STRUCTURE INFORMATION RETRIEVAL FROM SOLUTION X-RAY AND NEUTRON SCATTERING EXPERIMENTS

Vittorio Luzzati

Centre de Génétique Moléculaire, Centre National de la Recherche Scientifique, 91190 Gif-sur-Yvette, France.

The angular distribution of the intensity of X-ray and neutron beams scattered by macromolecules in solution usually displays two distinct regions : one, at "small" angles, is specifically sensitive to the long-range organization, the other, at "high" angles, to the short-range structural features. Most often this distinction, which mirrors the contrast between the sharp distribution of the shortest interatomic distances (in the 1-5 Å range) and the even distribution of the longer distances, is reasonably clear-out. Besides, if the macromolecules are large, the intensity decreases very rapidly with increasing scattering angle. We are mainly concerned here with the information relevant to the long-range structure, and thus with the "small" angle scattering ; we show below (see eq. 4) how to take into account, at least to a first approximation, the "small" angle effects of the short interatomic distances.

A typical solution scattering experiment consists of recording the intensity scattered by the solution, subtracting the intensity scattered by the solvent and by the instruments (sample cell, slits, etc.) and correcting for instrumental distortions, mainly collimation and polychroma-

---

Notation. Terminology and notation will be those commonly used in X-ray scattering studies[1] ; the results can be extended easily to neutrons. r (in Å) and s (in $Å^{-1}$) ($s = 2\sin\theta/\lambda$, 2θ being the scattering angle, λ the wavelength) specify positions in real and reciprocal space. i(s) is the distribution of the scattered intensity ; si(s) = $\int_{-\infty}^{+\infty} rp(r)\sin 2\pi rs\, dr$, where p(r) is the autocorrelation function, namely the spherical average of the convolution $\{\rho(\underline{r})*\rho(-\underline{r})\}$, $\rho(\underline{r})$ being the electron density distribution.

tism. The result, usually expressed in a digit form, is a set of intensities associated to a number of channels (see fig. 1). It is clear that the intensities recorded at the different channels may well not be statistically independent, especially when the number of channels is large ; it is worth while to note, in this respect, that some of the most common operations performed on the intensity curves - smoothing, interpolations, extrapolations - are based upon the very presence of such correlations. The nature of these correlations is closely linked to the information content of the data ; information content and data analysis are the themes of this paper [2, 3].

It is hardly possible to tackle these problems without making some assumption on the structure of the sample. The following conditions define the framework of our treatment [1, 2, 3]:

- the sample is an ideal solution of discrete particles ;
- the particles are all identical ;
- the particles are globular, or more precisely none of their dimensions is large with respect to $(s_{min})^{-1}$, s_{min} being the lower limit of the interval of s explored experimentally.

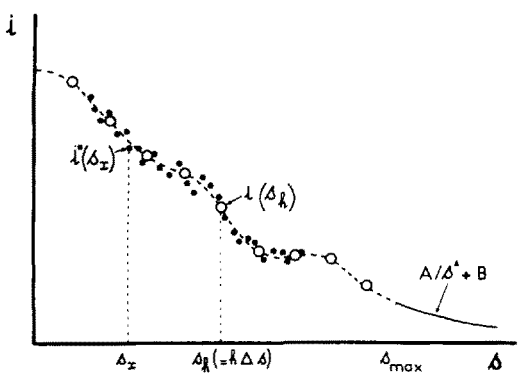

Fig. 1

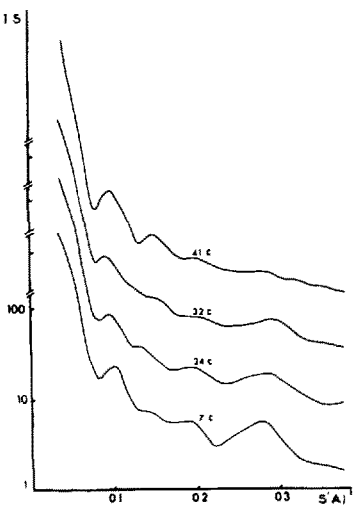

Fig. 2

Fig. 1 - Schematic representation of one intensity curve. The dotted line represents the correct curve, the small full dots the experimental points, the large open dots the intensities at the points of the lattice $s_h = h\Delta s$, the full line the asymptotic trend.

Fig. 2 - A few experimental curves, corrected for experimental distortions, obtained with simian low density serum lipoprotein, as a function of temperature [3].

Ideality is usually met by extrapolation to infinite dilution : monodispersity is a convenient simplification although the conclusions remain formally correct if the sample is heterogeneous. Globular shape is a strict requirement ; yet some of the results can be extended to rod-like and lamellar particles.

Within the framework of these hypotheses, the curvature at the origin of the intensity curve is porportional to one structural parameter, the radius of gyration R [4] :

$$i(s) = i(0)\{1 - (4/3)\pi^2 R^2 s^2 + \ldots.\} \qquad (1)$$

If, moreover, a few additional parameters are known - absolute scale, concentration, partial specific volume - the intensity at the origin yields the molecular weight M. Therefore, extrapolation to the origin - which involves some assumptions on the regularity of i(s) - can yield the values of R and M. Besides, it can often be assumed that the electron density inside the particle is fairly uniform ; in this case the asymptotic trend of i(s) takes the form [1, 5] :

$$\lim_{s\to\infty} s^4 i(s) \propto S \qquad (2)$$

where S is the area of the outer surface of the particle. Equation 2 can be tested against the data : if it is fulfilled the value of S can be determined. In this case it is also possible to determine the volume V of the particle [4]:

$$V = i(0) \Big/ 2\pi \int_0^\infty s^2 i(s) ds \qquad (3)$$

These four parameters - M, R, S, V - constitute the bulk of the information traditionnally retrieved from solution scattering experiments[4]; yet it is obvious, even upon visual inspection, that the experimental curves sometimes contain more information than that (see fig. 2).

It is particularly easy to discuss the problem of the information content when the particles are of globular shape. In this case the autocorrelation function vanishes beyond the maximal diameter of the particle, D_{max}. Since the function si(s) is the Fourier transform of rp(r) (see Notation), if $p(r) \equiv 0$ for $r > D_{max}$, then i(s) is completely defined by its values at the lattice points $h\Delta s$, with $\Delta s \leq (2D_{max})^{-1}$ (sampling theorem). Since the number of lattice points is infinite, a finite set of data is insufficient to define completely the function i(s). This problem can be circumvented if some assumption can be made about the mathematical form of i(s) at large s. For systems of biological interest (proteins, nucleic acids, lipids, etc.) the function

$$i(s) = A/s^4 + B \qquad (4)$$
$$s > s_{max}$$

has been shown to provide a good empirical description of the asymptotic trend [1, 6]. Equation 4 can also be justified from a structural viewpoint ; the term A corresponds to the average long-range electron density distribution, the term B to the short-range fluctuations around the average. If eq. 4 is adopted, then the entire function i(s), from s = 0 to infinity, is defined by the intensity at the lattice points of the interval $0<s<s_{max}$ plus the values of A and B. The number of these parameters is :

$$J = s_{max}\ 2D_{max} + 2 \qquad (5)$$

Thus, we reach the conclusions :

- within the restrictions specified above - ideal solution of identical globular particles - and with the assumption that the high angle form of the intensity is known, the whole of the experimental information provided by one solution scattering experiment is ideally equivalent to a finite number J of independent measurements ;
- the maximal structural information which can be retrieved from that experiment is expressed by the same number J of independent parameters.

The explicit expression for the intensity at any point s, in terms of the intensity at the lattice points in the interval $0<s<s_{max}$ and of the parameters A and B, is :

$$si(s) = \sum_{h=1}^{h_{max}} h\Delta s\ i(h\Delta s)\Phi(s,h) + A \sum_{h_{max}+1}^{\infty} (h\Delta s)^{-3}\Phi(s,h) + B \sum_{h_{max}+1}^{\infty} (h\Delta s)\Phi(s,h) \qquad (6)$$

where :

$$\Phi(s,h) = \frac{\sin\pi(s/\Delta s-h)}{\pi(s/\Delta s-h)} - \frac{\sin\pi(s/\Delta s+h)}{\pi(s/\Delta s+h)} \qquad (7)$$

Each experimental value of the intensity ($i^*(s_x)$ in fig. 1) is equivalent to one equation (6) ; if the number of experimental points is equal to, or larger than, the number of unknowns J, then the system can be solved. The function i(s) is completely defined by the values of those J parameters.

We can make a few comments :

a) - Such analysis of the information content is to some extent idealized, and in any event incomplete, if the limitations in the accuracy of the data are ignored. The treatment above can be completed by a statistical analysis of the propagation of the experimental errors.

b) - The mathematical treatment is heavily dependent upon the properties of the functions p(r) and i(s) when r and s are large. The assumption that p(r) vanishes beyond D_{max} is a logical consequence of the globular shape of the particles. The hypothesis that the asymptotic trend is a universal property of the intensity curves is more questionable, since,at all values of s,i(s) depends upon the precise structure of the sample. It must be stressed that the postulated asymptotic form is meant to be an empirical approximation, acceptable within the limits of the experimental errors ; besides, the errors due to this approximation can be estimated.

c) - Equations 6 is not very sensitive to the precise value of Δs (and thus of D_{max}), provided the value chosen is not too large. One possible way of choosing D_{max} is to solve the system of equations 6 for different values of D_{max}, using for example a least squares algorithm. As D_{max} increases the residual can be expected to drop first, and then to level off : the breaking point defines the minimum value of D_{max}. Other experiments (for example electron microscopy) can provide independent information on this point.

d) - The solution of the system of equations 6 does not require data recorded at a regular interval of s, or even obtained in one single experiment. In fact the algorithm is well suited for matching together experiments performed under different conditions, for example different sample-detector distances, different wavelengths.

e) - The raw intensities, before corrections for polychromatism and collimation distortions, are related to the same J parameters by equations similar to eq. 6, in which $\Phi(s,h)$ (see eq. 7) is replaced by an operator which takes into account the experimental distortions. These equations provide a convenient algorithm for the correction of polychromatism and collimation distortions.

f) - The treatment we sketch here provides a rational approach to operations like smoothing, interpolations, extrapolations, and also to the determination of structural parameters (for example M, R, V, S, see above). Indeed, these operations can be expressed in terms of the J experimental parameters ; these expressions are more accurate than those based upon local properties of the data (for example Guinier's plots[4].

g) - It may be wondered why worry, here and now, about these formal problems. The reason must be sought in the recently revived interest in some theoretical aspects of the solution scattering phenomenon prompted by a variety of technical developments. With regard to X-rays, first the

introduction of position sensitive detectors[7], later the use of synchrotron radiation[8] has had the effect, over the span of a few years, to shorten exposure time by a factor larger that 10^5. At almost the same time, the use of high flux reactors and of position sensitive detectors[9] have transformed neutron scattering into a powerful tool for structural studies of macromolecules in solution.

h) - It is common practice, in solution scattering studies, to perform experiments at variable solvent density; this is achieved by adding electron dense compounds (salts, sucrose, ..) in the case of X-rays and by varying the D_2O/H_2O ratio in the case of neutrons. These experiments are usually interpreted within the framework of the invariant volume hypothesis : a volume can be associated to each macromolecule in solution, inside of which the electron density distribution is independent of the density of the solvent. In this case,[1,3] the intensity scattered at any solvent density is a linear combination of three functions called the characteristic functions. In other words, the information which can be retrieved from any number of experiments performed under these conditions is contained in three independent intensity curves. Therefore the information content is equal to 3J[3].

i) - We can illustrate the previous results with a few examples. The first is an ideal protein, of spherical shape, which we assume to be 30% hydrated, and whose partial specific volume is 0.74 cm^3g^{-1}. If M is its molecular weight, its diameter is equal to 1.78 $M^{1/3}$ and $J = 3.56\ M^{1/3}s_{max}+2$. Assuming that $s_{max} = (25\ A)^{-1}$ one can expect, at best, to retrieve the value of five structural parameters - for example M, R, V, S, B - when $M \gtrsim 10{,}000$ Daltons. The information content increases (rather slowly, in fact) with M, and is also greater for particles which are anisometric (D_{max} increases in this case). Another example is low density serum lipoproteins[3]. A few intensity curves are shown in fig. 2. For these particles D_{max} and s_{max} are close to 300 Å and to $(25\ Å)^{-1}$ respectively. Therefore for one intensity curve the number of independent parameters is J = 26. Since this system has been studied systematically as a function of variable solvent density, and the invariant volume hypothesis has been shown to be fulfilled, the total information content is equivalent to 3J = 78 parameters. Such a wealth of information is quite unexpected for solution scattering studies : indeed in this case the experiments have been analyzed in terms of an elaborate model[3].

References

(1) - Luzzati, V., Tardieu, A., Mateu, L. & Stuhrmann, H.B. (1976), J. Mol. Biol., 101, 115-127

(2) - Luzzati, V. (1979), Ann. Rev. Biophys. Bioengin., 9, in press

(3) - Luzzati, V., Tardieu, A. & Aggerbeck, L.P. (1979) J. Mol. Biol., in press

(4) - Guinier, A. & Fournet, G. (1955). Small-angle scattering of X-rays, Wiley, New York

(5) - Porod, G. (1951), Kolloidzschr. 124, 83-114.

(6) - Luzzati, V., Witz, J. & Nicolaïeff, A. (1961), J. Mol. Biol., 3, 367-378

(7) - Gabriel, A. & Dupont, Y. (1972), Rev. Scient. Instr. 43, 1600-1603

(8) - Stuhrmann, H.B. (1978), Quart. Rev. Biophys., 11, 71-98

(9) - Ibel, K. (1976), J. Appl. Cryst., 9, 296-309