Accuracy and Resolution in Protein Crystallography: a Probabilistic Approach

BY VITTORIO LUZZATI
Centre de Génétique Moléculaire, Centre National de la Recherche Scientifique, 91190 Gif-sur-Yvette, France

AND DANIEL TAUPIN
Laboratoire de Physique des Solides, Centre Universitaire, 91405 Orsay, France

(Received 1 December 1983; accepted 5 March 1984)

Abstract
The purpose of this paper is to establish an explicit correlation between the values and probable deviations of the observed intensities and the potential accuracy of the structure eventually determined using those intensities. This problem involves several steps: the choice in real space of a reference stochastic model defining the ideal state of ignorance; the determination in reciprocal space of the probability density corresponding to that model and of the probability law corresponding to the experiment; the determination of the information; the transfer of the information from reciprocal to real space; the interpretation of the information in terms of structural resolution. A rigorous treatment of this problem is hindered by the inadequate knowledge of the statistical correlations between different reflections; more realistically, it is convenient to address a simpler problem, corresponding to an ideal experiment in which the intensities relevant to each reflection are measured independently of all the other reflections. In this case the mathematical problems can easily be solved and a parameter introduced — called projection information — whose value can be determined. This information is the sum of two terms, one associated with the modulus, the other with the phase of each reflection. By resorting to a mathematical model it is possible to make use of the projection information to determine the value of a parameter akin to structure resolution. The final result of this work is an operational definition of resolution, based upon the stochastic properties of the experimental observations. This resolution can be used to assess and compare the intrinsic quality of different structure analyses and of different stages of one structure determination, before the structure is solved.

1. Introduction
The problem of estimating the potential resolution of structure determinations and of evaluating the effective accuracy of the results is by no means trivial in protein crystallography. The various R factors and 'figures of merit' used for that purpose are most often empirical parameters of limited significance. A more satisfactory procedure, sometimes applied in the latest stages of structure analyses, is to let a refinement algorithm, whose main scope is to optimize the structure parameters with respect to the experimental observations, provide some statistical criterion of accuracy. All these methods come into play late in the process of structure determination; it would be highly desirable to be able to assess the potential accuracy of the eventual results before embarking on that venture.

From a mathematical viewpoint, the problem of solving a structure is equivalent to establishing a relationship between the probability law of the structure parameters and that of the experimental observations. If the problem is soluble then a correlation must exist between the stochastic structure of the experimental observations (i.e. their value and probable deviation) and the potential accuracy of the structure determination. This very correlation is the theme of this paper. More precisely, we seek some criterion to evaluate the quality of the experimental observations and to estimate the corresponding accuracy of the structure parameters. Ideally, the criterion should involve some parameter defined by the experimental observation and expressed in quantitative and absolute terms. This parameter would provide a test of the intrinsic quality of any structure determination at any stage of progress.

This paper involves the following logical steps.
(a) We first deal with the general problem of determining the amount of information associated with an experimental observation, given the prior state of ignorance. We also show that under proper conditions this information can be transferred from reciprocal to real space.
(b) We then show that the inadequate knowledge of the statistical correlations between different reflections sets a serious limitation on the practical interest of these general results.
(c) In order to circumvent the last problem we resort to the following procedure. We introduce in real space...
a virtual experiment dependent upon one stochastic parameter akin to structural resolution, and determine in reciprocal space the corresponding probability law for the structure factors. We then strive to determine the parameter defining the virtual experiment by comparing this law and the experimental probability law for the observation.

(d) In order to compare the two laws – this is by no means a trivial operation – we resort to the notion of projection information, namely the information associated with one reflection in an ideal experiment in which only that reflection is observed. We thus define and determine a parameter expressing structural resolution.

(e) The results of the crystallographic analysis of the structure of methionyl tRNA synthetase – a protein under current investigation in our laboratory (Zelwer, Risler & Brunie, 1982) – are used as an illustration.

II. Mathematical treatment

Notation

\( r \) and \( s \): vectors specifying position in real and reciprocal space. \( s = 2 \sin \theta / \lambda \), \( 2 \theta \) is the scattering angle, \( \lambda \) the wavelength. The units are respectively \( \text{Å} \) and \( \text{Å}^{-1} \).

\( h = \{h,k,l\} \): position of the reciprocal-lattice points.

\( F_h \): the true (and unknown) structure factor of the reflection \( h \); \( F_h \) is the modulus, \( \phi_h \) the phase.

\( A_h, B_h \): the real and imaginary components of \( F_h \).

\( I_h \): the observed value of the intensity of the native form (note that \( I_h \neq F_h^2 \)).

\( I_{j,h} \): the observed value of the intensity of the \( j \)th isomorphous derivative.

\( \langle W(s) \rangle \): the spherical average of the function \( W(s) \) in the space of the variable \( s \).

\( \Phi_k = \sum_{k=1}^{N} f_k^2(s) \): \( f_k(s) \) is the form factor of the \( k \)th atom, \( N \) is the number of atoms in the unit cell.

II-1. Ideal case

(a) Reciprocal space. We are concerned here with the diffraction of an X-ray beam by a crystal. The parameters of the phenomenon – namely the mathematical entities which are necessary and sufficient to describe it completely – are in reciprocal space the modulus \( F_h \) and the phase \( \phi_h \) of the reflections. The set of these parameters can be represented by a vector:

\[ X = \{F_h, \phi_h\}. \]

Another vector \( Y_0 \) represents the observations; \( Y_0 \) is one particular value of the vector \( Y \) representing all the possible intensities of the reflections for the native form as well as for the \( J \) isomorphous derivatives:

\[ Y_0 = \{I_{h}, \{I_{j,h}\}\}. \]

The following probability densities for the two variables \( X \) and \( Y \) are involved (Taupin & Luzzati, 1982):

\[ q(Y|X) \] (II-1-3a)

probability of \( Y \) knowing \( X \);

\[ t(X) \] (II-1-3b)

prior probability of \( X \);

\[ r(X|Y) = \frac{q(Y|X)t(X)}{\int q(Y|X)t(X) \, dx} \] (II-1-3c)

probability of \( X \) knowing \( Y \).

\( q(Y|X) \) has the properties of a transfer function. The result of an experiment is to yield one particular drawing \( Y_0 \) of the random vector \( Y \); in addition, the knowledge of the behaviour of the experimental device specifies the form of the function \( q(Y|X) \), at least in the vicinity of \( Y_0 \). In fact, an experiment is aimed at determining the probability density of \( X \) knowing \( Y \) – which is defined by the function \( r(X|Y) \) – rather than that of \( Y \) knowing \( X \), defined by \( q(Y_0|X) \). The determination of \( r(X|Y_0) \) (see II-1-3c) involves both \( q(Y_0|X) \) and \( t(X) \); only when \( t(X) \) is much flatter than \( q(Y_0|X) \) does \( q(Y_0|X) \) suffice to define \( r(X|Y_0) \).

The function \( t(X) \) expresses the prior probability density of \( X \), before any experiment. The form of \( t(X) \) is specified by the choice, in real space, of a reference stochastic model.

We are now set to formulate the following theorem and to introduce the notion of information (Taupin & Luzzati, 1982).

Theorem 1 (Kullback, 1968). The probability that the probability density for the variable \( X \), which a priori was \( t(X) \), becomes \( r(X) \) after the experimental observation of \( Y \) and of \( q(Y|X) \), is:

\[ P = 2^{-H} \]

\[ H = \int r(X) \log_2 \left[ \frac{r(X)}{t(X)} \right] \, dx. \] (II-1-4b)

\( H \) is the measure of the information – in short the information, expressed in bits – associated with the pair of probability densities \( r(X) \) and \( t(X) \). It is convenient to express \( H \) in terms of \( q(Y_0|X) \). This is done by inserting (II-1-3c) in (II-1-4b). The result is

\[ H = (\log 2)^{-1} \left[ -\log K + L/K \right] \] (II-1-5a)

\[ K = \int q(Y_0|X)t(X) \, dx \] (II-1-5b)

\[ L = \int q(Y_0|X)t(X) \log[q(Y_0|X)] \, dx. \] (II-1-5c)
Note that the function $t(X)$ is normalized – namely $\int t(X) \, dx = 1$ – whereas the function $q(X)$ is not.

(b) Real space. We now discuss how one can make use in real space of the information retrieved in reciprocal space. We assume that the structure is completely defined in real space by a set of structure parameters represented by a multidimensional vector $Z$. It is clear that one and only one vector $X$ corresponds to each vector $Z$:

$$X = f(Z).$$  \hspace{1cm} (II-1-6)

It is possible to introduce in real space probability densities of the variable $Z$ analogous to those introduced in reciprocal space (see II-1-3); we shall specify by the suffices $x$ and $z$ the functions in each space. A reference stochastic model defines the *a priori* probability density in real space, $t_x(Z)$. Its counterpart is the *a posteriori* probability density $t_z(X)$ (see II-1-3b). If $r_x(Z|Y)$ is the probability density in real space – namely the probability of $Z$ knowing $Y$ – if $r_z(X|Y)$ is its reciprocal-space counterpart, and if the function $q(Y|X)$ depends on the reciprocal space only, then the information conveyed by $Y$ is the same in reciprocal and in real space:

$$H_d(Y) = \int r_x(X|Y) \log_2 \left[ \frac{r_x(X|Y)}{t_x(X)} \right] \, dx \hspace{1cm} (II-1-7a)$$

$$H_r(Y) = \int r_z(Z|Y) \log_2 \left[ \frac{r_z(Z|Y)}{t_z(Z)} \right] \, dz \hspace{1cm} (II-1-7b)$$

$$H_{rz}(Y) = H_{xz}(Y), \hspace{1cm} (II-1-7c)$$

The statement above relies upon the following theorem (the proof is not given here).

**Theorem 2.** Given (a) two vectorial spaces $X$ and $Z$, (b) an *a priori* probability density $t_x(Z)$ in the space $Z$, (c) an experimental measurement $Y$ of probability $q(Y|X)$, which is a function of $X$, (d) a uniquely defined function $f(Z)$ relating $X$ to $Z$, then the information conveyed by $Y$ is the same in reciprocal and in real space:

$$H_d(Y) = H_r(Y) = H_{xz}(Q), \hspace{1cm} (II-1-8a)$$

where $H_d(Y)$ is given by (II-1-7a) and

$$H_r(Y) = \int p_z(Z|Q) \log_2 \left[ \frac{p_z(Z|Q)}{t_z(Z)} \right] \, dz \hspace{1cm} (II-1-8b)$$

Note that the parameter $Q$ can convey a notion akin to structural resolution.

II-2. Realistic case

The mathematical treatment above involves the multidimensional functions $q(Y_0|X)$ and $t(X)$. With regard to $t(X)$ Bricogne (1984) has recently shown that, contrary to previous expectations (see Klug, 1958), conspicuous correlations are likely to exist between different structure factors even when the number of atoms is large. The function $q(Y_0|X)$, which expresses the stochastic properties of the experimental procedures, can also be expected to display non-negligible correlations due to systematic errors (for example absorption) and to the methods used in the analysis of the data (for example multiple isomorphous replacements). It is difficult to take these correlations into proper account and thus to obtain satisfactory expressions of the functions $q(Y_0|X)$ and $t(X)$; consequently, the information cannot be reliably estimated using (II-1-5) (see in Appendix 1 a discussion of the effects of the correlations on the information).

On the other hand, if each reflection $h$ is considered independently of all the other reflections – in other words if the intensities of each reflection, for the native form as well as for the isomorphous derivatives, are measured in one independent experiment – then the relevant functions $\ell(F_h)$ and $\varphi(I_{h'}|I_{j,h'},F_h)$ can be determined fairly easily. One can contemplate the possibility of using these functions for the purpose of determining the value of some parameter related to structure resolution. The problem can be sketched as follows.

The function $\ell(F_h)$ is the projection of $t(X)$:

$$\ell(F_h) = \int t(X) \, dx \, \mathbf{F}_h, \hspace{1cm} (II-2-1)$$

and its expression is given by Wilson’s formulae (see § II-2a). The determination of the function

$$\varphi(I_{h'}|I_{j,h'},F_h)$$

is discussed in § II-2b. Besides, according to the procedure described in § II-1b we introduce in real space a virtual experiment defining a probability law $p_z(Z|Z';Q)$, and determine in reciprocal space the corresponding probability law $\varphi'(F_h|F_h,Q)$: $Z'$ and $F_h$ are the ‘observed’ values corresponding to one virtual experiment, $Q$ is a parameter specifying the stochastic properties of the virtual experiment. Finally, by comparing the experimental and the virtual probability laws $\varphi$ and $\varphi'$ we strive to optimize $Q$.

In fact, the forms of the two functions are so different – most often (see Fig. 1) $\varphi$ is narrow in the radial direction and wide in the angular direction, whereas $\varphi'$ is circularly symmetrical around the point $F_h/D_h$ – that a direct comparison is not possible. One may resort instead to a procedure based upon some global property of the two functions, not too rigidly...
dependent upon their forms and relative position. The information — namely the probability of observing the probability law \( \rho(Y_{oh}|F_{h}) \) when the prior probability density is \( g(F_{h}) \) — provides a convenient parameter for that purpose. We adopt here a procedure involving the projection information \( \rho(F_{h}) \) and \( \rho'(F_{h}|F_{h},Q) \): \( \rho(F_{h}) \).

Let us discuss the various steps of this procedure.

(a) The function \( \rho(F_{h}) \). The prior probability densities correspond to a large number of atoms uniformly and independently distributed in the unit cell; their expressions are given by Wilson's formulae (International Tables for X-ray Crystallography, 1952)

\[
\rho(F_{h}) = \frac{1}{\sqrt{2\pi q_{h}}} \exp\left(-\frac{F_{h}^{2}}{2q_{h}}\right) \quad (\text{II-2-2})
\]

for acentric reflections,

\[
\rho(F_{h}) = \frac{1}{\sqrt{2\pi q_{h}}} \exp\left(-\frac{F_{h}^{2}}{2q_{h}}\right) \quad (\text{II-2-3a})
\]

for centric reflections,

\[
\rho(F_{h}) = \frac{1}{\sqrt{2\pi q_{h}}} \exp\left(-\frac{F_{h}^{2}}{2q_{h}}\right) \quad (\text{II-2-3b})
\]

(b) The function \( \rho(Y_{oh}|F_{h}) \). According to an approximation introduced in this field by Blow & Crick (1959) we assume that the probability law of the modulus is defined by the observation \( I_{h} \) of the native form, whereas the probability law of the phase is defined by the observation of the derivatives \( \{I_{j,h}\} \):

\[
\rho(Y_{oh}|F_{h}) = \rho(I_{h}|F_{h}) \rho(\{I_{j,h}\}|I_{h}) \quad (\text{II-2-5})
\]

and that the probability law for the phase is the product of the contributions of each derivative:

\[
\rho(\{I_{j,h}\}|I_{h}) = P_{h}(\psi_{h}) = \prod_{j=1}^{J} p_{j}(\psi_{h}) \quad (\text{II-2-6})
\]

Strictly speaking, \( P_{h}(\psi_{h}) \) is not only a function of the reflection \( h \) but also of the other reflections, whose intensities are indeed involved in the determination of the heavy-atom parameters (position, occupancy, etc.). Quite often, yet, the uncertainty of these parameters depends less on the experimental errors of the intensity measurements than on other causes independent of the X-ray scattering experiments (degree of isomorphism, etc.). Under these conditions it can reasonably be assumed, at least to a first approximation, that the function \( P_{h}(\psi_{h}) \) is independent of the accuracy of the intensity measurements of all the reflections but \( h \).

With regard to \( \rho(I_{h}|F_{h}) \) we adopt a Gaussian form:

\[
\rho(I_{h}|F_{h}) = \exp\left[-\frac{(F_{h} - I_{h})^{2}}{2\sigma_{h}^{2}}\right] \quad (\text{II-2-7})
\]

With one exception (see Appendix 2) we assume that \( \sigma \) is a function of \( I \) (this question is discussed by Taupin & Luzzati, 1982). Note that in (II-2-7) \( I_{h} \) is allowed to take negative values, in keeping with the fact that the experimental intensity is the difference between signal and noise.

(c) The virtual experiment and the function \( \rho'(F_{h}|F_{h},Q) \). We assume a large number of atoms, each independently distributed with the probability density \( \rho(\Delta r_{k}) \) around its 'correct' position \( r_{k} \): \( \Delta r_{k} = r_{k} + \Delta r_{k} \) is the 'experimental' position. The function \( \rho(\Delta r) \) is the same for all the atoms. The probability law for \( F_{h} \) corresponding to this stochastic model has been determined by Luzzati (1952). For the real component of acentric reflections the law is

\[
p_{h}(\Delta A_{h}) = \left[\frac{\pi\Phi_{h}(1 - D_{h}^{2})}{D_{h}^{2}}\right]^{-1/2} \times \exp\left[-\frac{[(\Delta A_{h} - (D_{h} - 1)A_{h})^{2}/\Phi_{h}(1 - D_{h}^{2})]\} \right] \quad (\text{II-2-8a})
\]

where

\[
A_{h} = \sum_{k=1}^{N} f_{h}(k) \cos(2\pi r_{k}, h) \quad (\text{II-2-8b})
\]

Fig. 1. Schematic representation of the probability laws for one reflection considered independently of all the other reflections (§ II-2): case of an acentric reflection. The isodensity lines are drawn at 0.2-0.4-0.6-0.8 of the maximum. \( \rho(F_{h}) \) is the probability of observing the intensities \( I_{h}, \{I_{j,h}\} \) when \( F_{h} \) is known. \( \rho(F_{h}) \) is assumed to be the product of a radial function \( \rho_{v}(F_{h}) \) and of an angular function \( \rho_{a}(\psi_{h}) \) (§ II-2b). \( \rho_{v}(F_{h}) \) is much sharper than \( \rho_{a}(\psi_{h}) \). \( \rho_{v}(F_{h}) \) corresponds to the virtual experiment (§ II-2c, equations (II-2-10) and (II-2-11)): \( \rho_{v}(F_{h}) = \exp\left[-\frac{[(\Delta A_{h} - (D_{h} - 1)A_{h})^{2}/\Phi_{h}(1 - D_{h}^{2})]\} \right] \) with \( D_{h} = 0.85 \). \( \rho(F_{h}) \) is the prior probability density, corresponding to Wilson's statistics (§ II-2, equations (II-2-21) and (II-2-22)). \( \rho(F_{h}) = (\pi\Phi_{h})^{-1} \times \exp\left(-F_{h}^{2}/\Phi_{h}\right) \).
\[ \Delta A_h = \sum_{k=1}^{N} f_k(h) \left[ \cos(2\pi r_k^h) - \cos(2\pi r_k^h) \right]. \]

\[ D_h = \int \varphi(r) \cos(2\pi h) \, dv_r. \]

Assuming that \( \varphi(r) \) is centrosymmetric, \( D_h \) takes the form

\[ D_h = \int \varphi(r) \cos(2\pi h) \, dv_r. \]

According to the definition of § II-1a, \( A_h \) and \( A_h' \) (= \( A_h + \Delta A_h \)), which correspond respectively to the 'correct' \( (r_k) \) and the 'experimental' \( (r_k') \) atomic positions, are elements of the vectors \( X \) and \( Y \), respectively. Equation (II-2-8a) can thus be rewritten as follows [see (II-1-3a) and remember that the function \( g \) need not be normalized]:

\[ \varphi(A_h') A_h D_h = \exp \left[ - \frac{(A_h' - D_h A_h)^2}{\Phi_s(1 - D_h^2)} \right]. \]

Similarly, for the imaginary component of the acentric reflections one has:

\[ \varphi(B_h') B_h D_h = \exp \left[ - \frac{(B_h' - D_h B_h)^2}{\Phi_s(1 - D_h^2)} \right]. \]

and for the centric reflections (see Luzzati, 1952):

\[ \varphi(F_h') F_h D_h = \exp \left[ - \frac{(F_h' - D_h F_h)^2}{2\Phi_s(1 - D_h^2)} \right]. \]

These probability laws are asymptotically correct and the laws for \( A_h \) and \( B_h' \) are independent of each other, when \( N \to \infty \). Note that the laws above are functions of the 'experimental' positions \( r_k' \) of the probability density \( \varphi(r) \) and of \( \Phi_s \).

(d) The projection information \( K_h \): a function of \( g(Y_{0h}|F_h) \) and of \( \ell(F_h) \). \( K_h \) can be calculated by introducing the expressions for \( g(Y_{0h}|F_h) \) [(II-2-5) to (II-2-7)] and \( \ell(F_h) \) [(II-2-2) to (II-2-4)] in (II-1-4) and (II-1-5). The result is

\[ K_h = K_{F,h} + K_{A,h}. \]

with

\[ K_{F,h} = (\log 2)^{-1} \left( -\log K_{F,h} + L_{F,h}/K_{F,h} \right) \]

\[ K_{A,h} = (\log 2)^{-1} \left( -\log K_{A,h} + L_{A,h}/K_{A,h} \right) \]

\[ K_{F,h} = \int g_f(l_h|F_h) \ell_f(F_h) \, dF_h \]

\[ L_{F,h} = \int g_f(l_h|F_h) \ell_f(F_h) \log[g_f(l_h|F_h)] \, dF_h \]

\[ K_{A,h} = \int g_a(l_h|F_h) \ell_a(F_h) \, dF_h \]

\[ L_{A,h} = \int g_a(l_h|F_h) \ell_a(F_h) \log[g_a(l_h|F_h)] \, dF_h. \]

\[ K_{A,h} = \int g_a(l_h|F_h) \ell_a(F_h) \, dF_h. \]

\[ L_{A,h} = \int g_a(l_h|F_h) \ell_a(F_h) \log[g_a(l_h|F_h)] \, dF_h. \]

\[ K_{A,h} = \int g_a(l_h|F_h) \ell_a(F_h) \, dF_h. \]

\[ L_{A,h} = \int g_a(l_h|F_h) \ell_a(F_h) \log[g_a(l_h|F_h)] \, dF_h. \]

For the centric reflections, \( \ell_a(\alpha) \) takes non-zero values only when \( \alpha = 0 \) or \( \alpha = \pi \) [see (II-2-4)]; therefore only two values of \( P_a(\alpha) \) are needed, which we denote by \( p_+ \) for \( \alpha = 0 \) and by \( p_- \) for \( \alpha = \pi \). Thus

\[ L_{A,h} = (p_+ \log p_+ + p_- \log p_-)/2 \]

with

\[ p_+ + p_- = 1. \]

Inserting (II-2-4b) and (II-2-18a) in (II-2-14b) we obtain
\[ \tau \mathcal{H}_{z,h} = 1 + p_+ \log_2 p_+ + p_- \log_2 p_- \]  
(II-2-19)  

[Note that \( \tau \mathcal{H}_{z,h} = 0 \) for \( p_+ = p_- = 1/2 \) and that the maximal information – one bit – is obtained for \( p_+ \) (or \( p_- \)) = 1.]

The expressions above apply to the reflections whose intensity is large enough to be observed. The observation that the intensity of one reflection is smaller than some (known) threshold provides some information: the evaluation of this information is discussed in Appendix 2. As a rule, the phase of these weak reflections remains totally unknown: therefore \( \mathcal{H}_{z,h} = 0 \) for these reflections.

(e) The projection information \( \mathcal{H}_{z,w} \), a function of \( \varphi'(F_\hbar, D_\hbar) \) and of \( \varphi(F_\hbar) \). Let us consider first acentric reflections. When \( N \to \infty \) the probability laws \( \varphi' \) and \( \varphi \) for the real and imaginary components of \( F_\hbar \) become independent of each other. Therefore, the expressions of \( \varphi' \) and \( \varphi \) take the form:

\[
\varphi'(F_\hbar, D_\hbar) = \varphi'(A_\hbar, D_\hbar) \varphi'(B_\hbar, D_\hbar)
\]  
(II-2-20)

and the information associated with \( F_\hbar \) is the sum of the information associated with \( A_\hbar \) and with \( B_\hbar \). Introducing (II-2-10) and

\[
\varphi'(A_\hbar) = (\pi \varphi)^{-1/2} \exp(-A_\hbar^2/\varphi)
\]  
(II-2-22)

in the expression of the information (see II-1-5) one obtains:

\[
\langle \mathcal{H}_{z,w} \rangle = -\left( \frac{2 \log 2}{\log 1 - D_\hbar^2} \right) \left( \frac{2 \log 2}{\log 2} \right)
\]  
(II-2-23a)

Assuming that \( \varphi(r) \) – and consequently \( D_\hbar \) – is spherically symmetric, and that \( \langle A_\hbar^2 \rangle = \langle B_\hbar^2 \rangle = \varphi/2 \) (International Tables for X-ray Crystallography, 1952), then the spherical average of (II-2-23) takes a simpler form, independent of \( A_\hbar \):

\[
\langle \mathcal{H}_{z,w} \rangle = -\frac{1}{\log 2} \left( \frac{2 \log 2}{\log 1 - D_\hbar^2} \right)
\]  
(II-2-24a)

\[
\langle \mathcal{H}_{z,w} \rangle = \langle \mathcal{H}_{z,w} \rangle + \langle \mathcal{H}_{w,b} \rangle = -\frac{1}{\log 2} \left( \frac{2 \log 2}{\log 2} \right)
\]  
(II-2-24b)

For centric reflections the result is

\[
\langle \tau \mathcal{H}_{z,w} \rangle = -\log(1 - D_\hbar^2)/(2 \log 2).
\]  
(II-2-25)

Note, as expected, that the information tends to zero as \( \varphi(r) \) flattens out (in this case \( D_\hbar \to 0 \) for \( h \neq 0 \)), and that it tends to infinity as \( \varphi(r) \) sharpens up (\( D_\hbar \to 1 \) in this case).

(f) A practical determination of positional resolution.

The optimization of the function \( \varphi(r) \) which specifies the stochastic structure of the virtual experiment – is based upon a comparison of the projection information functions \( \mathcal{H}_{z,w} \) and \( \mathcal{H}_{w,b} \). As discussed above (see Fig. 1), the forms of the functions \( \varphi \) and \( \varphi' \) are so different that the possibility appears most unlikely of finding a physically and mathematically plausible expression for \( \varphi(r) \) which leads to a satisfactory agreement between \( \mathcal{H}_{z,w} \) and \( \mathcal{H}_{w,b} \) over an extended volume of reciprocal space. More realistically, the problem can be approached as follows. One postulates the mathematical form of \( \varphi(r) \), chosen to be a function of one parameter; then the value of that parameter is determined by equalling the value of the sums of \( \mathcal{H}_{z,w} \) and of \( \mathcal{H}_{w,b} \) over the whole reciprocal space, and for the same reciprocal lattice.

\( \mathcal{H}_{z,w} \) can be calculated for all the reflections taken into account by the experiment, including the weakest reflections (see § II.2d and Appendix 2). For all the other reflections, for example those at \( s > s_{\text{max}} \), \( \mathcal{H}_{w,b} \) is taken to be zero. Thus the sum

\[
\mathcal{R} = \sum_{h \neq 0} \mathcal{H}_{z,w}
\]  
(II-2-26)

can be calculated.

With regard to the virtual experiment, we choose a Gaussian form for \( \varphi(r) \):

\[
\varphi(r) dr = \left( \frac{3}{2 \pi} \right)^{3/2} U^3 \exp \left( -\frac{3 r^2}{2 U^2} \right) 4 \pi r^2 dr,
\]  
(II-2-27a)

where

\[
U^2 = (\Delta r)^2,
\]  
(II-2-27b)

which leads to the following expression of \( D_s^2 \) (see II-2-9):

\[
D_s^2 = \exp \left( -\frac{4}{3} \pi^2 U^2 s^2 \right).
\]  
(II-2-28)

For space group \( P1 \) the sum of the \( \mathcal{H}_{z,w} \) can be determined as follows (see II-2-25b):

\[
\sum_{h \neq 0} \mathcal{H}_{z,w} \approx -\frac{1}{\log 2} \frac{V}{4 \pi}
\]  

\[
\times \int_0^\infty s^2 \log \left[ 1 - \exp \left( -\frac{4}{3} \pi^2 U^2 s^2 \right) \right] ds
\]  

\[
= 0.113 V U^{-3} = \mathcal{H}^a.
\]  
(II-2-29)

\( V/2 \) is the volume of the unit cell divided by the multiplicity factor in reciprocal space.

It is useful to introduce a new parameter, the density of information \( \eta \) (in bits per \( \text{Å}^3 \)):

\[
\eta(s_{\text{m}}) = \sum_{h \neq 0} \mathcal{H}_{z,w}/V.
\]  
(II-2-30a)

\( U \) then takes the form

\[
U^3 \approx 0.113/\eta.
\]  
(II-2-30b)
The $U$ factor – namely the root-mean-square uncertainty of the atomic positions (see II-2-27b) – provides a convenient expression of positional resolution.

The points must be stressed that all the chemical information involved in the treatments of § II-2 is conveyed by the parameter $\Phi$, and that $\Phi$ is determined empirically (see § III). Therefore, the expressions of $\eta$ and of $U$ are invariant with respect to the number and nature of the chemical elements used in the representation of the structure – be these the individual atoms or clusters of atoms (corresponding, for example, to the individual amino acids) – provided that $N \gg 1$ and that each cluster is approximately spherical. On the contrary, the estimates of information in real space discussed in Appendix 1 and thus the results of that Appendix depend upon the number $N$.

### III. An illustration

The tryptic fragment of methionyl tRNA synthetase (MTS) (Zelwer et al., 1982) provides a convenient illustration of the results of § II. The molecular weight of this protein is 64 000 daltons; each molecule contains approximately 4750 non-hydrogen atoms. The space group is $P_{2_1}$, with two molecules per unit cell; the cell parameters are $a=78.1$, $b=46.2$, $c=87.9$ Å, $\beta=108.8^\circ$, the volume per asymmetric unit is 151 000 Å$^3$. The structure has been solved with the reciprocal space truncated at $s<0.4$ Å$^{-1}$, using several heavy atom and one anomalous scattering derivatives (Zelwer et al., 1982). The files available to us contained the experimental intensities of all the observed reflections, corrected for polarization and geometric factors, and the phase probability laws $P_b(q_\omega)$.

The functions $\Phi$ and $\sigma_f$ (see Notation and II-2-7) were determined empirically as explained in the legends of Figs. 2 and 3. The results are:

$$\Phi_s=4.1 \times 10^7 \exp(-14.2 \, s^2) \tag{III-1}$$

$$\sigma_f(l,s)=(4.6 \times 10^6)/(L_p)+8.2 \times 10^{-2} \, l \tag{III-2a}$$

where

$$L_p=(1+\cos^2 \, 2\theta)/\sin2\theta. \tag{III-2b}$$

In this work the observed intensities were truncated at 10$^6$. Therefore (see Appendix 2),

$$l_{\min}=10^6. \tag{III-3}$$

Once for one reflection $h$ the parameters $\Phi$ and $\sigma_f$ are known, as well as the intensity $l$, then the projection information $\mathcal{H}_{F,h}$ can be calculated [see (II-2-16) and (II-2-17), (A-2-4) and (A-2-5)]. Similarly, $\mathcal{H}_{F,h}$ can be calculated when $P_b(q_\omega)$ is known [see (II-2-16) and (II-2-19)]. These calculations were performed for all the reflections contained in the sphere of radius $s=0.4$ Å$^{-1}$. The results are reported in Table 1 and in Figs. 4 and 5.

It may be noted that as $s$ increases the projection information per zone increases up to ca 0.3 Å$^{-1}$, and then decreases rather abruptly (see Fig. 4). It may thus be concluded that in the case of MTS and with the crystallographic data used in this section little additional information is to be gained by extending $s$.

---

**Fig. 2.** Determination of the function $\Phi_s$. The reciprocal space was divided into eight spherical shells $[(n-1) \times 0.05 < s < n \times 0.05$ Å$^{-1}$] and the average $(\langle I_h \rangle)$ computed for each shell. These values are reported in the figure. The equation of the straight line is $4.1 \times 10^7 \exp(-14.2 \, s^2)$.

**Fig. 3.** Determination of the variance of the observed intensities. The observed amplitudes, $(I_h)^{1/2}$, corrected for the polarization and geometric factors, were available for several groups of symmetry-related reflections. The amplitude of these reflections, divided by the average Lorentz-polarization factor $(L_p)^{1/2}=[(1+\cos^2 \, 2\theta)/\sin2\theta]^{1/2}$, is called $F_{exp}(h)$. The averages $F_{exp}$ and $(F_{exp}-F_{exp})^2=\sigma_{exp}^2$ are calculated for all the symmetry-related reflections, and histogrammed as a function of $s$ and $F_{exp}$. The $s$ dependence turns out to be negligible. The results are shown by dots in the figure. The full line represents the function $\sigma_{exp}$ vs $F_{exp}$ corresponding to the equation $\sigma_1=(4.6 \times 10^6)/(L_p)+8.2 \times 10^{-2} \, l$, on the assumption that the probability law for $I$ takes the form $\exp[-(I-I)^2/2\sigma^2]$, see equation (II-2-7)).
The values of the global projection information (in bits) for all the independent reflections contained in the spherical shell of radii \( s_2 \) and \( s_b \). \( \mathcal{H}_p \) and \( \mathcal{H}_s \) are the values associated with the moduli and with the phases. The suffices 1 and \( \bar{1} \) refer to the acentric and to the centric reflections; the last two columns report the sum of \( \mathcal{H} \) for the two classes of reflections. For each spherical shell the upper and the lower rows of \( \mathcal{H}_p \) report the values relevant respectively to the observed \( (I > I_n) \) and to the unobserved \( (I \leq I_n) \) reflections. \( n \) is the number of independent reflections of each group.

<table>
<thead>
<tr>
<th>( s_2 ) (Å(^{-1}))</th>
<th>( s_b ) (Å(^{-1}))</th>
<th>( n )</th>
<th>( \mathcal{H}_p )</th>
<th>( \mathcal{H}_s )</th>
<th>( n )</th>
<th>( \mathcal{H}_p )</th>
<th>( \mathcal{H}_s )</th>
<th>( \mathcal{H}_p )</th>
<th>( \mathcal{H}_s )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05000</td>
<td>0.07500</td>
<td>64</td>
<td>214.09</td>
<td>101.30</td>
<td>18</td>
<td>65.802</td>
<td>12.693</td>
<td>82</td>
<td>279.89</td>
</tr>
<tr>
<td>0.07500</td>
<td>0.10000</td>
<td>21</td>
<td>100.84</td>
<td>94.651</td>
<td>26</td>
<td>14.197</td>
<td>168</td>
<td>57.813</td>
<td>226.66</td>
</tr>
<tr>
<td>0.10000</td>
<td>0.12500</td>
<td>27</td>
<td>123.60</td>
<td>50.076</td>
<td>19</td>
<td>63.130</td>
<td>72</td>
<td>287.34</td>
<td></td>
</tr>
<tr>
<td>0.12500</td>
<td>0.15000</td>
<td>47</td>
<td>204.21</td>
<td>114.78</td>
<td>25</td>
<td>19.877</td>
<td>405</td>
<td>123.94</td>
<td>627.53</td>
</tr>
<tr>
<td>0.15000</td>
<td>0.17500</td>
<td>64</td>
<td>262.09</td>
<td>70.183</td>
<td>29</td>
<td>123.94</td>
<td>627.53</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.17500</td>
<td>0.20000</td>
<td>98</td>
<td>375.75</td>
<td>142.04</td>
<td>35</td>
<td>80.025</td>
<td>133</td>
<td>455.77</td>
<td></td>
</tr>
<tr>
<td>0.20000</td>
<td>0.22500</td>
<td>144</td>
<td>465.71</td>
<td>242.45</td>
<td>73</td>
<td>72.288</td>
<td>180</td>
<td>537.99</td>
<td></td>
</tr>
<tr>
<td>0.22500</td>
<td>0.25000</td>
<td>193</td>
<td>563.13</td>
<td>224.45</td>
<td>48</td>
<td>103.33</td>
<td>165</td>
<td>518.14</td>
<td></td>
</tr>
<tr>
<td>0.25000</td>
<td>0.27500</td>
<td>238</td>
<td>788.68</td>
<td>123.60</td>
<td>51</td>
<td>94.651</td>
<td>14.197</td>
<td>57.813</td>
<td>226.66</td>
</tr>
<tr>
<td>0.27500</td>
<td>0.30000</td>
<td>289</td>
<td>2671.7</td>
<td>1222.4</td>
<td>73</td>
<td>242.45</td>
<td>728.28</td>
<td>180</td>
<td>537.99</td>
</tr>
<tr>
<td>0.30000</td>
<td>0.32500</td>
<td>345</td>
<td>3268.2</td>
<td>1338.5</td>
<td>71</td>
<td>213.36</td>
<td>3481.6</td>
<td>1374.2</td>
<td></td>
</tr>
<tr>
<td>0.32500</td>
<td>0.35000</td>
<td>391</td>
<td>3595.7</td>
<td>1557.0</td>
<td>71</td>
<td>206.02</td>
<td>3801.7</td>
<td>1591.9</td>
<td></td>
</tr>
<tr>
<td>0.35000</td>
<td>0.37500</td>
<td>447</td>
<td>3557.7</td>
<td>1982.8</td>
<td>70</td>
<td>162.24</td>
<td>1871.7</td>
<td>3719.9</td>
<td>2016.8</td>
</tr>
<tr>
<td>0.37500</td>
<td>0.40000</td>
<td>502</td>
<td>3268.2</td>
<td>2102.3</td>
<td>76</td>
<td>142.44</td>
<td>3470.7</td>
<td>1238.3</td>
<td></td>
</tr>
<tr>
<td>0.40000</td>
<td>0.42500</td>
<td>557</td>
<td>3595.7</td>
<td>2190.9</td>
<td>71</td>
<td>206.02</td>
<td>3801.7</td>
<td>1591.9</td>
<td></td>
</tr>
<tr>
<td>0.42500</td>
<td>0.45000</td>
<td>612</td>
<td>3557.7</td>
<td>1982.8</td>
<td>70</td>
<td>162.24</td>
<td>1871.7</td>
<td>3719.9</td>
<td>2016.8</td>
</tr>
<tr>
<td>0.45000</td>
<td>0.47500</td>
<td>667</td>
<td>3268.2</td>
<td>2102.3</td>
<td>76</td>
<td>142.44</td>
<td>3470.7</td>
<td>1238.3</td>
<td></td>
</tr>
<tr>
<td>0.47500</td>
<td>0.50000</td>
<td>722</td>
<td>3595.7</td>
<td>1982.8</td>
<td>70</td>
<td>162.24</td>
<td>1871.7</td>
<td>3719.9</td>
<td>2016.8</td>
</tr>
</tbody>
</table>

Beyond 0.4 Å\(^{-1}\), it is also clear that the most substantial fraction of the projection information is provided by the intensity of the observed reflections and that the contribution of the weakest reflections is small.

As discussed in § II the interpretation of \( \mathcal{H} \) in terms of structural resolution is based upon the choice of a virtual experiment and the specification of its stochastic parameters. This choice is to a large extent arbitrary. We adopt here the model discussed in § II.2c with a Gaussian distribution \( \sigma(r) \) (see § II.2f): this model depends on one parameter, the root-mean-square uncertainty \( U \) of the atomic positions [see (II-2-27)]. In this case each experimental value of \( \mathcal{H} \) defines one value of \( U \). The evolution of \( \eta \) and of \( U \) as a function of \( s_m \) [see (II-2-30)] is plotted in Fig. 5.

As expected, the structural resolution improves – namely the \( U \) factor decreases – as \( s_m \) increases. Moreover, at all values of \( s_m \) the overall resolution is better than the resolution determined using only the amplitude or the phase of the reflections. Note that in this example the smallest value of the \( U \) factor is \( \frac{U}{10} \) Å. This figure should be interpreted to mean the best resolution which can possibly be attained if the structure of MTS were determined using only the crystallographic data reported by Zelwer et al. (1982) and in the absence of additional (for example chemical) information.

The MTS data can also be used to illustrate the effects of the cross correlations on the statistical analysis (see § II-2 and Appendix 1). If the cross correlations were negligible then projection information and plain information would coincide and by virtue of theorem 2 of § II-1b the experimental value of \( \mathcal{H} \) could be transferred to real space. The information in real space can be determined by reference to the model introduced in Appendix 1 (\( N \) points each thrown at random in a volume \( v \)); the expression of \( H_z \) is given by (A-1-3). Since \( \mathcal{H} \) can be equalled to \( H_z \), (A-1-3) can be solved for \( v \) and the value of the \( U \) factor determined assuming that \( v \) is spherical.

\[
U^3 = \left( \frac{3v}{4\pi} \right) \left( \frac{3}{5} \right)^{3/2}. \tag{III-4}
\]

The values of \( U \) corresponding to (A-1-3) and (III-4) are plotted in Fig. 6 along with the values of \( U \) determined according to (II-2-30). For \( s_m \) small the two values remain pretty close (their ratio is \( 1 \pm 0.4 \)) and then sharply depart from each other as \( s_m \) becomes larger than 0.280 Å\(^{-1}\). As discussed in Appendix 1, \( s_m = 0.280 \) Å\(^{-1}\) is the point at which the number of degrees of freedom associated with all the
reflections with \( s \leq s_m \) is the same as the number of degrees of freedom of all the atoms of one asymmetric unit. Note that for the whole of the reflections of the present study of MTS \( (s_m = 0.4 \, \text{Å}^{-1}) \) the two resolutions differ by a factor close to six: in this case the number of degrees of freedom is approximately three times larger in reciprocal than in real space.

We have also plotted in Fig. 6 the value of the conventional 'resolution' \( s^{-1} \). The fact is noteworthy that the ratio \( s^{-1}_m / U \) is almost constant (and close to 4) over the whole range of \( S_m \).

We wish to mention some recent - and preliminary - results of the structure refinement of MTS which provide the most rewarding indication that our estimate of resolution is realistic. Dr S. Brunie, who is carrying out the refinement - she uses Hendrickson & Konnert's procedure (Konnert, 1976) which involves a great deal of chemical information - kindly informs us that according to her latest (but not yet final) results the atoms which originally had been located in the multiple isomorphous replacement (MIR) map (approximately 70% of the total) have undergone displacements whose root-mean-square value is 0.93 Å. Since electron density maps are notoriously sensitive to the phase and insensitive to the modulus of the structure factors, we presume that the MIR map mirrors the phase rather than the modulus information. Therefore, the observed root-mean-square displacement should be compared with the accuracy estimated using only the phase information. The agreement is surprisingly good (0.93 and 1.10 Å, see Fig. 5).

Note that all the results of this section, with the exception of the determination of \( U \) via \( (A-1-3) \) and \( (III-4) \) (open dots in Fig. 6) are independent of \( N \) (see § II-2f).

IV. Conclusions

(a) We approach in this paper some stochastic aspects of protein crystallography. Information emerges as a privileged notion to deal with these problems. The most significant properties are expressed by the two theorems of § II: (1) information is a clearly defined mathematical parameter; (2) information can be transferred from reciprocal to real space.

(b) A practical application of these theorems hinges upon an adequate knowledge of the probability densities in the multidimensional space of the variables. Most often, in fact, only the probability laws relevant to each reflection considered independently of all the other reflections can be determined: the statistical correlations between different reflections are much more difficult to estimate. Under these circumstances it is not possible to determine the overall information conveyed in the multidimensional space of all the reflections (see also point e below).

(c) In order to overcome this drawback we introduce another probabilistic parameter associated with each reflection – which we name projection information – whose value we show how to determine. The interpretation in real space of this projection information is not as straightforward as when theorem 2 can

Fig. 4. Global projection information (●) for the sets of reflections contained in the successive spherical shells of thickness 0.025 Å\(^{-1}\). The contributions of the phases (×) and of the moduli (+) of the observed reflections, as well as of the moduli of the unobserved reflections (⊙), are reported separately.

Fig. 5. \( \eta \) is the sum of the projection information (●) of the reflections contained inside the sphere of radius \( s_m \) divided by the volume of the asymmetric unit [see equation (II-2-30a)]; the contributions of the moduli (+) and of the phases (×) are reported separately. The \( U \) factor is the root-mean-square uncertainty of the atomic positions [see equations (II-2-27b) and (II-2-30b)]. The dotted circles correspond to the root-mean-square displacement of the atomic positions at the present stage of the structure refinement of MTS (see text).
be relied upon. As a solution of the problem, we propose to compare the experimental projection information with the projection information corresponding to a virtual experiment introduced in real space, the two sets of information being determined with the same reference stochastic model. By this procedure the experimental projection information can be interpreted in terms of one parameter specifying the stochastic properties of the virtual experiment: we use here the root-mean-square deviation of the atomic positions, a parameter akin to structural resolution.

(d) It is timely to survey the approximations involved in our treatment. The effects of the cross correlations on the function $t$ are eliminated (see § II-2 and point e below). Cross correlation effects could also distort the function $q$; we discuss in § II-2b why we believe that these effects are mild. Other approximations involved in the determination of the function $P(x)$ are commonplace in protein crystallography (see also § II-2b). More seriously, the scope of our treatment, and thus the precise meaning of resolution are restricted by the particular choice of the virtual experiment (see §§ II-2c and f).

(e) An interesting question is to assess the effects of the cross correlations in structure-factor statistics. The simulation discussed in Appendix I points to a highly accurate, but inapplicable to real data. It is obviously highly desirable to be able to test the results of this work on some example in which accuracy could be assessed by other procedures. In this respect let us emphasize the very purpose of this work, namely to predict accuracy and resolution from the data before solving the structure, and point out that all standard methods to assess accuracy are based upon a comparison of observed and calculated structure factors (and thus come into play only after the structure is solved). Yet, an analysis of the current refinement of MTS (see § III) provides a most rewarding preliminary confirmation that our estimate of accuracy is realistic.

(h) It is barely necessary to point out how improper is the conventional use of the position of the cut off in reciprocal space to express resolution. Any parameter expressing resolution should indeed fulfill at least two conditions: operationally, its definition should involve the accuracy of the experimental observations; conceptually, its significance should be related to the potential accuracy of the structure parameters eventually determined using those observations. The parameters introduced in this work – $H$, $U$, $U$ – fulfill these conditions.

(i) It is obvious that the observation that $\tilde{R}$ is much smaller than $R$ (see § III and Fig. 5) does not mean that the crystallographer’s efforts to improve the knowledge of the phases are ill invested or poorly rewarded, but rather that at the late stages of structure refinement only a modest gain in accuracy can be
expected from involving the phases, in addition to the moduli, in the refinement algorithm. On the contrary, at the early stages of structure analyses the phase information is likely to play a predominant role, since the quality of the MIR maps — whose interpretation is the decisive step in the process of solving the structure — is far more sensitive to the phase than to the modulus of the structure factors.

(j) All the results obtained in reciprocal space, namely the determination of the projection information and the analysis in terms of resolution (essentially the whole of § II-2) are invariant with respect to the number and nature of the structure elements (the ‘atoms’). On the contrary, the results relevant to real space (see Appendix 1) are obtained by reference to models in which the number N of ‘atoms’ must be specified.

(k) No account is taken in this work of the chemical information — bond lengths and angles, peptide sequence, etc. The remark can be made, in this respect, that usually hardly any chemical argument is involved in the early stages of protein structure analysis. Chemical arguments play their full role in the interpretation of the electron density maps and have the effect of increasing — perhaps dramatically — the information in real space.

(l) From a practical viewpoint, it may be noted that the computations involved in the determination of the projection information are simple. The remark can also be made that the results of this work are not restricted to proteins.

(m) Some results of this work may have potential applications in the strategy of phase determination. We intend to discuss this problem in a forthcoming paper.

We thank Dr G. Bricogne for the communication of unpublished results and for highly valuable suggestions. We are grateful to Drs S. Brunie, J. L. Risler and C. Zelwer for generously providing the data on MTS. This work was supported in part by a grant from the Délegation Générale à la Recherche Scientifique et Technique.

APPENDIX 1

An analysis of the influence of the statistical correlations between different reflections

Theorem 2 (see § II-1b), which states that the information calculated in real and in reciprocal space must take the same value, can be used to assess the influence of the statistical correlations between different reflections.

The information associated with the virtual experiment of § II-2c and with the reference stochastic model of § II-2a can be easily determined in real space if the probability density $\phi(r)$ is a step function over the volume $v$. This information $H_z$ can be compared to the sum of the projection information $H^p$: the two should be equal if the reflections were statistically independent.

The information $H_z$ is determined as follows. The probability that any point of the unit cell of volume $V$ be external to $n$ boxes of equal volume $v$, randomly distributed in $V$, is $(1-v/V)^n$. Therefore, the probability that one point drawn at random in $V$ falls inside any of the $n$ boxes is

$$p_n = 1 - (1 - v/V)^n. \quad (A-1-1)$$

The overall probability is

$$P = \prod_{n=1}^{N} p_n \quad (A-1-2)$$

and the information

$$H_z = - \sum_{n=1}^{N} \log_2(1 - (1 - v/V)^n). \quad (A-1-3)$$

For $Nv \ll V$, $(A-1-3)$ can be simplified:

$$H_z \approx - N \log_2(Nv/V). \quad (A-1-4)$$

In reciprocal space, when the reflections are statistically independent of each other the information is equal to the sum of the projection information $H^p$ (see §§ II-2e and f). For space group $P1$ the sum of the associated with the reciprocal-lattice points contained in the sphere of radius $s_m$ is

$$\mathcal{H}^p(s_m) = \sum_{h \neq 0} \mathcal{H}^p \simeq -2\pi V \int_0^{s_m} s^2 \log_2(1 - D_k^2) \, ds. \quad (A-1-5)$$

The number $M(s_m)$ of those lattice points is

$$M(s_m) \simeq (2/3) \pi s_m^3 V. \quad (A-1-6)$$

We shall consider two cases.

(a) The volume $v$ is small

More precisely, we assume that all the dimensions of $v$ are much smaller than $(s_m)^{-1}$. Under these conditions we can use Guinier’s approximation:

$$\langle 1 - D_k^2 \rangle = (3/4) \pi^2 R_g^2 s^2, \quad (A-1-7)$$

where $R_g$ is the radius of gyration of the volumes $v$. Replacing $(A-1-7)$ in $(II-2-24b)$ and performing the calculation, we obtain

$$\mathcal{H}^p(s_m)/H_z \xrightarrow{v s_m^3 \to 0} (2/3) M(s_m)/N. \quad (A-1-8)$$

It is noteworthy that for $v s_m^3 \to 0$ the value of $s_m$ at which $\mathcal{H}^p(s_m)$ is equal to $H_z$ corresponds to the number of degrees of freedom being equal in real $(3N)$ and in reciprocal space $[2M(s_m)]$. This points to an interesting
correlation between the number of reflections involved in the statistical analysis and the relative importance of the cross terms of \( t(X) \) and \( q(X) \).

(b) \( v \) is a sphere of radius \( R \)

In this case

\[
D(s) = 3(\sin t - t \cos t) / t^3 \quad (A-1-9a)
\]

\[
t = 2\pi Rs. \quad (A-1-9b)
\]

We can determine the expressions of the ratio of the information \( \mathcal{H}((s_m)/H_z \) by replacing \((A-1-9) \) into \((II-2-24b) \) and by taking into account \((A-1-5) \). This ratio is a function of \( R, V, N \) and \( s_m \). We have plotted in Fig. 7 the values of \( \mathcal{H}((s_m)/H_z \) vs \( s_m \) for a few values of \( R; V \) and \( N \) are those of MTS (see §III). This figure suggests several comments:

(i) at all values of \( R \) the ratio increases monotonically with \( s_m \);

(ii) as discussed above [see \((A-1-8)\)], for \( R \) small the ratio becomes equal to one at the value of \( s_m \) at which the number of degrees of freedom is the same in real and in reciprocal space;

(iii) as \( R \) increases the value of \( s_m \) at which \( \mathcal{H}((s_m)/H_z \) becomes equal to \( H_z \) increases - in other words the constraints imposed by the cross correlations are relaxed;

(iv) in the case of MTS (as in other proteins, since \( V/N \) is almost constant in proteins) the condition \( \mathcal{H}((s_m)/H_z = 1 \) is fulfilled at \( s_m = 0.279 \text{ Å}^{-1} \), with the truncation of the present analysis \( s_m = 0.4 \text{ Å}^{-1} \), Zelwer et al., 1982), the assumption that the cross terms of \( t(X) \) and of \( q(X) \) are negligible would lead to an over-estimation of the information by a factor close to 3.

Finally, we wish to stress that although the results of this Appendix correspond to particular choices of the virtual experiment and of the function \( \phi(r) \), yet their significance is likely to be of rather general bearing. Note also that the results obtained here for space group \( P1 \) can be easily extended to space group \( P1 \).

APPENDIX 2

Projection information associated with the weakest reflections

Strictly speaking, \((II-2-16) \) and \((II-2-17) \) apply to all the reflections, provided the intensity \( I \) is determined for each of them, even the weakest ones (for some of these reflections \( I \) takes a negative value). Yet, according to a widespread (and unfortunate) practise the measured intensities are truncated at some threshold \( I_m \) - namely \( I \) is taken to be zero whenever its value is smaller than \( I_m \). The expression of \( q_{lF}(I/F) \) corresponding to this situation takes the form

\[
q_{lF}^*(I/F) = \delta(I) \int \frac{dI}{\sigma^2} \exp\left[-\frac{(I - F^2)^2}{2\sigma^2}\right] dI.
\]

(A-2-1)

In order to carry out an explicit calculation it is convenient to assume here that \( \sigma \) is a function of \( F^2 \) (and not of \( I \) as before). In this case \( q_{lF}^* \) becomes

\[
q_{lF}^*(I/F) = \delta(I) \frac{\sigma^2}{(\pi/2) \sigma^2} \int T(t) \, dz
\]

(A-2-2a)

with

\[
T(t) = 1 + (2/\pi)^{1/2} \int_0^t e^{-y^2} \, dy.
\]

(A-2-2b)

Replacing \((II-2-3a) \) (or \( b \)) and \((A-2-2a) \) in \((II-2-14a) \) and \((II-2-14b) \), and introducing the following variables and parameters:

\[
z = F/\Phi^{1/2}
\]

\[
\gamma^2(z) = \sigma(F^2)/\sigma(z^2\Phi)/\Phi
\]

\[
z_m^2 = I_m/\Phi,
\]

we obtain

\[
1K_{lF}^* = (2\pi)^{1/2} \int_0^\infty z \sigma(z^2\Phi) T\left[\frac{z_m^2 - z^2}{2^{1/2} \gamma^2(z)}\right] e^{-z^2} \, dz
\]

(A-2-4a)

\[
1L_{lF}^* = (2\pi)^{1/2} \int_0^\infty z \sigma(z^2\Phi) T\left[\frac{z_m^2 - z^2}{2^{1/2} \gamma^2(z)}\right] \exp(-z^2) \times \log\left\{(\pi/2)^{1/2} \sigma(z^2\Phi) T\left[\frac{z_m^2 - z^2}{2^{1/2} \gamma^2(z)}\right]\right\} \, dz
\]

(A-2-4b)
\[ \begin{align*}
\Gamma K^z &= \int_0^\infty \sigma(z^2 \Phi) T \left[ \frac{z_m^2 - z^2}{2^{1/2} \sqrt{z}(z)} \right] \exp(-z^2/2) \, dz \\
\Gamma L^z &= \int_0^\infty \sigma(z^2 \Phi) T \left[ \frac{z_m^2 - z^2}{2^{1/2} \sqrt{z}(z)} \right] \exp(-z^2/2) \\
&\quad \times \log \left\{ (\pi/2)^{1/2} \sigma(z^2 \Phi) T \left[ \frac{z_m^2 - z^2}{2^{1/2} \sqrt{z}(z)} \right] \right\} \, dz.
\end{align*}\] 

References


