

## Sur l'Emploi des Méthodes Statistiques dans l'Étude de la Structure Cristalline des Protéines

PAR V. LUZZATI

*Protein Structure Project, Polytechnic Institute of Brooklyn, Brooklyn, N.Y., U.S.A.  
et Centre de Recherches sur les Macromolécules, 6 rue Boussingault, Strasbourg, France\**

(Reçu le 9 mars 1955, revu le 13 juillet 1955)

It is shown that the statistical methods of Wilson are inadequate to explain the intensity distribution experimentally observed with certain protein crystals. The limitations inherent in Wilson's treatment are examined and new theoretical laws are deduced corresponding to more exact models and taking account of the actual distribution of atoms in the unit cell and of the correlation between the atomic positions. It is shown that these laws are in good agreement with the intensity distributions obtained from crystals of ribonuclease and methemoglobin.

### Notation

$N$  — nombre d'atomes contenus dans une maille élémentaire.

$V$  — volume de la maille élémentaire.

$\mathbf{r}_j$  — vecteur de position de l'atome  $j$ .

$\mathbf{s}$  — vecteur de position dans l'espace réciproque ( $|\mathbf{s}| = 2 \sin \theta / \lambda$ ).

$f_j$  — facteur de structure de l'atome  $j$ .

$\mathbf{F}$  — facteur de structure de la maille élémentaire.

$A, B$  — composantes réelle et imaginaire de  $\mathbf{F}$ .

$a_j = f_j \cos(2\pi \mathbf{r}_j \times \mathbf{s})$  — contribution à  $A$  de l'atome  $j$ .

$b_j = f_j \sin 2\pi(\mathbf{r}_j \times \mathbf{s})$  — contribution à  $B$  de l'atome  $j$ .

$\Phi_n = \sum_{j=1}^N f_j^n$ .

$\langle \psi(\mathbf{s}) \rangle$  — moyenne des valeurs de  $\psi(\mathbf{s})$  pour tous les points de la surface d'une sphère de rayon  $|\mathbf{s}|$ ,  $\psi(\mathbf{s})$  étant une fonction quelconque.

### Introduction

Lorsqu'on applique les méthodes statistiques aux intensités diffractées par un réseau cristallin, il convient de ne pas confondre les deux problèmes suivants:

*Problème I.*—On choisit les positions des atomes d'une maille élémentaire selon une certaine loi aléatoire: chaque choix des positions atomiques définit une structure. En se plaçant en un point fixe du réseau réciproque on cherche à déterminer la fonction

$$p_I(X) dX \quad (1)$$

qui représente la probabilité que  $X$  soit compris entre  $X$  et  $X+dX$  ( $X$  peut être  $I, A$  ou  $B$ ).

*Problème II.*—Soit la même maille élémentaire, dans laquelle toutes les positions atomiques sont fixes. Considérons un ensemble  $M$  de nœuds du réseau réciproque, contenus dans une région limitée de l'espace

réciproque: le nombre de nœuds de l'ensemble soit grand. On veut calculer la fonction

$$p_{II}(X) dX \quad (2)$$

qui représente la fraction des nœuds de l'ensemble  $M$  dans lesquels  $X$  a une valeur comprise entre  $X$  et  $X+dX$ .

Les méthodes mathématiques adoptées par tous les auteurs qui ont appliqué les raisonnements statistiques aux problèmes des structures (en particulier Wilson, 1949; Karle & Hauptman, 1952, 1953) sont basées sur l'existence de variables aléatoires: elles se réfèrent au problème I ci-dessus. Au contraire le cristallographe s'intéresse au problème II, car il cherche à déterminer la distribution de  $X$  pour une certaine structure, les positions atomiques étant fixes.

En principe les fonctions  $p_I(X)$  et  $p_{II}(X)$  sont différentes: on a seulement prouvé qu'elles coïncident (Karle & Hauptman, 1952) dans le cas où  $p_I(X)$  est calculé en laissant au hasard le choix des positions atomiques, et l'ensemble  $M$  est formé par tous les nœuds contenus dans un volume suffisamment étendu de l'espace réciproque. En pratique toutefois l'accord entre les lois théoriques de Wilson et les distributions expérimentales est généralement satisfaisant, même lorsque le choix des nœuds se limite à une mince couche sphérique de l'espace réciproque, pourvu que le rayon moyen de cette couche ne soit pas trop petit. Au contraire on observe un écart systématique entre la distribution des intensités et les lois de Wilson lorsqu'on s'approche de l'origine de l'espace réciproque: ces écarts sont particulièrement importants pour les cristaux des protéines (voir le chapitre des exemples), car les taches de diffraction se trouvent concentrées autour de l'origine de l'espace réciproque, à cause des grandes dimensions des mailles élémentaires et de l'effacement rapide des intensités en fonction de l'angle de diffraction.

C'est en examinant la distribution statistique des

\* Adresse actuelle.

intensités de quelques cristaux de protéines que nous nous sommes aperçu des limitations qui affectent les lois calculées par Wilson. Nous nous sommes alors proposé de calculer d'autres lois théoriques, qui soient en meilleur accord avec la distribution des intensités données par l'expérience. Nous avons choisi des modèles statistiques plus élaborés que celui adopté par Wilson, en tenant compte de deux propriétés des cristaux de protéines que le modèle de Wilson néglige: la distribution réelle des atomes, dont les fluctuations, dans les protéines, sont plus régulières que celles dues au hasard (Harker, 1953) et la corrélation des positions atomiques due aux liaisons chimiques et aux empêchements stériques.

Chaque modèle est caractérisé par la loi aléatoire qui règle le choix des positions atomiques (voir ci-dessus la définition du problème I):

*Modèle 1.*—Le choix des positions atomiques se fait au hasard: chaque atome a une probabilité  $dv/V$  d'occuper un élément de volume  $dv$  (un élément de volume est 'occupé' lorsqu'il contient le centre de l'atome).

*Modèle 2.*—On choisit les positions atomiques en attribuant à chaque atome une probabilité d'occuper un élément de volume, qui est une fonction de la position de l'élément de volume dans la maille élémentaire. Dans ces deux modèles le choix de chaque position atomique est indépendant de l'emplacement de tous les autres atomes.

*Modèle 3.*—On partage la maille élémentaire en un nombre donné de parallélépipèdes égaux, et on admet que chaque parallélépipède contient un nombre fixe (et grand) d'atomes. À l'intérieur de chaque parallélépipède les positions atomiques sont choisies au hasard, les unes indépendantes des autres.

*Modèle 4.*—On partage la maille élémentaire en un grand nombre de cases identiques, dont le volume est suffisamment petit pour qu'un seul atome puisse y être contenu. On attribue à chaque case une certaine probabilité d'être occupée: cette probabilité est une fonction de la position de la case dans la maille élémentaire.

Le modèle 1 est celui adopté par Wilson (1949). C'est en cherchant à établir des lois valables pour des structures formées d'atomes distribués avec une certaine régularité que nous avons été amené à choisir le modèle 2: les modèles 3 et 4 ont été proposés pour tenir compte en outre de la corrélation des positions atomiques. Le modèle 3 fixe les fluctuations du nombre d'atomes contenus dans des fractions de maille élémentaire dont les dimensions sont grandes par rapport à celles des atomes: la corrélation est donc à grande distance. Le modèle 4, au contraire, tient compte de la corrélation à courte distance, due à l'empêchement stérique des atomes et aux liaisons chimiques.

Nous n'essayerons pas de prouver de manière rigoureuse que les lois statistiques relatives aux modèles 2, 3 ou 4 coïncident en général avec la distri-

bution des intensités des protéines; nous examinerons l'accord entre la distribution des intensités diffractées par les cristaux de ribonucléase et de méthémoglobine et les lois théoriques, en tachant ainsi d'établir un rapport empirique entre les problèmes I et II ci-dessus. On peut remarquer, d'ailleurs, que cette même restriction limite la validité de la plupart des applications des méthodes statistiques à d'autres structures que celle des protéines; il est rare en effet que l'analyse statistique soit faite en respectant les conditions établies par Karle & Hauptman (1952), qui sont les seules dans lesquelles il ait été prouvé que les deux problèmes coïncident.

### Calcul des lois de distribution

Les lois de distribution statistique que nous calculons ci-dessous seront employées dans la région de l'espace réciproque voisine de l'origine: Karle & Hauptman (1952) ont montré, et la valeur asymptotique des lois calculées ci-dessous va le confirmer, que lorsque  $|s|$  est suffisamment grand, les lois de distribution coïncident avec celles calculées par Wilson.

Les lois de distribution sont calculées en utilisant un théorème connu du calcul des probabilités que Wilson (1949) et nous même (Luzzati, 1952) avons utilisé dans un problème analogue. Nous faisons l'hypothèse que le nombre  $N$  d'atomes est grand. Nous supposons en outre que les atomes sont tous identiques, pour rendre plus simple la déduction des lois mathématiques. Toutefois cette restriction n'est pas nécessaire: les lois sont valables même si les atomes sont d'espèces différentes, pourvu qu'aucun atome ne soit prépondérant.

#### Modèle 1

Chaque position atomique est une variable aléatoire et indépendante, distribuée uniformément dans le volume de la maille élémentaire.

Ce cas est celui traité par Wilson (1949) et Karle & Hauptman (1952, 1953). Les lois de distribution des intensités sont:

$$p(I)dI = (1/\Phi_2) \exp(-I/\Phi_2)dI \quad (\text{non centrosymétrique}), \quad (3)$$

$$p(I)dI = (2\pi\Phi_2 I)^{-\frac{1}{2}} \exp(-I/2\Phi_2)dI \quad (\text{centrosymétrique}). \quad (4)$$

#### Modèle 2

Nous calculons les lois de distribution statistique en supposant que chaque position atomique est une variable aléatoire et indépendante, dont la loi de distribution dans la maille élémentaire est:

$$q(\mathbf{r})dv. \quad (5)$$

$q(\mathbf{r})$  est la même pour tous les atomes.

*Cas non centrosymétrique.*—Calculons la loi de distribution de  $A$ :

$$A = f \sum_{j=1}^N \cos 2\pi(\mathbf{r}_j \times \mathbf{s}) = \sum_{j=1}^N a_j. \quad (6)$$

Chaque terme  $a_j$  est ici une variable aléatoire et indépendante; sa loi de distribution est une fonction de  $q(\mathbf{r})$  (5).

$$\bar{a}_j = f \int_V \cos 2\pi(\mathbf{r} \times \mathbf{s}) q(\mathbf{r}) dv = f A_q(\mathbf{s}), \quad (7)$$

$$\begin{aligned} \overline{(a_j - \bar{a}_j)^2} &= f^2 \int_V [\cos 2\pi(\mathbf{r} \times \mathbf{s}) - A_q]^2 q(\mathbf{r}) dv \\ &= \frac{1}{2} f^2 [1 + A_q(2\mathbf{s}) - 2A_q^2(\mathbf{s})], \end{aligned} \quad (8)$$

$$\bar{A} = \sum_{j=1}^N \bar{a}_j = \Phi_1 A_q, \quad (9)$$

$$\overline{(A - \bar{A})^2} = \sum_{j=1}^N \overline{(a_j - \bar{a}_j)^2} = \frac{1}{2} \Phi_2 [1 + A_q(2\mathbf{s}) - 2A_q^2(\mathbf{s})]. \quad (10)$$

Si le contenu total de la fonction  $q(\mathbf{r})$  n'est pas concentré dans un petit volume de la maille élémentaire  $A_q(\mathbf{s})$  et  $A_q(2\mathbf{s})$  sont négligeables en comparaison de l'unité, car le contenu total de  $q(\mathbf{r})$  dans une maille élémentaire est égal à 1. On peut donc simplifier l'équation (10):

$$\overline{(A - \bar{A})^2} \cong \frac{1}{2} \Phi_2. \quad (11)$$

$\bar{A}$  et  $\overline{(A - \bar{A})^2}$  définissent la loi de distribution statistique de  $A$ :

$$p(A) dA = (\pi \Phi_2)^{-\frac{1}{2}} \exp[-(A - \Phi_1 A_q)^2 / \Phi_2] dA. \quad (12)$$

D'une manière analogue, on calcule la loi de distribution de  $B$ :

$$p(B) dB = (\pi \Phi_2)^{-\frac{1}{2}} \exp[-(B - \Phi_1 B_q)^2 / \Phi_2] dB, \quad (13)$$

où

$$B_q = \int_V q(\mathbf{r}) \sin 2\pi(\mathbf{r} \times \mathbf{s}) dv. \quad (14)$$

Déterminons la loi de distribution de  $I$ . La probabilité que  $A$  et  $B$  aient simultanément des valeurs comprises entre  $A$  et  $A+dA$ , et entre  $B$  et  $B+dB$ , est:

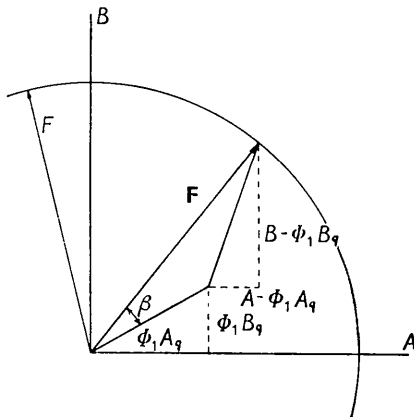


Fig. 1. Représentation schématique du plan  $A, B$ .

$$\begin{aligned} p(A)p(B) dA dB &= (\pi \Phi_2)^{-1} \\ &\times \exp[-\{(A - \Phi_1 A_q)^2 + (B - \Phi_1 B_q)^2\} / \Phi_2] dA dB. \end{aligned} \quad (15)$$

Effectuons le changement de variables (voir Fig. 1):

$$\begin{aligned} (A - \Phi_1 A_q)^2 + (B - \Phi_1 B_q)^2 \\ = |F|^2 + \Phi_1^2 (A_q^2 + B_q^2) - 2\Phi_1 |F| (A_q^2 + B_q^2)^{\frac{1}{2}} \cos \beta, \end{aligned} \quad (16)$$

$$dA dB = |F| d|F| d\beta. \quad (17)$$

$A_q$  et  $B_q$  sont les composantes réelle et imaginaire de la transformée de Fourier de  $q(\mathbf{r})$ :

$$A_q^2 + B_q^2 = I_q. \quad (18)$$

La probabilité que  $|F|$  ait une valeur comprise entre  $|F|$  et  $|F|+d|F|$  est égale à la probabilité que l'extrémité du vecteur  $\mathbf{F}$  soit contenu entre les cercles de rayon  $|F|$  et  $|F|+d|F|$  (voir Fig. 1):

$$\begin{aligned} p(|F|) d|F| &= (\pi \Phi_2)^{-1} \\ &\times \left\{ \int_0^{\beta} \exp \left[ -\frac{|F|^2 + \Phi_1^2 I_q - 2\Phi_1 |F| \sqrt{I_q} \cos \beta}{\Phi_2} \right] |F| d\beta \right\} \\ &\times d|F|. \end{aligned} \quad (19)$$

En remplaçant  $|F|^2$  par  $I$  on a

$$\begin{aligned} p(I) dI &= \Phi_2^{-1} \exp \left[ -\frac{I + \Phi_1^2 I_q}{\Phi_2} \right] \\ &\times \left\{ \frac{1}{2\pi} \int_0^{2\pi} \exp \left[ \frac{2\Phi_1 \sqrt{II_q}}{\Phi_2} \cos \beta \right] d\beta \right\} dI. \end{aligned} \quad (20)$$

L'intégrale entre crochets dans (20) est égale à la fonction de Bessel à variable imaginaire, d'ordre zéro:

$$\frac{1}{2\pi} \int_0^{2\pi} \exp(z \cos \beta) d\beta = I_0(z). \quad (21)$$

En remplaçant dans (20) on a:

$$p(I) dI = \Phi_2^{-1} \exp \left[ -\frac{I + \Phi_1^2 I_q}{\Phi_2} \right] I_0 \left[ \frac{2\Phi_1 \sqrt{II_q}}{\Phi_2} \right] dI. \quad (22)$$

*Cas centrosymétrique.*—

$$F = 2f \sum_{j=1}^{N/2} \cos 2\pi(\mathbf{r}_j \times \mathbf{s}) = 2 \sum_{j=1}^{N/2} a_j. \quad (23)$$

D'une manière analogue à celle employée dans le cas non-centrosymétrique nous calculons  $\bar{F}$  et  $\overline{(F - \bar{F})^2}$ .

$$\bar{a}_j = 2f \int_V \cos 2\pi(\mathbf{r} \times \mathbf{s}) q(\mathbf{r}) dv = 2f F_q, \quad (24)$$

$$\begin{aligned} \overline{(a_j - \bar{a}_j)^2} &= 4f^2 \int_V [\cos 2\pi(\mathbf{r} \times \mathbf{s}) - F_q]^2 q(\mathbf{r}) dv \\ &= 2f^2 [1 + F_q(2\mathbf{s}) - 2F_q^2(\mathbf{s})] \cong 2f^2, \end{aligned} \quad (25)$$

$$\bar{F} = \sum_{j=1}^{N/2} \bar{a}_j = \Phi_1 F_q, \quad (26)$$

$$\overline{(F-\bar{F})^2} = \sum_{j=1}^{N/2} \overline{(a_j - \bar{a}_j)^2} \simeq \Phi_2. \quad (27)$$

On obtient ainsi

$$p(F) dF = (2\pi\Phi_2)^{-\frac{1}{2}} \exp[-(F-\Phi_1 F_q)^2/2\Phi_2] dF. \quad (28)$$

Calculons la loi de distribution de  $I$  en remplaçant  $F$  par  $\sqrt{I}$  dans l'équation (28) et en tenant compte du fait qu'une valeur de  $I$  correspond à  $F$  et à  $-F$ :

$$p(I) dI = (2\pi\Phi_2)^{-\frac{1}{2}} \left\{ \exp\left[-\frac{(\sqrt{I}-\Phi_1 F_q)^2}{2\Phi_2}\right] + \exp\left[-\frac{(-\sqrt{I}-\Phi_1 F_q)^2}{2\Phi_2}\right] \right\} \frac{dI}{2\sqrt{I}}. \quad (29)$$

### Modèle 3

Partageons la maille élémentaire en  $T$  parallélépipèdes égaux de volume  $v$ . Supposons que  $n_i$  atomes sont distribués au hasard à l'intérieur du parallélépipède  $t$ .

Calculons la loi de distribution de  $A$ :

$$A = f \sum_T \sum_{n_i} \cos 2\pi(\mathbf{r}_i + \mathbf{r}_i \times \mathbf{s}) = \sum_T \sum_{n_i} a_{ii}. \quad (30)$$

$\mathbf{r}_i$  définit la position du centre du parallélépipède  $t$ ,  $\mathbf{r}_i$  indique la distance d'un atome  $i$  au centre du parallélépipède  $t$ , à l'intérieur duquel il est contenu. La somme double s'étend à tous les atomes d'un parallélépipède et à tous les parallélépipèdes.  $\mathbf{r}_i$  est la variable aléatoire.

$$\bar{A} = \sum_T \sum_{n_i} \bar{a}_{ii}, \quad (31)$$

$$\bar{a}_{ii} = f \overline{\cos 2\pi(\mathbf{r}_i + \mathbf{r}_i \times \mathbf{s})} = f \cos 2\pi(\mathbf{r}_i \times \mathbf{s}) A_v. \quad (32)$$

$A_v$  est la transformée de Fourier d'un parallélépipède de volume  $v$  ayant une densité uniforme égale à  $1/v$ :

$$A_v = \frac{1}{v} \int_v \cos 2\pi(\mathbf{r} \times \mathbf{s}) dv, \quad (33)$$

$$\bar{A} = \sum_T \sum_{n_i} \bar{a}_{ii} = f A_v \sum_T n_i \cos 2\pi(\mathbf{r}_i \times \mathbf{s}) = N f A_v A'_q. \quad (34)$$

Nous définissons ainsi:

$$A'_q = \sum_T \frac{n_i}{N} \cos 2\pi(\mathbf{r}_i \times \mathbf{s}), \quad (35)$$

dont l'analogie avec  $A_q$  (7) est évidente. La différence entre  $A_q$  et  $A'_q$  est d'autant plus petite que les variations de  $q(\mathbf{r})$  (5) sont plus lentes, et que  $|s|$  est plus petit (voir Sayre, 1951).

On a donc:

$$\bar{A} = \Phi_1 A_v A'_q. \quad (36)$$

Calculons ensuite  $\overline{(A-\bar{A})^2}$ :

$$\overline{(A-\bar{A})^2} = \sum_T \sum_{n_i} \overline{(a_{ii} - \bar{a}_{ii})^2}, \quad (37)$$

$$\begin{aligned} \overline{(a_{ii} - \bar{a}_{ii})^2} &= f^2 [\cos 2\pi(\mathbf{r}_i + \mathbf{r}_i \times \mathbf{s}) - A_v \cos 2\pi(\mathbf{r}_i \times \mathbf{s})]^2 \\ &= \frac{1}{2} f^2 \{1 + A_v(2\mathbf{s}) \cos 4\pi(\mathbf{r}_i \times \mathbf{s}) \\ &\quad - A_v^2 [1 + \cos 4\pi(\mathbf{r}_i \times \mathbf{s})]\}, \quad (38) \end{aligned}$$

$$\overline{(A-\bar{A})^2} = \frac{1}{2} N f^2 \{1 + A_v(2\mathbf{s}) A'_q(2\mathbf{s}) - A_v^2 [1 + A'_q(2\mathbf{s})]\}. \quad (39)$$

On peut simplifier (voir passage des équations (10) à (11)):

$$\overline{(A-\bar{A})^2} \simeq \frac{1}{2} \Phi_2 (1 - A_v^2). \quad (40)$$

Le calcul de  $\bar{B}$  et  $\overline{(B-\bar{B})^2}$  est analogue:

$$\bar{B} = \Phi_1 A_v B'_q, \quad (41)$$

$$\overline{(B-\bar{B})^2} = \frac{1}{2} \Phi_2 (1 - A_v^2). \quad (42)$$

Si on compare les équations (36) et (9), (40) et (11), on remarque que les lois de distribution de l'intensité correspondant au modèle 3 ont la même forme que celles du modèle 2, en remplaçant  $\Phi_1^2 I_q$  par  $\Phi_1^2 I_q A_v^2$  et  $\Phi_2$  par  $\Phi_2 (1 - A_v^2)$ . Ces lois sont portées plus loin (équations (57) et (58)).

### Modèle 4

Divisons la maille élémentaire en un nombre  $L$  de cases égales dont le volume  $w$  est suffisamment petit pour que chacune ne puisse contenir qu'un seul atome. Soit  $l$  la case dont le centre se trouve au point  $\mathbf{r}_l$ . Introduisons une variable aléatoire  $z_l$ , qui représente le fait qu'un atome soit contenu à l'intérieur de la case  $l$ . La loi de distribution statistique de  $z_l$  est définie par les conditions suivantes:

$$p(z_l) = 0 \quad \text{si } z \neq 0 \quad \text{ou } z \neq 1, \quad (43)$$

$$\lim_{\epsilon \rightarrow 0} \int_{-\epsilon}^{+\epsilon} p(z_l) dz_l = 1 - N w q''(\mathbf{r}_l), \quad (44)$$

$$\lim_{\epsilon \rightarrow 0} \int_{1-\epsilon}^{1+\epsilon} p(z_l) dz_l = N w q''(\mathbf{r}_l). \quad (45)$$

$q''(\mathbf{r}_l)$  est une fonction analogue à  $q(\mathbf{r}_l)$  (5).

On calcule les lois de distribution en considérant la contribution de chaque case comme une variable aléatoire, fonction des deux variables indépendantes  $z_l$  et  $\mathbf{r}'$ .  $\mathbf{r}'$  représente la position de l'atome à l'intérieur de  $w$ : on suppose que la distribution de  $\mathbf{r}'$  est uniforme.

$$\bar{A} = \sum_L \bar{a}_l, \quad (46)$$

$$a_l = z_l f \cos 2\pi(\mathbf{r}_l + \mathbf{r}' \times \mathbf{s}), \quad (47)$$

$$\begin{aligned} \bar{a}_l &= z_l f \overline{\cos 2\pi(\mathbf{r}_l + \mathbf{r}' \times \mathbf{s})} \\ &= N w q''(\mathbf{r}_l) f \cos 2\pi(\mathbf{r}_l \times \mathbf{s}) A_w, \quad (48) \end{aligned}$$

$$\bar{A} = N f A_w w \sum_L q''(\mathbf{r}_l) \cos 2\pi(\mathbf{r}_l \times \mathbf{s}) = \Phi_1 A_w A'_q. \quad (49)$$

$A_w$  et  $A'_q$  sont analogues aux fonctions  $A_v$  (33) et  $A'_q$  (35) définies pour le modèle 3: il faut toutefois remarquer que le volume d'une case  $l$  est bien plus petit que celui d'un parallélépipède du modèle 3.

Calculons la valeur de  $\overline{(A-\bar{A})^2}$ :

$$\begin{aligned} \overline{(A-\bar{A})^2} &= \sum_L \overline{(a_l - \bar{a}_l)^2}, \quad (50) \\ \overline{(a_l - \bar{a}_l)^2} &= f^2 [z_l \cos 2\pi(\mathbf{r}_l + \mathbf{r}' \times \mathbf{s}) - z_l A_w \cos 2\pi(\mathbf{r}_l \times \mathbf{s})]^2 \\ &= \frac{1}{2} f^2 \{ z_l^2 [1 + A_w(2\mathbf{s}) \cos 4\pi(\mathbf{r}_l \times \mathbf{s})] \\ &\quad - \bar{z}_l^2 A_w^2 [1 + \cos 4\pi(\mathbf{r}_l \times \mathbf{s})] \}. \quad (51) \end{aligned}$$

Puisque (voir (44) et (45)):

$$\bar{z}_l^2 = z_l = Nwq''(\mathbf{r}_l), \quad (52)$$

$$\begin{aligned} \overline{(A-\bar{A})^2} &= \frac{1}{2} f^2 \sum_L \{ Nwq''(\mathbf{r}_l) [1 + A_w(2\mathbf{s}) \cos 4\pi(\mathbf{r}_l \times \mathbf{s})] \\ &\quad - N^2 w^2 q''^2(\mathbf{r}_l) A_w^2 [1 + \cos 4\pi(\mathbf{r}_l \times \mathbf{s})] \} \\ &\cong \frac{1}{2} f^2 \{ Nw \sum_L q''(\mathbf{r}_l) - N^2 w^2 A_w^2 \sum_L q''^2(\mathbf{r}_l) \}. \quad (53) \end{aligned}$$

En tenant compte des conditions  $V = wL$  et  $\bar{q}'' = 1/V$  on a:

$$\overline{(A-\bar{A})^2} = \frac{1}{2} \Phi_2 \{ 1 - A_w^2 (N/L) (\bar{q}''^2 / q''^2) \}. \quad (54)$$

On calcule  $\bar{B}$  et  $\overline{(B-\bar{B})^2}$  d'une manière analogue:

$$\bar{B} = \Phi_1 A_w B'_q, \quad (55)$$

$$\overline{(B-\bar{B})^2} = \frac{1}{2} \Phi_2 \{ 1 - A_w^2 (N/L) (\bar{q}''^2 / q''^2) \}. \quad (56)$$

Les lois de distribution de l'intensité ont donc la forme (22) et (29), en remplaçant  $\Phi_1^2 I_q$  par  $\Phi_1^2 I'_q A_w^2$  et  $\Phi_2$  par  $\Phi_2 \{ 1 - A_w^2 (N/L) (\bar{q}''^2 / q''^2) \}$ .

On remarque que  $\{ (N/L) (\bar{q}''^2 / q''^2) \}$  est plus petit ou égal à 1 (équations 44 et 45).

### Résumé

Les lois de distribution correspondant aux quatre modèles statistiques ont toutes la même forme: non centrosymétrique:

$$p(I) dI = \frac{1}{\alpha \Phi_2} \exp \left( - \frac{I + \beta^2 I_q \Phi_1^2}{\alpha \Phi_2} \right) I_0 \left( \frac{2 \Phi_1 \beta \sqrt{II_q}}{\alpha \Phi_2} \right) dI; \quad (57)$$

centrosymétrique:

$$\begin{aligned} p(I) dI &= (8\pi\alpha\Phi_2 I)^{-\frac{1}{2}} \left\{ \exp \left[ - \frac{(|I - \Phi_1 \beta / I_q|)^2}{2\alpha\Phi_2} \right] \right. \\ &\quad \left. + \exp \left[ - \frac{(|I + \Phi_1 \beta / I_q|)^2}{2\alpha\Phi_2} \right] \right\} dI. \quad (58) \end{aligned}$$

La valeur de  $\bar{I}$  est calculée dans l'Appendice: c'est la même pour les deux cas centrosymétrique et non-centrosymétrique:

$$\bar{I} = \alpha \Phi_2 + \beta^2 \Phi_1^2 I_q. \quad (59)$$

Les paramètres  $\alpha$  et  $\beta$  ont les valeurs suivantes:

$$\begin{aligned} \left. \begin{aligned} \alpha &= 1 \\ \beta &= 0 \end{aligned} \right\} && \text{modèle statistique 1;} \\ \left. \begin{aligned} \alpha &= 1 \\ \beta &= 1 \end{aligned} \right\} && \text{modèle statistique 2;} \\ \left. \begin{aligned} \alpha &= 1 - A_w^2 \\ \beta &= A_w \end{aligned} \right\} && \text{modèle statistique 3;} \\ \left. \begin{aligned} \alpha &= 1 - A_w^2 (N/L) (\bar{q}''^2 / q''^2) \\ \beta &= A_w \end{aligned} \right\} && \text{modèle statistique 4.} \end{aligned}$$

$I_q$  représente  $I'_q$  et  $I''_q$ , pour les modèles 3 et 4.

### Normalisation des lois de distribution

Selon le procédé indiqué dans l'introduction, nous cherchons à décrire la distribution des intensités expérimentales, correspondant à un ensemble  $M$  de nœuds du réseau réciproque, au moyen des lois statistiques (57) et (58). Il convient de séparer deux types d'ensembles: L'ensemble  $M_1$  est formé par des nœuds dans lesquels  $\alpha \Phi_2$  et  $\beta^2 \Phi_1^2 I_q$  sont constants. L'ensemble  $M_2$  est formé par des nœuds dans lesquels  $\alpha \Phi_2$  et  $\beta^2 \Phi_1^2$  sont constants, mais  $I_q$  varie.

En principe il n'est pas difficile de choisir des nœuds du réseau réciproque où  $\alpha$ ,  $\beta$ ,  $\Phi_1$  et  $\Phi_2$  sont constants: ce sont tous les nœuds ayant une valeur de  $|s|$  en commun, si les atomes sont isotropes, et si l'on néglige l'anisotropie des transformées de Fourier des volumes  $v$  et  $w$  des modèles 3 et 4. Au contraire il est difficile de choisir des nœuds où  $I_q$  est constant. Par ailleurs on verra dans les exemples que les lois calculées dans l'hypothèse que  $I_q$  varie rendent beaucoup mieux compte de la distribution des intensités expérimentales que les lois déterminées en supposant que  $I_q$  est constant.

Nous allons considérer d'abord l'ensemble  $M_1$ , qu'on peut traiter facilement, du point de vue mathématique: les résultats ainsi obtenus nous serviront à discuter le cas de l'ensemble  $M_2$ .

### Ensemble $M_1$

Les lois de distribution des intensités (57) et (58) dépendent de  $\alpha \Phi_2$  et de  $\beta^2 \Phi_1^2 I_q$ : elles ont la même forme dans tous les points de l'ensemble  $M_1$ .

Dans un cas pratique, où on connaît les intensités seulement en valeur relative, il est commode de normaliser les lois de distribution en remplaçant  $I$  par  $I/\bar{I}$  (Howells, Phillips & Rogers, 1950). On définit

$$x = I/\bar{I}, \quad (60)$$

$$K = \beta^2 \Phi_1^2 I_q / \alpha \Phi_2. \quad (61)$$

La valeur de  $\bar{I}$  est donnée par l'équation (59).

En remplaçant  $x$  et  $K$  dans les équations (57) et (58), on obtient respectivement:

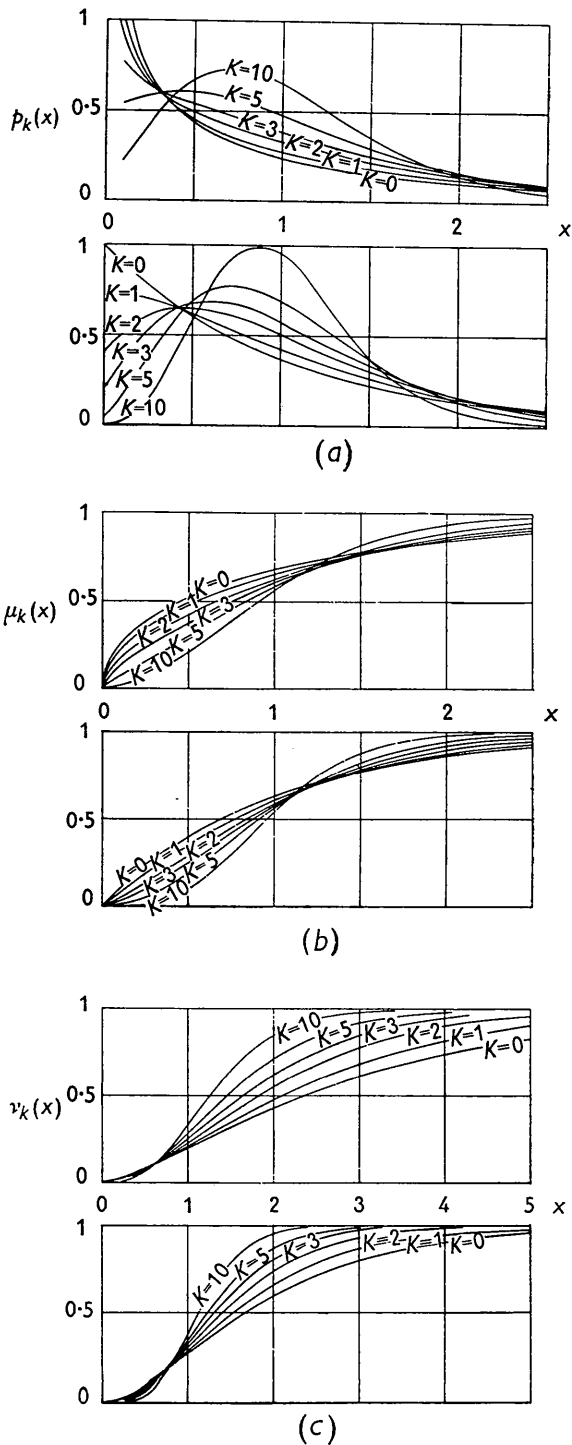


Fig. 2. (a) Familles de courbes  $p_K(x)$ , (b) familles de courbes  $\mu_K(x)$ , (c) familles de courbes  $\nu_K(x)$ , correspondant au cas non-centrosymétrique (inférieures) et centrosymétrique (supérieures).

$$p_K(x) dx = \exp[-x - K(1+x)] I_0(2\sqrt{xK(1+K)}) (1+K) dx \quad (62)$$

dans le cas non-centrosymétrique, et

$$p_K(x) dx = \sqrt{\frac{1+K}{2\pi x}} \times \frac{1}{2} \left\{ \exp\left[-\frac{1}{2}(\sqrt{x(1+K)} - \sqrt{K})^2\right] + \exp\left[-\frac{1}{2}(\sqrt{x(1+K)} + \sqrt{K})^2\right] \right\} dx \quad (63)$$

dans le cas centrosymétrique.

Les équations (62) et (63) peuvent être représentées par deux familles de courbes,  $p$  en fonction de  $x$ , chaque courbe correspondant à une valeur du paramètre  $K$  (Fig. 2(a)). Pour  $K = 0$  on a les équations obtenues par Wilson (1949):

$$p_0(x) dx = e^{-x} dx, \quad (64)$$

$$p_0(x) dx = (2\pi x)^{-\frac{1}{2}} e^{-\frac{1}{2}x} dx. \quad (65)$$

Si  $K$  tend vers l'infini (62) et (63) tendent vers une fonction dont tout le contenu (égal à l'unité) est concentré dans le point  $x = 1$ . Dans les Figs. 2(b) et 2(c) nous avons porté les moments d'ordre zéro

$$\mu_K(x) = \int_0^x p_K(x) dx \quad (66)$$

et les premiers moments

$$\nu_K(x) = \int_0^x x p_K(x) dx \quad (67)$$

des fonctions (62) et (63). Les fonctions (66) et (67) ont été calculées par une intégration graphique. Nous discuterons ci-dessous l'emploi de ces moments (voir chapitre sur les applications).

*Ensemble  $M_2$*

On cherche à établir la loi de distribution des intensités dans les points d'un ensemble  $M_2$  où  $I_q$  n'est pas constant: on peut ramener ce cas au cas précédent. En effet, on peut décomposer l'ensemble  $M_2$  en un certain nombre d'ensembles partiels  $M_1$ , dans tous les points desquels  $I_q$  a la même valeur: dans chaque ensemble  $M_1$  la loi de distribution des intensités est  $p(I)$  (57, 58) définie ci-dessus. La loi de distribution dans l'ensemble  $M_2$  est donc égale à la moyenne de  $p(I)$  dans tous les points de  $M_2$ . De même les moments d'ordre zéro et un des lois de distribution sont égaux aux valeurs moyennes des moments (66) et (67) dans tous les points de  $M_2$ .

Normalisons les lois de distribution d'une manière analogue à celle employée dans le cas précédent. Nous indiquons par une barre les valeurs moyennes de  $I_q$  et de  $K$  dans tous les points de l'ensemble  $M_2$ .

$$x' = (I/\bar{I}) = I/(\alpha\bar{\Phi}_2 + \beta^2\bar{\Phi}_1^2\bar{I}_q), \quad (68)$$

$$\bar{K} = \beta^2\bar{\Phi}_1^2\bar{I}_q/\alpha\bar{\Phi}_2. \quad (69)$$

Soit  $g(K)$  la loi de distribution de  $K$  dans  $M_2$ . Les contributions à la loi de distribution de  $I$  et à ses moments d'une fraction  $g(K)dK$  de points de  $M_2$ , où  $K$  a une valeur comprise entre  $K$  et  $K+dK$ , sont:

$$p_K(x)g(K)dK, \quad (70)$$

$$\mu_K(x)g(K)dK, \tag{71}$$

$$\nu_K(x)g(K)dK, \tag{72}$$

où  $p_K(x)$ ,  $\mu_K(x)$  et  $\nu_K(x)$  sont les fonctions (62) ou (63), (66) et (67).

Puisque nous voulons déterminer les lois de distribution en fonction de  $x'$  (68), nous remplaçons  $x$  (60) par son expression en fonction de  $x'$  dans (70), (71) et (72).

$$I/\alpha\Phi_2 = x(1+K) = x'(1+\bar{K}), \tag{73}$$

$$x = x'(1+\bar{K})/(1+K). \tag{74}$$

On calcule les nouvelles lois correspondant à l'ensemble  $M_2$  en intégrant pour toutes les valeurs de  $K$ :

$$p_{g(K)}(x') = \int_0^\infty p_K\left(x' \frac{1+\bar{K}}{1+K}\right) g(K) dK, \tag{75}$$

$$\mu_{g(K)}(x') = \int_0^\infty \mu_K\left(x' \frac{1+\bar{K}}{1+K}\right) g(K) dK, \tag{76}$$

$$\nu_{g(K)}(x') = \int_0^\infty \frac{1+K}{1+\bar{K}} \nu_K\left(x' \frac{1+\bar{K}}{1+K}\right) g(K) dK. \tag{77}$$

Le facteur  $(1+K)/(1+\bar{K})$  de l'intégrale (77) met à la même échelle les contributions des groupes de points ayant différentes valeurs de  $K$ , et effectue la normalisation (quand  $x'$  tend vers l'infini,  $\nu_{g(K)}(x')$  tend vers un).

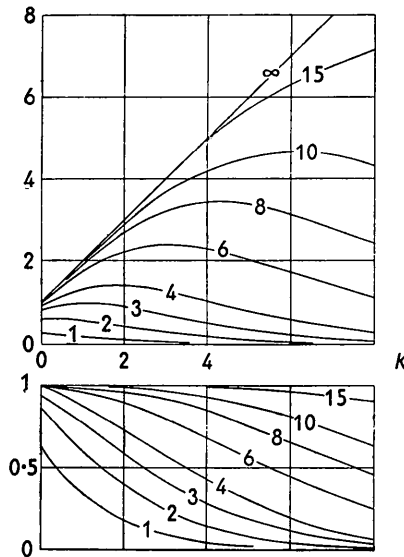


Fig. 3. Familles de courbes  $\mu_K\left(x' \frac{1+\bar{K}}{1+K}\right)$  (inférieures) et  $(1+K)\nu_K\left(x' \frac{1+\bar{K}}{1+K}\right)$  (supérieures). Les valeurs de  $x'(1+\bar{K})$  sont indiquées sur les courbes.

On peut donner une interprétation graphique de l'intégration avec laquelle on obtient les fonctions (76) et (77). Traçons la famille des courbes (Fig. 3),

$$\mu_K\left(x' \frac{1+\bar{K}}{1+K}\right) \tag{78}$$

et

$$(1+K)\nu_K\left(x' \frac{1+\bar{K}}{1+K}\right). \tag{79}$$

Chaque courbe de la Fig. 3 correspond à une valeur déterminée de  $x'(1+\bar{K})$ , et représente les fonctions (78) et (79) en fonction de  $K$ . On obtient la valeur de (76) et (77) en un point  $x'$  en déterminant la valeur moyenne des fonctions (78) et (79) correspondant à  $x'(1+\bar{K})$ ,  $g(K)$  étant le poids à attribuer à chaque point  $K$ : dans le cas de  $\nu_{g(K)}(x')$  il faut en outre diviser par  $(1+\bar{K})$ .

Essayons de déterminer quelques propriétés des fonctions  $\mu_{g(K)}(x')$  (76) et  $\nu_{g(K)}(x')$  (77). Si  $K$  a seulement la valeur  $\bar{K}$ ,  $\mu_{g(K)}(x')$  et  $\nu_{g(K)}(x')$  coïncident avec les fonctions des familles (66) et (67) qui correspondent à  $K = \bar{K}$ :  $\mu_{\bar{K}}(x')$  et  $\nu_{\bar{K}}(x')$ . Admettons qu'en général  $g(K)$  a une valeur non nulle seulement dans le domaine fini  $K_1 < K < K_2$

$$\int_{K_1}^{K_2} g(K) dK = \int_0^\infty g(K) dK = 1. \tag{80}$$

Si la courbe de la Fig. 3, correspondant à une valeur de  $x'(1+\bar{K})$ , était une droite dans le domaine  $K_1-K_2$ , la valeur de  $\mu_{g(K)}(x')$  (ou de  $\nu_{g(K)}(x')$ ) serait égale à  $\mu_{\bar{K}}(x')$  (ou  $\nu_{\bar{K}}(x')$ ), au point  $x'$ . En effet, en remplaçant  $\mu_K\left(x' \frac{1+\bar{K}}{1+K}\right)$  par les deux premiers termes de son développement en série de Taylor, on a:

$$\begin{aligned} \mu_{g(K)}(x') &= \int_{K_1}^{K_2} \mu_K\left(x' \frac{1+\bar{K}}{1+K}\right) g(K) dK \\ &= \int_{K_1}^{K_2} [\mu_{\bar{K}}(x') + a(K-\bar{K})] g(K) dK = \mu_{\bar{K}}(x'). \end{aligned} \tag{81}$$

Il est évident que si dans le domaine  $K_1-K_2$  une courbe de la Fig. 3 se maintient respectivement au-dessous ou au-dessus de la droite qui lui est tangente au point  $\bar{K}$ , la valeur de  $\mu_{g(K)}(x')$  (ou  $\nu_{g(K)}(x')$ ) est plus petite ou plus grande que celle de  $\mu_{\bar{K}}(x')$  (ou  $\nu_{\bar{K}}(x')$ ). Or on remarque que toutes les courbes de la Fig. 3, correspondant aux valeurs de  $x'(1+\bar{K})$  plus petites qu'un certain maximum, se trouvent au dessus de la droite qui leur est tangente au point  $\bar{K}$ , et que toutes les courbes où  $x'(1+\bar{K})$  est plus grand qu'un certain minimum se trouvent au dessous de la tangente, quelle que soit l'étendue du domaine  $K_1-K_2$ . En conséquence l'allure d'une fonction  $\mu_{g(K)}(x')$  (ou  $\nu_{g(K)}(x')$ ) quelconque, par rapport à la fonction de la famille (66) ou (67) correspondant à  $\bar{K}$  est la suivante:

- (a) pour  $x'$  petit  $\mu_{g(K)}(x')$  est plus grand que  $\mu_{\bar{K}}(x')$ ;
- (b) quand  $x'$  croît  $\mu_{g(K)}(x')$  et  $\mu_{\bar{K}}(x')$  se croisent, et  $\mu_{g(K)}(x')$  devient plus petit que  $\mu_{\bar{K}}(x')$ ;

(c) pour des grandes valeurs de  $x'$  l'écart entre  $\mu_{g(K)}(x')$  et  $\mu_{\bar{K}}(x')$  diminue, et les deux courbes tendent à coïncider.

Les fonctions  $\mu_{g(K)}(x')$  et  $\nu_{g(K)}(x')$  ont une allure analogue, mais non identique (voir Fig. 3): les régions de  $x'$  où  $\mu_{g(K)}(x')$  est plus grand ou plus petit que  $\mu_{\bar{K}}(x')$  ne coïncident pas nécessairement avec celles où  $\nu_{g(K)}(x')$  est plus grand ou plus petit que  $\nu_{\bar{K}}(x')$  (sauf pour  $x'$  très petit ou très grand, où  $\mu_{g(K)}(x')$  et  $\nu_{g(K)}(x')$  sont respectivement plus grand et plus petit que  $\mu_{\bar{K}}(x')$  et  $\nu_{\bar{K}}(x')$ ). Pour une valeur donnée de  $\bar{K}$  l'écart entre  $\mu_{g(K)}(x')$  et  $\mu_{\bar{K}}(x')$ , et entre  $\nu_{g(K)}(x')$  et  $\nu_{\bar{K}}(x')$ , est d'autant plus grand que la loi de distribution de  $K$  est plus étendue.

Nous avons discuté ci-dessus le cas des structures non-centrosymétriques, qui sont les plus importantes dans l'étude des protéines: les résultats qu'on obtient dans le cas des structures centrosymétriques sont analogues. En particulier l'allure relative des courbes  $\mu_{g(K)}(x')$  et  $\mu_{\bar{K}}(x')$ , et  $\nu_{g(K)}(x')$  et  $\nu_{\bar{K}}(x')$ , est la même dans les deux cas.

### Applications et exemples

#### Propriétés statistiques des intensités

Résumons et comparons les propriétés statistiques des intensités correspondant aux quatre modèles définis dans l'introduction.

Dans le modèle 1  $\langle I \rangle$  coïncide avec  $\Phi_2$ , les lois de distribution des intensités sont données par les équations (3) et (4).

Le modèle 2 est compatible avec une allure moins régulière de  $\langle I \rangle$  en fonction de  $|s|$ : il impose néanmoins à  $\langle I \rangle$  la condition d'être partout égal ou plus grand que  $\Phi_2$ :

$$\langle I \rangle = \Phi_2 + \Phi_1^2 \langle I_q \rangle. \quad (82)$$

Nous avons discuté dans le chapitre précédent l'allure des lois de distribution et de leurs moments, pour les deux cas où  $I_q$  est constant et varie.

Les modèles 3 et 4 imposent encore moins de restrictions à la forme de la courbe  $\langle I \rangle$ :  $\langle I \rangle$  et  $\Phi_2$  sont différentes, et leur position relative peut être n'importe laquelle:

$$\langle I \rangle = \Phi_2(1 - A_v^2) + \Phi_1^2 \langle I_q' \rangle A_v^2 \quad \text{modèle 3,} \quad (83)$$

$$\langle I \rangle = \Phi_2 \left( 1 - \frac{N}{L} \frac{q''^2}{q'^2} A_w^2 \right) + \Phi_1^2 \langle I_q'' \rangle A_w^2 \quad \text{modèle 4.} \quad (84)$$

Examinons ces deux équations.  $A_v^2$  dans (83), est égal à 1 quand  $|s| = 0$ : pour  $|s|$  croissant  $A_v^2$  tend vers 0 plus vite que  $\Phi_2$ .\* Donc la courbe  $\langle I \rangle$  correspondant

au modèle 3 est inférieure à celle du modèle 2: la différence est grande pour  $|s|$  petit, et s'estompe rapidement avec  $|s|$  croissant.

Le terme  $\{1 - (N/L)(q''^2/q'^2)A_w^2\}$  de l'équation (84) dépend de deux facteurs indépendants de  $|s|$ , à savoir le 'degré de remplissage'  $N/L$  des cases entre lesquelles on a partagé la maille élémentaire et les fluctuations  $q''^2/q'^2$  de  $q(\mathbf{r})$  (équation (5)), et du facteur  $A_w$ , fonction de  $|s|$ , qui est la transformée de Fourier du volume d'une case.  $A_w$  est égal à 1, si  $|s| = 0$ , et décroît à peine plus vite que  $\Phi_2$ , pour  $|s|$  croissant.\* Nous avons montré que  $\{(N/L)(q''^2/q'^2)\}$  ne peut pas être plus grand que 1. Donc la courbe  $\langle I \rangle$  correspondant au modèle 4 est intermédiaire entre celles des modèles 2 et 3, dans la région où  $|s|$  est petit; pour  $|s|$  croissant  $\langle I \rangle$  du modèle 4 tend vers  $\langle I \rangle$  du modèle 2 plus lentement que  $\langle I \rangle$  du modèle 3.

Les lois de distribution des intensités normalisées, et leurs moments, ont été discutées au chapitre précédent: elles sont les mêmes pour les trois modèles (2, 3 et 4).

Si  $|s|$  est suffisamment grand,  $\langle I \rangle$  tend vers  $\Phi_2$  et les lois de distribution tendent vers celles de Wilson dans les trois modèles, car le facteur de  $\Phi_1^2$  dans (82), (83) et (84) tend vers 0 plus vite que  $\Phi_2$ , si l'on admet que les variations de  $q(\mathbf{r})$  sont lentes dans le volume qu'occupe un atome.

#### Détermination de $\Phi_2$ et détection d'un centre de symétrie

Il est peut-être opportun de faire quelques remarques relatives à l'application de certaines méthodes statistiques dérivées des travaux de Wilson aux intensités des protéines: en particulier la détermination du facteur de température et de l'échelle absolue (Wilson, 1942), et la détection des centres de symétrie (Howells *et al.*, 1950), à partir des intensités, connues en valeur relative.

Le premier de ces problèmes équivaut à déterminer la courbe  $\Phi_2$ , à partir de  $\langle I \rangle$ . La méthode la plus directe serait celle d'utiliser la portion de  $\langle I \rangle$  correspondant à  $|s|$  grand, où  $\langle I \rangle$  et  $\Phi_2$  coïncident, pour extrapoler vers l'origine (Wilson, 1942). En pratique ce procédé est affecté de graves erreurs, car l'intensité moyenne devient très faible, lorsque  $|s|$  croît, et les valeurs de  $\langle I \rangle$  sont peu précises. Un moyen différent est celui d'utiliser d'autres données expérimentales: nous en donnons un exemple dans le cas de la ribonucléase (voir ci-dessous). L'hypothèse que  $\langle I \rangle$  ne s'écarte pas de  $\Phi_2$  (Corey, Donohue & Trueblood, 1952), dans la région où  $|s|$  est petit, est arbitraire, et peut mener à des conclusions fausses.

Pour déceler la présence d'un centre de symétrie, on utilise habituellement la méthode proposée par Howells *et al.* (1950), qui consiste à comparer les points expérimentaux  $\mu(x')$  (voir (86) ci-dessus) avec les courbes théoriques  $\mu_{K=0}(x')$  des cas centrosymétrique et non-centrosymétrique, dans la région  $x' < 1$ . Or nous avons montré que lorsque  $K$  n'est pas constant,

\* Le parallélépipède (voir modèle 3) dont  $A_v$  est la transformée de Fourier est grand, en comparaison des dimensions d'un atome; au contraire le volume d'une case (voir modèle 4) dont  $A_w$  est la transformée de Fourier est du même ordre de grandeur de celui occupé par un atome.



les points expérimentaux  $\mu(x')$  se trouvent au-dessus de la courbe  $\mu_{\bar{K}}(x')$ , dans la région où  $x'$  est petit: il est possible, dans ces conditions, que les points  $\mu(x')$  d'une structure non-centrosymétrique atteignent ou dépassent la courbe  $\mu_{\bar{K}=0}(x')$  des structures centrosymétriques. Cela est d'autant plus probable que la variation soit de  $I_q$  soit de  $\Phi_2$  est plus grande; en effet on peut montrer sans difficulté (voir le chapitre sur l'ensemble  $M_2$ ) que la variation de  $\Phi_2$  dans un ensemble  $M_2$  est formellement équivalente à la variation de  $I_q$ .

#### Exemple: ribonucléase

Nous avons choisi cette protéine comme exemple car nous disposons d'un ensemble de mesures expérimentales faites au 'Protein Structure Project' (Magdoff, Crick & Luzzati, 1956) au moyen d'une technique perfectionnée (Furnas & Harker, 1955): Magdoff & Crick (1955) ont discuté la précision de ces mesures.

Le cristal dont nous avons utilisé les intensités appartient au système monoclinique, groupe de symé-

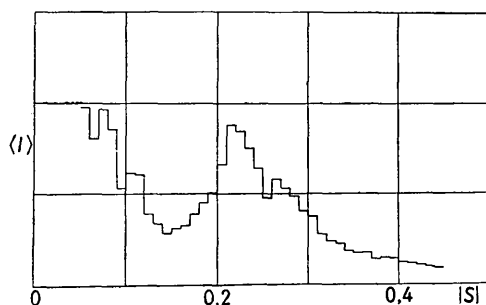


Fig. 4.  $\langle I \rangle$  de la ribonucléase. Chaque palier d'étend sur la région de  $|s|$  pour laquelle on a calculé une valeur de l'intensité moyenne.

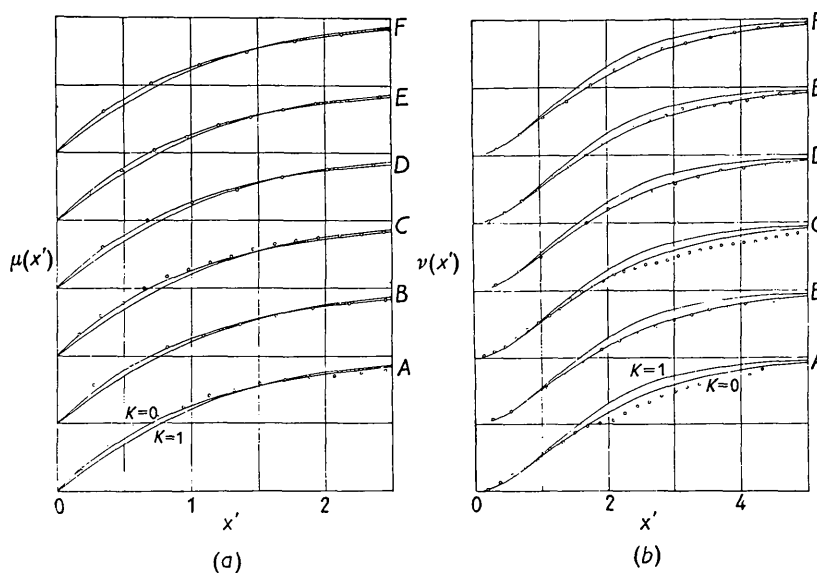


Fig. 5. (a) Fonctions  $\mu(x')$  expérimentales, (b) fonctions  $\nu(x')$  expérimentales: ribonucléase.

trie  $P2_1$ ; les dimensions de la maille élémentaire sont:  $a = 30,24$ ,  $b = 28,43$ ,  $c = 53,21$  Å,  $\beta = 105^\circ 85'$ .

Nous avons porté dans la Fig. 4 la courbe  $\langle I \rangle$  expérimentale. L'allure de  $\langle I \rangle$ , en fonction de  $|s|$ , est incompatible avec le modèle statistique proposé par Wilson, qui impose la condition  $\langle I \rangle = \Phi_2$ . Pour choisir entre les autres modèles celui qui est en meilleur accord avec  $\langle I \rangle$  il faudrait comparer les courbes  $\langle I \rangle$  et  $\Phi_2$ : or nous avons montré ci-dessus qu'il n'est pas facile de déterminer la forme de  $\Phi_2$  à partir des intensités.

Nous avons essayé d'éviter cette difficulté en comparant  $\langle I \rangle$  avec la courbe de diffraction de la ribonucléase amorphe\*: nous supposons pour cela que la distribution des atomes est approximativement la même dans le cristal et dans la poudre amorphe. Cette méthode, telle que nous l'avons employée, ne permet pas de tracer  $\Phi_2$  avec une précision comparable à celle de  $\langle I \rangle$ ; nous avons pu déterminer néanmoins que la courbe  $\Phi_2$  se trouve par endroits au-dessus et par endroits au-dessous de  $\langle I \rangle$ , et que les deux courbes se croisent au moins une fois dans la région  $|s| \approx 0,2$  Å $^{-1}$ . Cela est incompatible avec le modèle 2, selon lequel  $\Phi_2$  devrait se trouver partout au-dessous de  $\langle I \rangle$ : le choix entre les modèles 3 et 4 dépend d'une meilleure connaissance de la courbe  $\Phi_2$ .

Nous avons examiné ensuite la distribution des intensités expérimentales. Nous avons effectué l'analyse statistique avec toutes les intensités contenues à l'intérieur d'une sphère de rayon  $|s| = 0,445$  Å $^{-1}$ , ayant son centre à l'origine de l'espace réciproque, à l'exclusion de la strate  $h0l$ ; nous les avons réunies en plusieurs groupes, dont chacun est formé des inten-

\* Nous devons à la complaisance du Dr D. P. Riley et du Dr V. W. Arndt d'avoir pu utiliser les données expérimentales relatives à la ribonucléase amorphe, qu'ils n'ont pas encore publiées.

sités des  $N$  nœuds du réseau réciproque contenus entre deux sphères de rayons  $|s_1|$  et  $|s_2|$ :

Groupe	$ s_1 $	$ s_2 $	$N$
A	0	0,147	173
B	0,115	0,193	320
C	0,193	0,293	1040
D	0,293	0,337	778
E	0,337	0,392	1269
F	0,392	0,445	1694

Pour chaque groupe nous avons calculé l'intensité moyenne  $\bar{I}$ , et nous avons normalisé les intensités:

$$x' = I/\bar{I}. \quad (85)$$

En ordonnant les intensités en fonction de  $x'$  croissant nous avons calculé quelques points expérimentaux des moments  $\mu(x')$  et  $\nu(x')$ :

$$\mu(x') = m(x')/m(\infty), \quad (86)$$

$$\nu(x') = n(x')/n(\infty). \quad (87)$$

$m(x')$  et  $n(x')$  représentent respectivement le nombre de nœuds dans lesquels  $I/\bar{I}$  est plus petit que  $x'$ , et la somme des valeurs de  $x'$  dans ces nœuds. Ces points sont portés dans la Fig. 5 avec quelques courbes des familles  $\mu_K(x)$  (66) et  $\nu_K(x)$  (67), relatives au cas non-centrosymétrique.

La distribution des intensités des groupes A et C est nettement différente des lois de Wilson ( $K = 0$  dans la Fig. 5): l'allure des courbes expérimentales est au contraire en excellent accord avec les lois statistiques relatives aux modèles 2, 3 et 4 pour le cas où  $I_q$  n'est pas constant. En effet les points expérimentaux  $\mu(x')$  et  $\nu(x')$ , qui pour  $x'$  petit se trouvent au-dessus des courbes  $\mu_K(x)$  et  $\nu_K(x)$ , les croisent lorsque  $x'$  croît, passent au-dessous et tendent enfin vers l'une des courbes  $\mu_{\bar{K}}(x')$  et  $\nu_{\bar{K}}(x')$ . Selon la théorie  $\mu(x')$  et  $\nu(x')$  ont  $\mu_{\bar{K}}(x)$  et  $\nu_{\bar{K}}(x)$  pour asymptote: en pratique les derniers points expérimentaux de la Fig. 5 sont encore éloignés de l'asymptote, donc on ne peut pas déterminer la valeur de  $\bar{K}$ .

L'écart entre la distribution des intensités du groupe B et les lois de Wilson n'est pas grand: toutefois son allure est analogue à celle des groupes A et C, quoique beaucoup moins accentuée.

Les intensités des groupes D, E et F sont distribuées selon des lois pratiquement identiques à celles de Wilson.

Or on remarque que les groupes A et C coïncident avec les régions où  $\langle I \rangle$  a ses valeurs les plus grandes (Fig. 4), et que le groupe B, au contraire, correspond à la région où  $\langle I \rangle$  atteint son minimum. Cela est en accord avec la théorie, qui établit que les lois de distribution s'écartent d'autant plus des lois de Wilson que  $K$  (61) est plus grand: il est évident que  $K$  est grand lorsque  $\langle I \rangle$ , donc  $\langle I_q \rangle$ , est grand. Par ailleurs  $K$  tend vers zéro avec  $\langle I_q \rangle$ , quand  $|s|$  croît: cela ex-

plique la distribution des intensités des groupes D, E et F.

#### Exemple: méthémoglobine

Le Dr Perutz ayant obligeamment mis à notre disposition les intensités obtenues avec des cristaux de méthémoglobine de cheval (Perutz, 1949), nous les avons utilisées pour faire une analyse statistique analogue à celle effectuée avec la ribonucléase.

Les cristaux de méthémoglobine appartiennent au système monoclinique, groupe de symétrie  $C2(a=109, b=63,2, c=54,4 \text{ \AA}, \beta=111^\circ)$ .

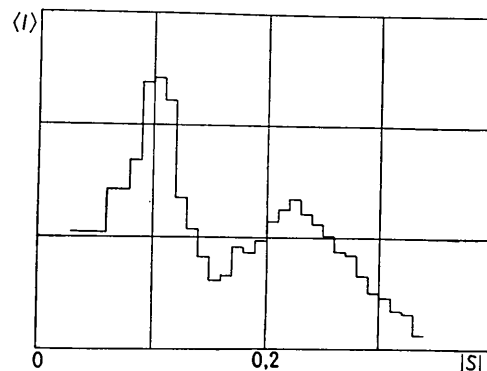


Fig. 6.  $\langle I \rangle$  de la méthémoglobine.

La courbe  $\langle I \rangle$ , représentée dans la Fig. 6, est analogue à celle de la ribonucléase: l'allure de  $\langle I \rangle$  suffit à indiquer le désaccord avec le modèle statistique de Wilson. Ne disposant pas d'autres données expérimentales, nous n'avons pas pu déterminer la position de  $\Phi_2$ .

Nous avons fait l'analyse statistique avec les taches correspondant aux nœuds du réseau réciproque contenus à l'intérieur de la sphère de rayon  $|s| = 0,310$  (les taches  $h0l$  exclues), que nous avons groupées de la manière suivante:

Groupe	$ s_1 $	$ s_2 $	$N$
A	0,080	0,120	200
B	0	0,150	552
C	0,130	0,190	791
D	0,170	0,210	759
E	0,200	0,240	1004
F	0,240	0,280	1447
G	0,280	0,310	1379

Nous avons porté dans la Fig. 7 les points expérimentaux  $\mu(x')$  et  $\nu(x')$  et quelques courbes des familles  $\mu_K(x)$  et  $\nu_K(x)$ .

Dans tous les cas où les points expérimentaux s'écartent des lois de Wilson, ils se rangent par rapport aux courbes  $\mu_K(x)$  et  $\nu_K(x)$  selon les lois établies pour les modèles 2, 3 et 4 (voir chapitre précédent et exemple sur la ribonucléase). En particulier l'allure des points expérimentaux du groupe A fournit la meilleure illustration, entre tous les groupes des deux exemples.

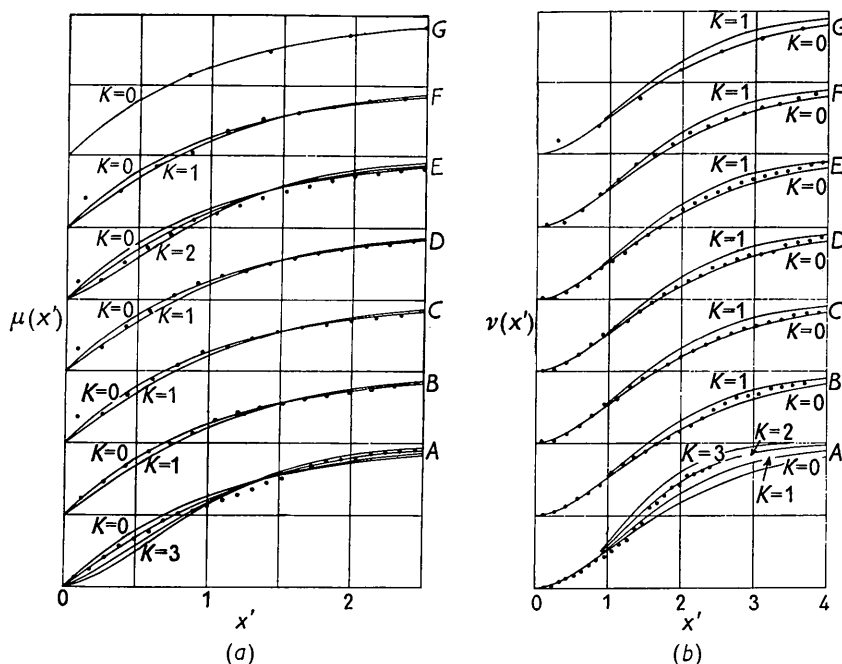


Fig. 7. (a) Fonctions  $\mu(x')$  expérimentales, (b) fonctions  $\nu(x')$  expérimentales: méthémoglobine.

des résultats théoriques du chapitre précédent. Pour la méthémoglobine, comme pour la ribonucléase, l'écart entre points expérimentaux et lois de Wilson est grand pour les groupes dont l'intensité moyenne est grande (voir Fig. 6): on peut remarquer, en particulier, la différence entre la distribution des intensités du groupe A, correspondant au premier maximum de  $\langle I \rangle$ , et celle du groupe C, qui contient le minimum de  $\langle I \rangle$  situé autour de  $|s| = 0,150$ .

J'ai développé ce travail pendant un séjour d'un an au Protein Structure Project. Que son Directeur, le Dr D. Harker, veuille bien trouver ici l'expression de ma profonde reconnaissance pour l'accueil qu'il m'a réservé dans son Laboratoire, et l'intérêt qu'il a porté à mon travail. Je tiens à remercier également le Dr B. S. Magdoff, qui m'a aidé de ses conseils et de ses encouragements. Il m'est enfin particulièrement agréable de remercier le Dr F. H. C. Crick, dont les critiques fort pertinentes ont corrigé de graves erreurs et ont contribué à donner à ce travail sa forme définitive.

## APPENDICE

### Calcul de $\bar{I}$

Dans le cas non-centrosymétrique on a

$$\bar{I} = \iint_{-\infty}^{\infty} (A^2 + B^2) p(A) p(B) dA dB, \quad (88)$$

où  $p(A)$  et  $p(B)$  sont respectivement les équations (12) et (13). Dans le cas centrosymétrique on a

$$\bar{I} = \int_{-\infty}^{\infty} F^2 p(F) dF, \quad (89)$$

où  $p(F)$  est l'équation (28).

On effectue un changement de variable dans les intégrales (88) et (89), de manière à obtenir des intégrales ayant la forme:

$$\bar{I} = \iint_{-\infty}^{\infty} \varphi(u, v) \exp[-u^2 - v^2] du dv, \quad (90)$$

où  $\varphi(u, v)$  est un polynôme en  $u$  et  $v$ .

On calcule (90) en utilisant l'intégrale connue:

$$\int_{-\infty}^{\infty} t^2 \exp[-t^2] dt. \quad (91)$$

Dans les deux cas, non-centrosymétrique et centrosymétrique, on obtient:

$$\bar{I} = \alpha \Phi_2 + (\beta^2 \Phi_1^2 - \alpha \Phi_2) I_q. \quad (92)$$

Nous avons admis que le nombre d'atomes est très grand: dans ce cas  $\Phi_2$  est négligeable en comparaison de  $\Phi_1^2$  et (92) devient

$$\bar{I} = \alpha \Phi_2 + \beta^2 \Phi_1^2 I_q. \quad (59)$$

### References

- COREY, R. B., DONOHUE, J. & TRUEBLOOD, K. N. (1952). *Acta Cryst.* **5**, 701.  
 FURNAS, T. C. & HARKER, D. (1955). *Rev. Sci. Instrum.* **26**, 449.  
 HARKER, D. (1953). *Acta Cryst.* **6**, 731.

- HOWELLS, E. R., PHILLIPS, D. C. & ROGERS, D. (1950). *Acta Cryst.* **3**, 210.
- KARLE, J. & HAUPTMAN, H. (1952). *Acta Cryst.* **5**, 48.
- KARLE, J. & HAUPTMAN, H. (1953). *Acta Cryst.* **6**, 131.
- LUZZATI, V. (1952). *Acta Cryst.* **5**, 802.
- MAGDOFF, B. S. & CRICK, F. H. C. (1955). *Acta Cryst.* **8**, 461.
- MAGDOFF, B. S., CRICK, F. H. C. & LUZZATI, V. (1956). *Acta Cryst.* Sous presse.
- PERUTZ, M. F. (1949). *Proc. Roy. Soc. A*, **195**, 474.
- SAYRE, D. (1951). *Acta Cryst.* **4**, 362.
- WILSON, A. J. C. (1942). *Nature, Lond.* **150**, 152.
- WILSON, A. J. C. (1949). *Acta Cryst.* **2**, 318.

*Acta Cryst.* (1955). **8**, 806

## The Geometry of Lattice Planes. II. Crystallography of Edge Dislocations

BY M. A. JASWON AND D. B. DOVE

*Department of Mathematics, Imperial College, London S.W. 7, England*

(Received 17 June 1955 and in revised form 22 July 1955)

A crystallographic procedure is given for constructing a dislocation of given Burgers vector in any crystal structure. The theory is applied to glissile dislocations in the c.p.h. and diamond structures, and to sessile dislocations in the f.c.c. and b.c.c. lattices.

### Introduction

Given that a dislocation has the form of a straight line lying in a particular lattice plane, referred to hereafter as the slip plane, with a prescribed Burgers vector  $\mathbf{b}$ , where  $\mathbf{b}$  is a lattice vector in this plane, it is well understood that there are two degrees of freedom at our disposal in the construction of the dislocation and three in its crystallographic description. For the direction of the dislocation line may be that of any lattice vector  $\mathbf{d}$  in the slip plane, and the dislocation plane\* may be any lattice plane (other than the slip plane) containing  $\mathbf{d}$ ; if  $\mathbf{p}$  is any lattice vector in the dislocation plane (other than  $\mathbf{d}$ ), a map of the plane  $\mathbf{p}$ ,  $\mathbf{b}$ —referred to hereafter as the elevation plane—provides the most convenient and suggestive picture of the lattice distortion. It is the purpose of the present paper to analyse how the crystallographic construction of the dislocation depends on the choice of the dislocation plane, to show how the description depends on the choice of the elevation plane, and to apply the results to the more general dislocation consisting of a sequence of straight-line segments of differing directions. The analysis is extended to any crystal structure, however complex, and should hence provide the theory of dislocations with a rigorous crystallographic setting which is at present lacking.

### Crystallographic construction

According to the preceding considerations, we set up a triplet of vectors  $\mathbf{b}$ ,  $\mathbf{d}$ ,  $\mathbf{p}$  at any lattice point  $\circ$  of

\* This may be formally defined as the lattice plane parallel to which half-planes are inserted or removed.

the dislocation line:  $\mathbf{b}$  is a given lattice vector in the slip plane;  $\mathbf{d}$  lies in the slip plane, but is otherwise arbitrary, and defines the dislocation line through  $\circ$ ;  $\mathbf{p}$  does not lie in the slip plane, but is otherwise arbitrary, together with  $\mathbf{d}$  defining the dislocation plane and together with  $\mathbf{b}$  defining the elevation plane. The slip properties of the dislocation depend only on  $\mathbf{b}$ . For a given  $\mathbf{b}$ , the construction of the dislocation depends on the choice of the dislocation plane. In the simplest possible case, the vectors  $\mathbf{b}$ ,  $\mathbf{d}$ ,  $\mathbf{p}$  constitute a primitive triplet. This implies that the stacking pattern of the dislocation planes along  $\mathbf{b}$  is

... 1 1 1 1 1 ...;

removal or insertion of a half-plane does not alter this stacking at distances far removed from  $\circ$ , and hence enables us to construct the dislocation in the usual manner (Cottrell, 1953). The stacking pattern of the elevation planes along  $\mathbf{d}$  is

... 1 1 1 1 1 ... ,

so that the map of only one such plane suffices for a complete description of the lattice distortion. Finally, to follow the disturbance along the dislocation line, a map of the slip plane is also necessary.

If  $\mathbf{b}$ ,  $\mathbf{d}$ ,  $\mathbf{p}$  do not constitute a primitive triplet, e.g. the parallelepiped defined by them contains a lattice point in the interior or on one of the sides, it is convenient to consider three distinct possibilities for the situation of this point:

(1) It lies in the parallelogram  $\mathbf{d}$ ,  $\mathbf{p}$ , as illustrated in Fig. 1(a). In this case the stacking of the elevation planes along  $\mathbf{d}$  is