M. (1975). *J. Biochem. (Tokyo)*, **77**, 147–162.