directions (that is, in the directions towards the closest neighbours); therefore, it is not surprising that the three-fold faces are usually absent on the surfaces of growing quasicrystals.

The relationship between the icosahedral quasicrystals and the CsCl structure (*via* A_{-3} cells) was demonstrated above. Now, it is interesting to note that pieces of the (110) atomic plane of CsCl have been suggested by Dong, Dubois, Kang & Audier (1992) as the structural units for the *decagonal* phase. Perhaps this is an explanation for why the icosahedral, decagonal and CsCl-like phases have close compositions in many alloys.

The author is grateful to N. N. Devnina and E. Kuklina for their help in the bibliography search for crystal structures. At different stages of this work, discussions with M. A. Fradkin, R. V. Galiulin, M. Kléman, L. S. Levitov and H.-R. Trebin were very fruitful. The support of l'Université Pierre et Marie Curie (Paris VI), France, where the final part of the work was done, is acknowledged.

References

BERGMAN, G., WAUGH J. L. T. & PAULING, L. (1957). *Acta Cryst.* **10**, 254-259.
BOUDARD, M., DE BOISSIEU, M., JANOT, C., HEGER, G., BEELI, C., NISSEN, H.-U., VINCENT, H., IBBERSON, R., AUDIER, M. & DUBOIS J. M. (1992). *J. Phys. Condens. Matter*, **4**, 10149-10168.
CENZUAL, K., CHABOT, B. & PARTHÉ, E. (1985). *Acta Cryst.* **C41**, 313-319.
COOPER, M. & ROBINSON, K. (1966). *Acta Cryst.* **20**, 614-617.
CORNIER-QUIQUANDON, M., QUIVY, A., LEFEBRE, S., ELKAIM, E., HEGER, G., KATZ, A. & GRATIAS, D. (1991). *Phys. Rev. B*, **44**, 2071-2084.
DMITRIENKO, V. E. (1987). *Pis'ma Zh. Eksp. Teor. Fiz.* **45**, 31-34. Engl. Transl: *JETP Lett.* **45**, 38-42.
DMITRIENKO, V. E. (1990). *J. Phys. (Paris)*, **51**, 2717-2732.
DMITRIENKO, V. E. (1992). *Pis'ma Zh. Eksp. Teor. Fiz.* **55**, 388-391. Engl. transl: *JETP Lett.* **55**, 391-395.
DMITRIENKO, V. E. (1993). *J. Non-Cryst. Solids*, **153&154**, 150-154.
DONG, C., DUBOIS, J. M., KANG, S. S. & AUDIER, M. (1992) *Philos. Mag. B*, **65**, 107-126.
DUBOIS, J. M., KANG, S. S. & VON STEBUT, J. (1991). *J. Mater. Sci. Lett.* **10**, 537-541.
ELSER, V. & HENLEY, C. L. (1985). *Phys. Rev. Lett.* **55**, 2883-2886.
GUYOT, P., KRAMER, P. & DE BOISSIEU, M. (1991). *Rep. Prog. Phys.* **54**, 1373-1425.
HENLEY, C. L. (1988). *Philos. Mag. Lett.* **58**, 87-89.
HENLEY, C. L. (1991). *Phys. Rev. B*, **43**, 993-1020.
HIRAGA, K., HIRABAYASHI, M., INOUE, A. & MASUMOTO, T. (1985). *J. Phys. Soc. Jpn*, **54**, 4074-4080.
JANOT, C., DUBOIS, J. M., PANNETIER, J., DE BOISSIEU, M. & FRUCHART, R. (1988). *Quasicrystalline Materials*, edited by C. JANOT & J. M. DUBOIS, pp. 107-125. Singapore:World Scientific.
JONES, H. (1960). *Theory of Brillouin Zones and Electronic States in Crystals*. New York: Interscience.
KALUGIN, P. A., KITAEV, A. YU. & LEVITOV, L. S. (1985). *Pis'ma Zh. Eksp. Teor. Fiz.* **41**, 119-121. Engl. transl: *JETP Lett.* **41**, 145-149.
KLÉMAN, M. (1989). *Adv. Phys.* **38**, 605-667.
KURIYAMA, M., LONG, G. G. & BENDERSKY, L. (1985). *Phys. Rev. Lett.* **55**, 849-851.
LANDAU, L. D. & LIFSHITZ, E. M. (1968). *Statistical Physics*, 2nd ed. New York: Pergamon.
MACKAY, A. L. (1981). *Kristallografiya*, **26**, 910-919. Engl. transl: *Sov. Phys. Crystallogr.* **26**, 517-522.
MACKAY, A. L. (1986). *Scr. Metall.* **20**, 1205-1210.
NIIZEKI, K. (1990). *J. Phys. A*, **23**, L1069-L1072.
PENROSE, R. (1979). *Math. Intell.* **2**, 32-37.
POON, S. J. (1992). *Adv. Phys.* **41**, 303-363.
STEURER, W. (1990). *Z. Kristallogr.* **190**, 179-234.
TRESSAND, A., SOUBEYROUX, J. L., TOUHARA, H., DEMAZEAU, G. & LANGLAIS, F. (1981). *Mater. Res. Bull.* **16**, 207-214.
VILLARS, P. & CALVERT, L. D. (1985). *Pearson's Handbook of Crystallographic Data for Intermetallic Phases*, Vols. 1-3. Metal Park, Ohio: American Society of Metals.

Acta Cryst. (1994). A**50**, 526–537

The Heavy-Atom Problem: a Statistical Analysis. I. *A Priori* Determination of Best Scaling, Level of Substitution, Lack of Isomorphism and Phasing Power

BY PHILIPPE DUMAS

UPR de Biologie Structurale, Institut de Biologie Moléculaire et Cellulaire, 15 rue René Descartes, 67084 Strasbourg CEDEX, France

(*Received* 22 *September* 1993; *accepted* 7 *February* 1994)

Abstract

The classical problem of determining heavy-atom parameters in single or multiple isomorphous replacement methods is reconsidered in two related papers. This first paper systematically examines how to derive *a priori* statistical information concerning heavy atoms and lack of isomorphism (LOI). By *a priori* is meant without any knowledge other than that of the measured intensities (and their estimated σ's) of a 'native' and 'derivative' crystal pair, that is to say *before* any potential site of substitution has been determined. First, both the terms $\Sigma_H = \sum_{i=1}^{N} f_i^2$, where f_i is the scattering factor of the ith heavy atom and N is the number of sites and, simultaneously, the best scale factor between the 'native' and 'derivative' data are estimated *a priori* as

a function of $\sin\theta/\lambda$. It is then shown how to derive a quantitative estimate of the respective contributions to Σ_H of 'signal' due to heavy atoms and 'noise' from LOI. The actual heavy-atom contribution is obtained from estimates of both a global isotropic temperature factor and a global absolute occupancy factor. The noise contribution is obtained as a 'LOI factor' analogous to a Debye-Waller term as shown by Read [*Acta Cryst.* (1990), **A46**, 900-912; *Crystallographic Computing* 5 (1991), edited by D. Moras, A. D. Podjarny & J. C. Thierry, pp. 158-167. IUCr/Oxford Univ. Press]. As an important consequence, the variation with resolution of both the 'lack of closure' and derivative phasing power can be estimated.

1. Notation

The following notation is used in this paper:

$P(X)$	Probability density that X lies between X and $X + \mathrm{d}X$
$P(X\|Y)$	Conditional probability density that X lies between X and $X + \mathrm{d}X$, Y being known
$\langle X\rangle$	Expected values of X calculated with the required probability density
$\overline{X}$	Ensemble average
$\Sigma_T = \sum_{i=1}^{N_T} f_i^2$	Sum of the squares of the scattering factors of atoms of type T [$T = P$ for the native, $T = H$ for the heavy atom(s), $T = PH$ for the derivative, $T = N$ for the dummy atoms modelling the noise due to lack of isomorphism, $T = HN$ for the heavy plus dummy atoms]
$\varphi_{T/T'} = \Sigma_T/\Sigma_{T'}$	
$\mathbf{F}_T$	Structure factor corresponding to the contribution of the atoms of type T (same meanings as above)
F_T	Modulus of $\mathbf{F}_T$
$X = 2F_PF_{PH}/\Sigma_H$	
$Y = F_P^2/\Sigma_H$	
$Z = (F_P^2 + F_{PH}^2)/\Sigma_H$	
$J_n(X)$	Bessel function of the first kind of the nth order
$I_n(X) = \exp[-in(\pi/2)] \times J_n(iX)$	Modified Bessel function of the nth order
$G(X) = XI_1(X)/I_0(X)$	
$F(\alpha,\beta;X)$	Confluent (or degenerate) hypergeometric function
B_N	'LOI factor', analogous to a Debye-Waller factor for the quantification of the contribution from LOI to Σ_{HN}
B_H	'Global' isotropic temperature factor of the heavy atoms
$Q_H = (\sum_{j=1}^{N_{\text{site}}} Q_j^2)^{1/2}$	Q_j is the individual absolute occupancy of the jth heavy-atom site (all supposed to be of the same kind)
ε	Lack of closure
$W = \text{r.m.s.}(F_H)/\text{r.m.s.}(\varepsilon)$	Phasing power

2. Introduction

The most common strategy for solving the phase problem in macromolecular crystallography remains that of multiple isomorphous replacement. As a first step, this method requires the determination, as accurately as possible, of the heavy-atom parameters: coordinates, occupancies and temperature factors. To achieve this, the difference Patterson synthesis is most useful. It was first used with the simple differences $(F_{PH}^2 - F_P^2)$ as coefficients at the suggestion of Perutz (1956). Despite its simple interpretation, it has a major drawback owing, in particular, to a high sensitivity to scaling errors. It is well known that the coefficients $(F_{PH} - F_P)^2$, the squares of the isomorphous differences, are much less sensitive to this problem (Rossmann, 1960). These terms are intended to represent the F_H^2, necessary for calculating the Patterson function of the heavy atoms alone. This is rigorous for most centric reflections (apart from experimental errors) but only a crude approximation for acentric ones. Furthermore, the seminal hypothesis of isomorphism is often violated in practice, which makes LOI another (and commonly the major) source of noise. In practice, not only the level of substitution (through the occupancy and temperature factors of each site), but also both the best relative scale factor and the level of LOI (through the lack of closure) are estimated at the refinement step of the heavy-atom parameters. It is commonplace to recall the important correlation between the three terms occupancy, temperature factor and relative scale factor and, therefore, the inherent oscillatory character of their refinement. More subtle is the slow drift of the solution, particularly in cases of a high level of LOI, owing to the lack of closure being taken as a refinable parameter (Bricogne, 1991). This paper addresses several questions related to these problems from the opposite point of view, namely obtaining information *prior to the determination of any site*

of substitution. This information concerns the relative scale factor, the level of substitution, the importance of LOI and, consequently, the lack of closure and the phasing power. Such results could therefore be used to circumvent the factor of circularity introduced by their determination at the refinement step.

3. Lack of isomorphism: previous results and some comments

In many instances, where it is not essentially due to modification of thermal agitation, LOI may be modelled with additional dummy atomic 'dipoles', *i.e.* with atoms occuring in pairs with opposite occupancies. The orientation of these dipoles can vary from total randomness to complete lack of randomness depending on whether there is alteration of the molecular structure and/or rigid-body movement and/or unit-cell modification. Such dipoles have a contribution to Σ_H increasing with resolution, which is the mark of LOI. Indeed, one can easily show that the scattering factor of such a dipole is $i2\pi f(\mathbf{h})\mathbf{h}\cdot\mathbf{u}$ with $i^2 = -1$, $f(\mathbf{h})$ the common scattering factor of the two atoms for a reflection $\mathbf{h}$ and $\mathbf{u}$ the dipole interatomic vector (this expression is valid for small values of $|\mathbf{u}|$ relative to $d = 1/|\mathbf{h}|$). A common hypothesis on this contribution from LOI is to consider it as noise, *i.e.* that it is unrelated to both the heavy-atom and native structure factors. In fact, Read (1990, 1991) recognized that there exists a component of the error due to LOI negatively correlated with the native structure factor. That is to say, on average,

$$\mathbf{F}_{PH} = D\mathbf{F}_P + \mathbf{F}_H + \mathbf{F}_N, \tag{1}$$

where D is a positive factor whose value is 1 at low resolution and decreases with resolution and $\mathbf{F}_N$ is the contribution of the true noise due to LOI. An extremely simple explanation of this fact is as follows: let us consider two hypothetical non-isomorphous crystals of the same molecule, the LOI being essentially due to some structural changes without significant modification for thermal agitation. If $\mathbf{F}_P$ and $\mathbf{F}'_P$ are their respective Fourier transforms, one has

$$\mathbf{F}'_P = \mathbf{F}_P + \mathbf{F}_N. \tag{2}$$

But from the Parseval theorem $\overline{F'^2_P} = \overline{F^2_P}$ and, therefore, the noise contribution $\mathbf{F}_N$ cannot be considered as uncorrelated with $\mathbf{F}_P$, which would lead to $\overline{F^2_P} = \overline{F^2_P} + \overline{F^2_N}$.* On the contrary, if the cosine term D is included, one correctly obtains

$$\overline{F'^2_P} = \overline{D^2F^2_P} + \overline{F^2_N}. \tag{3}$$

If $\overline{F'^2_P} = \overline{F^2_P}$ holds, and since the D's and F_P's are uncorrelated, this leads to

$$\overline{F^2_N} = \overline{(1 - D^2)F^2_P} = (1 - D^2)\Sigma_P. \tag{4}$$

This result is identical to one obtained in a different manner by Read (1990). Because of the statistical independence of $\mathbf{F}_H$ *versus* $\mathbf{F}_P$ and $\mathbf{F}_N$, it is clear that replacing $\mathbf{F}'_P$ and $\overline{F'^2_P}$ by $\mathbf{F}_{PH}$ and F_{PH}, only leads to adding $\mathbf{F}_H$ and $\overline{F^2_H}$ to the right-hand sides of, respectively, (2) and (3).

Two other facts emerge from previous work. First (Luzzati, 1952; Read, 1990), the term $D(s)$ can be understood as a Debye–Waller term, that is to say

$$D(s) = \exp(-B_N s^2) \tag{5}$$

and, when LOI is due to structural changes, B_N may be related to the mean square deviation $\langle\Delta\mathbf{r}^2\rangle$ by

$$B_N = (8\pi^2/3)\langle\Delta\mathbf{r}^2\rangle. \tag{6}$$

Second (Read, 1991), even in the case of LOI requiring a set of nonrandom 'atomic dipoles', *e.g.* for lattice change or rigid-body movement, the noise usually follows a Gaussian distribution. That, in such situations, the latter could be essentially anisotropic is readily dealt with by replacing the scalar B_N by a tensor as for anisotropic agitation or disorder. Interestingly, such a tensor could be theoretically determined for simple cases of LOI by following the spirit of the paper by Crick & Magdoff (1956). In the present paper, we do not consider this aspect but rather tackle the practical determination of B_N as a scalar. The basis of the method for the determination of B_N lies in the *a priori* determination of $\overline{F^2_H} + \overline{F^2_N} = \Sigma_H + \Sigma_N = \Sigma_{HN}$ in shells of resolution. This is examined in §§ 4 and 5.

4. Determination of Σ_H

It is shown in the following how acentric and centric estimates of Σ_H can be obtained in shells of resolution for an ideal isomorphous case.* The following considerations are an analytical version of the numerical method proposed by Nixon & North (1976). Interestingly, Read (1986) showed by numerical tests that it leads to the best results when compared with other methods. It is

* In other words, the noise power, $\overline{F^2_N}$, increases at the expense of the original signal power, $\overline{F^2_P}$. This interpretation is the only valid one, otherwise passing from a perfectly isomorphous crystal pair to a fully nonisomorphous pair would never make their mutual correlation vanish.

* The following results on the determination of Σ_H, the derivative-to-native scale factor and the global quantities Q_H and B_H were presented at the Crystallographic Computing School held at Bischenberg, France, in 1990 (Dumas, 1991).

shown in the next section how the influence of LOI can be dealt with.

4.1. *Determination of* Σ_H *from acentric reflections*

The method is based on calculating $\langle (F_{PH} - F_P)^2 \rangle$ for any given value of F_P, that is to say

$$\langle (F_{PH} - F_P)^2 \rangle = \int_0^\infty (F_{PH} - F_P)^2 P(F_{PH}|F_P) \mathrm{d}F_{PH}. \tag{7}$$

This quantity* can be calculated by considering for $P(F_{PH}|F_P)$ the conditional density probability from Sim's (1959) paper,

$$P(F_{PH}|F_P) = 2(F_{PH}/\Sigma_H) \exp[-(F_P^2 + F_{PH}^2)/\Sigma_H] \times I_0(2F_P F_{PH}/\Sigma_H). \tag{8}$$

Developing the squared difference in (7), one obtains

$$\langle (F_{PH} - F_P)^2 \rangle = \langle F_{PH}^2 \rangle - 2F_P \langle F_{PH} \rangle + F_P^2. \tag{9}$$

The two integrations, for $\langle F_{PH}^2 \rangle$ and $\langle F_{PH} \rangle$, respectively, are detailed in the Appendix and give†

$$\langle F_{PH}^2 \rangle = F_P^2 + \Sigma_H, \tag{10}$$

which is equivalent to

$$\langle F_{PH}^2 - F_P^2 \rangle = \Sigma_H, \tag{11}$$

since the average is calculated for a fixed value of F_P and

$$\langle F_{PH} \rangle = \tfrac{1}{2}(\pi \Sigma_H)^{1/2} F(-\tfrac{1}{2}, 1; -Y), \tag{12}$$

where $F(-\frac{1}{2}, 1; -Y)$ is the confluent hypergeometric function and Y is the reduced variable F_P^2/Σ_H. Putting everything together, one obtains

$$\langle (F_{PH} - F_P)^2 \rangle = \Sigma_H \Gamma_a(Y) \tag{13}$$

with

$$\Gamma_a(Y) = 1 + 2Y - (\pi Y)^{1/2} F(-\tfrac{1}{2}, 1; -Y), \tag{14}$$

where the index a denotes acentric reflections. The function $\Gamma_a(Y)$ is plotted in Fig. 1. The resulting curve shows the existence of two domains: first $0 \leq Y \leq 0.91$, in which the function varies rapidly, and, second, $0.91 \leq Y$, in which it is nearly constant and goes asymptotically to $\frac{1}{2}$. This feature has the following interesting consequence: if we consider all acentric reflections in a given shell of resolution with F_P greater than, say, the average value of F_P in this range, then Σ_H can be approximated by $2\overline{(F_{PH} - F_P)^2}$ in this shell. This first (and well known) approximation, allowing calculation of Y for any reflection, can then be used in a cyclic procedure to recalculate Σ_H as $\overline{(F_{PH} - F_P)^2}/\Gamma_a(Y)$ [from (13)] until convergence.

It can be noted that this result is a 'rigorous' derivation (as far as the underlying statistical premises hold) replacing the common approximation $\Sigma_H = 2\overline{(F_{PH} - F_P)^2}$.

4.2. *Determination of* Σ_H *from centric reflections*

We are still concerned with the average value $\langle (F_{PH} - F_P)^2 \rangle$ for a given value of F_P. The previous statistical treatment could be used by replacing the acentric probability density by the centric one. It is equally possible to consider the different values of $(F_{PH} - F_P)^2$ as a function of F_P and F_H. First, $\mathbf{F}_P$ and $\mathbf{F}_H$ can be of the same or opposite orientation with a probability of $\frac{1}{2}$ for each. Then, in the case of the same orientation, $(F_{PH} - F_P)^2 = F_H^2$ but, in the case of an opposite orientation, $(F_{PH} - F_P)^2 = F_H^2$ if $F_P \geq F_H$ and $(F_{PH} - F_P)^2 = (2F_P - F_H)^2$ if $F_P \leq F_H$. Therefore, from the probability density given by Wilson (1949) for centric terms,

$$P(F_H) = (2\pi \Sigma_H)^{-1/2} \exp(-F_H^2/2\Sigma_H), \tag{15}$$

the average value sought can be written as

$$\langle (F_{PH} - F_P)^2 \rangle = \tfrac{1}{2}\left[\langle F_H^2 \rangle + \int_0^{F_P} F_H^2 P(F_H) \mathrm{d}F_H + \int_{F_P}^\infty (2F_P - F_H)^2 P(F_H) \mathrm{d}F_H \right]. \tag{16}$$

These integrals are easily calculated and we eventually obtain

$$\langle (F_{PH} - F_P)^2 \rangle = \Sigma_H \Gamma_c(Y) \tag{17}$$

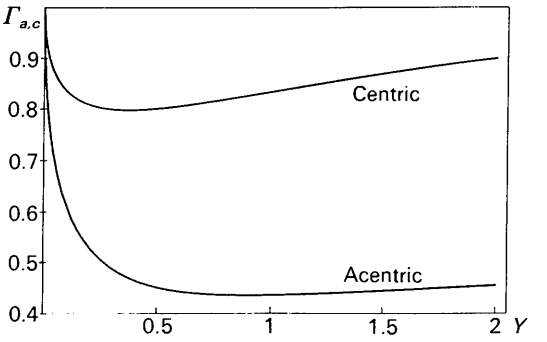

Fig. 1. variation *versus* Y of $\Gamma_a(Y)$ for acentric terms [(14)] and of $\Gamma_c(Y)$ for centric terms [(18)]. $\Gamma_a(Y)$ tends asymptotically towards $\frac{1}{2}$.

* In fact, this quantity should be more precisely denoted $\langle (F_{PH} - F_P)^2 \rangle_{F_P}$, to recall that the average is calculated for a fixed value of F_P. However, the suffix F_P has been dropped for the sake of simplicity.

† Indeed, this result is more a verification than anything new, since it can be obtained *a priori* by using the Parseval theorem.

with

$$\Gamma_c(Y) = 1 + \{1 - \text{erf}\,[(Y/2)^{1/2}]\}Y - [(2/\pi)Y\exp(-Y)]^{1/2}, \quad (18)$$

where the index c denotes centric reflections, Y has the same meaning as in the previous section and erf denotes the error function. The function $\Gamma_c(Y)$ is plotted in Fig. 1. As for acentric terms [(11)], it can be verified that $\langle(F_{PH}^2 - F_P^2)\rangle = \Sigma_H$.

5. Determination of the best derivative-to-native scale factor

5.1. *Case without LOI*

This part is presented separately from the previous considerations only for the sake of clarity. Indeed, both theoretically and practically, determination of Σ_H and of the relative scale factor cannot be dissociated. It has been seen [(11)] that, for centric as well as for acentric reflections, $\langle F_{PH}^2 - F_P^2\rangle = \Sigma_H$. It seems that the simplest way of determining Σ_H would be to consider this estimate. However, in practice, one observes that this often leads to a negative value. This fact is understandable in terms of insufficiently good scaling of the derivative data set *versus* the native one. This corresponds exactly to the problem examined by Rossmann (1960) (see *Introduction*) and to his conclusion that coefficients $(F_{PH}^2 - F_P^2)$ are much more sensitive to scaling problems than $(F_{PH} - F_P)^2$. As a consequence, one can determine the scale factor (either an overall one or per shell of resolution) by the imposition of identical values of Σ_H obtained from $\overline{(F_{PH} - F_P)^2}$ and $\overline{(F_{PH}^2 - F_P^2)}$, respectively. This is achieved in practice by the use of an iterative procedure that converges in two or three cycles.

To test the method, we used synthetic data for which the scale factor is perfectly known (equal to 1). The results are shown in Table 1 and clearly demonstrate its effectiveness. In particular, the difference between this and ordinary methods is dramatic in the case of a high level of substitution (*cf.* the last line of Table 1). In practice, it has been observed that an initial scale factor (to multiply the F_{PH}'s) obtained by other usual methods is most often increased by a few percent. It is worth noting that this method is strongly reminiscent of that using considerations of the heights of two difference Patterson origin peaks (Blundell & Johnson, 1976, p. 335). It was first used (without any explanations) by Kraut, Sieker, High & Freer (1962). Tickle (1991) reconsidered it with more details with the usual approximation $\overline{F_H^2} = 2\overline{(F_{PH} - F_P)^2}$ for acentric reflections.

5.2. *An important comment on the exact meaning of scaling*

It thus appears that the proposed procedure for scaling two data sets amounts to determining the power, *i.e.* $\Sigma_H = \overline{F_H^2}$, of the component of the derivative signal uncorrelated with the native signal. In fact, with it in its more general form, one is led to considering Σ_H as the power of some signal, in fact any signal, provided it is uncorrelated with the native signal. Whether or not it actually originates from heavy atoms is unimportant: it could also be mere noise. This may seem secondary but reveals itself to be crucial for the understanding of the influence of LOI on scaling (see below).

Another important feature is also worth discussing because, at first sight, the theory seems to be inconsistent. Let us consider any theoretical case shown in Table 1 from which the correct unit scale factors were correctly retrieved. Suppose we simply interchange the two data sets, *i.e.* we treat the native data set as the derivative one and *vice versa*. We also expect to obtain unit scale factors, for we implicitly consider that native and derivative data sets should remain correctly scaled whatever the order in which they are taken into consideration. However, the result is contrary to this naïve assumption and we obtain scale factors greater than 1, that is to say the power of the native is increased as if it contained the heavy atoms. The explanation of this apparent paradox is that the proposed procedure only deals with Σ_H, the power of the heavy atoms. Therefore, by the Parseval theorem, there is no way to discriminate heavy atoms with positive or negative occupancy. Thus, there are no inconsistencies, since heavy atoms with positive occupancies in the derivative crystal can be considered formally as atoms in the native crystal but with negative occupancies.

5.3. *How does LOI affect the scaling procedure?*

The answer to this question is deceptively simple: this scaling procedure takes care perfectly of the presence of LOI without any other correction! The reason is simply that this procedure, as stated above, amounts to determining the power of the component of the derivative signal uncorrelated with the native signal. Indeed, there is a component of LOI that is negatively correlated with the native signal, the remaining component being true noise (see § 3). Therefore, the scaling procedure will determine this true noise component of LOI by correctly estimating the term $D(s)$ [(4)] as a contribution to the derivative-to-native scale factor.

It is shown in the next section not only that such interpretation is correct but also how it is possible to 'deconvolute' the respective contributions from LOI, heavy atom(s) and experimental errors to the overall 'noise' making the native and derivative signals different.

5.4. *Extension to the scaling of calculated and observed structure factors*

It is common in practice to have to scale calculated and observed structure factors. The calculated terms may come from an atomic model (with errors and/or incomplete) or even from a density map after modification (*e.g.* after solvent flattening). The present method,

Table 1. *Results of various test calculations with increasing values of* $\varphi_{H/P} = \Sigma_H/\Sigma_P$

Each line of the table corresponds to a given value of $\varphi_{H/P}$. F_P's are calculated with 920 light atoms in the asymmetric unit, F_{PH}'s are calculated with four additional Au atoms whose occupancies and isotropic temperature factors are Q_i and B_i, $i = 1, 4$. $a = 50$, $b = 60$ and $c = 70$ Å. Space group $P2_12_12_1$. The data have been calculated and used up to 3 Å resolution. Lines 1 to 8: $Q_2 = 0.8Q_1$, $Q_3 = 0.5Q_1$, $Q_4 = 0.2Q_1$ and $B_i = 30$ Å^2. Line 9: $Q_i = 1$, $B_i = 0$ Å^2 for $i = 1, 4$. The results are used to assess the accuracy of determination of:

(1) The scaling procedure (one single scale factor: column 6) in comparison with a standard method from the *CCP*4 package (one scale factor: column 5; and isotropic temperature factor: column 4). In all cases, the correct scale factor is 1 with a null temperature factor. It is clearly apparent that the standard method gives increasingly wrong results while the proposed procedure gives very satisfactory results at any level of substitution. Indeed, the overall scale factor (scale factor plus temperature factor) from the standard procedure can be wrong by 55% at the highest resolution in comparison with the maximum error of 1.8% at any resolution for the proposed procedure.

(2) A global isotropic temperature factor B_H for the heavy atom (column 7). The correct values are 30 Å^2 from lines 1 to 8 and 0 Å^2 for line 9.

(3) A global absolute occupancy term Q_H for the heavy atom(s) (column 9 to be compared with column 8).

1	2	3	4	5	6	7	8	9
Line no.	Q_1	$\varphi_{H/P}$	B_{scale} (Å^2)	K_{scale}	K_{scale},* this method	B_H*(Å)	Q_H, exact	Q_H,* estimated
1	0.3	0.025	0.7	0.999	1.002 1.003	30.4 23.9	0.42	0.35 0.40
2	0.4	0.046	1.2	0.996	1.002 1.003	30.4 24.7	0.56	0.48 0.54
3	0.5	0.072	1.7	0.992	1.003 1.004	30.7 25.9	0.69	0.60 0.69
4	0.6	0.103	2.3	0.987	1.003 1.004	30.2 27.6	0.83	0.71 0.85
5	0.7	0.142	2.9	0.980	1.003 1.004	30.4 28.2	0.97	0.84 1.00
6	0.8	0.186	3.5	0.972	1.003 1.004	29.7 27.3	1.11	0.95 1.13
7	0.9	0.236	4.2	0.961	1.003 1.005	31.0 26.8	1.25	1.09 1.27
8	1.0	0.289	4.8	0.948	1.003 1.004	31.2 26.3	1.39	1.21 1.39
9	1.0	1.318	27.0	0.961	0.982 0.992	8.2 5.4	2.00	1.81 2.17

* The upper and lower values correspond to the results obtained with data between 70 and 3 Å and between 10 and 3 Å, respectively.

clearly related to the method of Read (1986), can be used for such purposes with the native and derivative data sets being replaced, respectively, by the calculated and experimental data sets. In view of the comments made in § 5.2, the effects of both the missing atoms (corresponding to Σ_H) and the error on coordinates (corresponding to Σ_N) are obtained as Σ_{HN} by the scaling procedure.

6. *A priori* estimation of the level of LOI and of global heavy-atom parameters

6.1. *Estimation of a 'LOI factor' and of a global heavy-atom temperature factor*

From the preceding sections, one may state that some 'observed' quantity Σ_{obs} can be determined that is the sum of heavy-atom contribution(s) and of various statistically independent errors of experimental origin, Σ_{exp}, and owing to LOI, Σ_N. In other words,

$$\Sigma_{\text{obs}} = \Sigma_H + \Sigma_{\text{exp}} + \Sigma_N. \qquad (19)$$

One takes care of Σ_{exp} by simply subtracting from Σ_{obs} in each shell of resolution the average value of $m[\sigma^2(F_P) + \sigma^2(F_{PH})]$ with $m = 1$ and $m = 2$ for, respectively, centric and acentric reflections. It was verified by a numerical test (not shown) that this correctly leads to $\Sigma_H \simeq 0$ when the F_P's and F_{PH}'s differ only by 'experimental' normal errors. Therefore, for real data, one should insist on the importance of a correct estimate of the uncertainties of experimental origin. This requires not only the correct estimation of $\sigma(I)$ at the data

reduction step but also that $\sigma(F)$ is correctly obtained from $\sigma(I)$.*

In order to separate contributions of both remaining terms, Σ_H and Σ_N, one is led to express them explicitly as a function of $s = \sin\theta/\lambda$. Although the heavy atoms may be too few to allow the safe application of Wilson statistics (Wilson, 1942), it is difficult to avoid doing so. Therefore, we make the (over)simplifying assumption

$$\Sigma_H = \Sigma_{H_0} \exp(-2B_H s^2). \quad (20)$$

Then, from the previous considerations [(4) and (5)], one is led for Σ_N to the simple expression

$$\Sigma_N = \Sigma_P[1 - \exp(-2B_N s^2)]. \quad (21)$$

The values of Σ_{H_0}, B_H and B_N (the 'LOI factor') are to be determined to fit the variation of Σ_{obs} with resolution. Practically speaking, Σ_{obs} is calculated in shells of resolution by considering the acentric and centric estimates obtained as described previously, and the three unknowns are then determined after a systematic three-dimensional search followed by a nonlinear refinement. The residual function to minimize is obtained from Σ_{obs} and Σ_{calc} [from (19), (20) and (21)] and its gradient and Hessian matrix† are analytically determined. The inverse of the Hessian matrix also allows the estimation of the errors on the obtained values.

On the one hand, the residual function is well conditioned because it has a single minimum (at least in the examples examined). However, on the other hand, this function is extremely flat in one direction around this minimum owing to important correlation of the three parameters. There is, in particular, a serious danger of overfitting or underfitting the obtained set of Σ_H values owing to the difficulty of correctly estimating a confidence interval for each of them and, consequently, a proper weighting scheme. Clearly, some progress is necessary in this area. Related to this weighting problem is the extreme simplification of invoking Wilson statistics for the contribution of heavy atom(s). High-symmetry space groups could be more favourable than low-symmetry space goups in this respect (Shmueli & Weiss, 1987). Despite this obvious limitation, the proposed method gives rather satisfactory results in test cases with different levels of LOI.

In real cases, the results are generally, if not always, consistent with the refined heavy-atom parameters. These results, for both test and real cases, are shown in § 6.3.

6.2. *Determination of a global absolute occupancy factor for the heavy atom(s)*

The term Σ_{H_0} [(20)] obtained from the previous study cannot yet be taken as being on an absolute scale. However, one can easily derive an approximate value of the global quantity $Q_H = (\sum_{j=1}^{N_{site}} Q_j^2)^{1/2}$, where Q_j is the absolute occupancy of the jth site. This is achieved by the following Wilson-like analysis:

$$\overline{F_P^2}/\overline{F_H^2} = \Sigma_P/\Sigma_H = \sum_i N_i \int f_i^2 s^2 ds / Q_H^2 \int f_H^2 s^2 ds, \quad (22)$$

with N_i being the number of atoms of the ith type in the native molecule. The integrals of the third member can be calculated by considering the standard representation of all atomic structure factors in terms of a sum of Gaussians (*International Tables for X-ray Crystallography*, 1974), and Q_H is thus obtained. In fact, their exact calculation requires the knowledge of the overall temperature-factor values for the macromolecule, B_P, and the heavy atoms, B_H. B_H is known from the previous study. However, B_P cannot be obtained from a Wilson plot if experimental data do not extend much beyond 3 Å resolution. In such a case, one must rely on a guess value. Theoretically, the previous expressions are

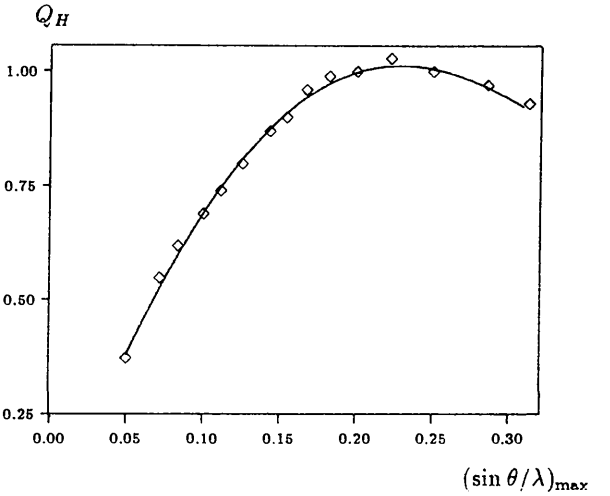

Fig. 2. Evolution of the theoretically determined value of Q_H *versus* the upper limit of resolution used in calculations for a test case in $P2_12_12_1$. For each point, all low-resolution data have been used. The theoretical F_{PH}^2 terms have been calculated with one heavy atom with $Q_H = 1$ and $B_H = 0$ Å^2, without any influence of LOI. The existence of a maximum, instead of a plateau, at $Q_H = 1$ is due to the slight imprecision in the determination of B_H, whose importance increases with resolution.

* As mentioned by Tickle (1991), there is often miscalculation of $\sigma(F)$ from $\sigma(I)$ in many data-processing programs for low values of F. A more accurate value of it for the limiting case $F = 0$ seems to be $\sigma(F) = [\sigma(I)/2^{1/2}]^{1/2}$ instead of $\sigma(F) = [\sigma(I)]^{1/2}$ as stated by Tickle (in any case, the two values differ only by 16%). This results from considering a Gaussian probability density truncated to zero for negative values of I and renormalized for taking care of this fact. This yields a somewhat complex expression, valid for any situation ranging from $I/\sigma(I) = 0$ to $I/\sigma(I) = \infty$, for which one retrieves the usual approximation $\sigma(F) = \sigma(I)/2F$.

† The Hessian matrix is evaluated by keeping only the product of first derivatives for the cross terms (Press, Flannery, Teukolsky & Vetterling, 1986).

Table 2. *Comparison for test cases with LOI of the estimated values of Q_H, B_H (Å^2) and B_N (Å^2) with their exact values*

Space group $P4_12_12$. Native structure factors calculated with the coordinates of the bleomycin resistance protein, derivative structure factors calculated with the six sites of its europium derivative with the following absolute occupancies and temperature factors Q_i, B_i: $Q_1 = 1$, $B_1 = 55$ Å^2; $Q_2 = 0.39$, $B_2 = 90$ Å^2; $Q_3 = 0.31$, $B_2 = 50$ Å^2; $Q_4 = 0.28$, $B_4 = 50$ Å^2; $Q_5 = 0.20$, $B_5 = 25$ Å^2; $Q_6 = 0.1$, $B_6 = 40$ Å^2. These values allow us to calculate the theoretical values $Q_H = 1.15$ and $B_H = 55$ Å^2 to minimize the squared residual difference between $\sum_i Q_i^2 s^2 \exp(-2B_i s^2)$ and $Q_H^2 s^2 \exp(-2B_H s^2)$, in reasonable agreement with those obtained when correcting for the effect of LOI.

Low LOI. LOI modelled by the movement of one loop (ten residues). The resulting overall r.m.s. deviation is $\langle \Delta \mathbf{r}^2 \rangle^{1/2} = 0.15$ Å yielding a theoretical value of $B_N = 0.6$ Å^2 from (6), in excellent agreement with the value found.

Medium LOI. Same changes as above plus a rigid-body translation of −0.10, 0.05, −0.15 Å along the axes and a rotation of 0.15° about an arbitrary axis. The resulting overall r.m.s. deviation is $\langle \Delta \mathbf{r}^2 \rangle^{1/2} = 0.294$ Å yielding a theoretical value of $B_N = 2.3$ Å^2 from (6), in excellent agreement with the value found.

High LOI. Same changes as above plus unit-cell modification from $a = b = 48.4$, $c = 111.5$ Å to $a = b = 48.6$, $c = 110.5$ Å. The resulting value of $\langle \Delta \mathbf{r}^2 \rangle$ has been estimated simply by adding to $(0.294)^2$ the contributions $\langle \Delta X_i^2 \rangle = (\delta a_i)^2 \int_0^1 x_i^2 \mathrm{d}x_i = (\delta a_i)^2/3$ from the three mutually orthogonal directions $i = 1, 3$. This makes the simplifying assumption of a constant density of atoms within the unit cell. The result is $\langle \Delta \mathbf{r}^2 \rangle = (0.294)^2 + 2 \times 0.04/3 + 1/3 = 0.446$ Å^2, yielding a theoretical value of $B_N = 11.7$ Å^2 from (6), in good agreement with the value found.

LOI	B_H*	B_H (e.s.d.)	B_H, theoretical	Q_H*	Q_H	Q_H, theoretical	B_N (e.s.d.)	B_N, theoretical
Low	44	77 (6)	55	1.91	1.16	1.15	0.55 (12)	0.6
Medium	24	60 (12)	55	1.92	1.16	1.15	2.2 (5)	2.3
High	10	85 (37)	55	4.32	1.42	1.15	15.8 (22)	11.7

* Values obtained without correction for LOI.

valid when considering infinitely many Fourier coefficients. Practically, only an approximate value of Q_H is obtained. Test calculations show that Q_H tends to be underestimated until the resolution is less than 2.5 Å (Fig. 2), but also that Q_H tends to be overestimated if low-resolution data are ignored (not shown). It turns out that these two effects cancel almost exactly for data lying in the usual resolution range of 10-4.0 Å. Interestingly, this low limit of 10 Å is also quite convenient for the elimination of the deleterious effects of the disordered solvent.

6.3. *Results concerning the determination of B_H, Q_H and B_N*

Results from the two previous sections concern both test and real cases. A first test case has been performed to assess the results without LOI but at varying levels of substitution (Table 1, columns 7 and 9). It clearly gives a very satisfying result. A second test has been performed to assess our ability to differentiate between true signal and noise from LOI. For that, three sets of data were calculated with, respectively, low, medium and high LOI as described in the legend of Table 2. Furthermore, 'experimental' errors were artificially added in order to simulate real data as realistically as possible. With that aim, an attempt has been made to mimic the behaviour of errors *versus* intensity and resolution as observed for a real case, namely for the experimental data from the bleomycin resistance protein (BRP) (Dumas, Bergdoll, Cagnon & Masson, 1994). The results are given in Table 2. They show that both B_H and Q_H are correctly recovered when the LOI contribution has been estimated and removed. One just notices a tendency to obtain too high a value for B_H. This may not be general but just the result of this particular case. An extremely interesting result concerns the remarkable quality of the values obtained for B_N, allowing the level of LOI to be quantified. This quality can be assessed by comparing the value of $\langle \Delta \mathbf{r}^2 \rangle$ obtained from (6) with that obtained directly from the known modification of the coordinates for modelling LOI.

Results using real cases are given in Table 3. They concern several structures (RNA and proteins) solved in our laboratory. There is obviously no way to compare the obtained value for B_N with its exact counterpart. However, the highest values obtained for B_N concern ASPA(Au) and ATIII(Pt). This is in agreement with the low phasing power of these derivatives. More specifically, for ASPA(Au), this agrees with the LOI induced on soaking, which was characterized after refinement (Westhof, Dumas & Moras, 1985; Dumas, 1986).

7. Consequence for the *a priori* estimation of the lack of closure and of the phasing power

The previous results have an important practical consequence. Indeed, knowledge of the terms Σ_H, Σ_{exp} and Σ_N allows the calculation of a very early estimate of the 'lack of closure' and of the signal-to-noise ratio, usually referred to as the 'phasing power'. The latter can be defined as

$$W = (\overline{F_H^2})^{1/2}/(\overline{\varepsilon^2})^{1/2}, \qquad (23)$$

Table 3. *Results for real cases*

Determination of the influence of LOI with B_N (Å^2) and comparison of the global temperature factor B_H with the calculated values $B_{H_{calc}}$ (Å^2) from the refined values of both occupancy (Q_i) and temperature factor (B_i) for each site i, as explained in the legend of Table 2. ALD: aldose reductase (Rondeau *et al.*, 1987). ASPA: yeast aspartic tRNA (Comarmond, Giegé, Thierry & Moras, 1986). ATIII: anti-thrombin III (Samama, Delarue, Mourey, Choay & Moras, 1989). BRP: bleomycin resistance protein (Dumas *et al.*, 1994). $B_{H_{calc}}$ for ASPA(Au) was not available since one site had its temperature factor arbitrarily blocked at 100 Å^2 (Comarmond *et al.*, 1986). However, the correct value is certainly in better agreement with the value of 90 Å^2 obtained after consideration of the influence of LOI than with the null value obtained without consideration of the influence of LOI. Finally, the important disagreement between the values B_H = 145 Å^2 and $B_{H_{calc}}$ = 58 Å^2 for BRP(Eu) has no obvious explanation.

Crystal	Space group	Resolution range (Å)	B_H* (e.s.d.)	B_H (e.s.d.)	$B_{H_{calc}}$	B_N (e.s.d.)
ALD (Hg)	$P4_12_12$	10–3.5	22 (2)	37 (4)	29	0.48 (20)
ASPA (Gd)	$C222_1$	12–3.3	37 (4)	55 (15)	45	0.45 (50)
ASPA (Au)	$C222_1$	12–4.0	0 (9)	90 (65)	?	1.4 (5)
ATIII (Pt)	$P4_32_12$	12–5.0	122 (11)	193 (26)	157	2.1 (7)
BRP (Eu)	$P4_12_12$	10–3.5	37 (7)	145 (14)	58	0.40 (7)
BRP (Hg)	$P4_12_12$	10–3.3	29 (2.5)	33 (4)	28	0.040 (25)

* Values obtained without consideration of the influence of LOI. These are in excellent agreement with the calculated ones when B_N is close to zero.

with ε being the so-called lack of closure. Its r.m.s. is given by

$$(\overline{\varepsilon^2})^{1/2} = (\Sigma_{exp} + \Sigma_N)^{1/2} \tag{24}$$

and the phasing power $W(s)$, as a function of resolution, thus reads, from (20) and (21),

$$\begin{aligned} W(s) &= \Sigma_H^{1/2}/(\Sigma_{exp} + \Sigma_N)^{1/2} \\ &= \Sigma_{H_0}^{1/2} \exp(-B_H s^2) \\ &\quad \times \{\Sigma_{exp} + \Sigma_P[1 - \exp(-2B_N s^2)]\}^{-1/2}. \end{aligned} \tag{25}$$

We thus obtain explicit analytical expressions for the dependence of lack of closure and phasing power on resolution.

These results have been checked with test data in order to obtain exact values of both $(\overline{\varepsilon^2})^{1/2}$ and W for the comparison (Fig. 3). They are extremely encouraging since the lowering of the phasing power with increasing LOI at a given resolution and with increasing resolution at a given level of LOI is rather well recovered. It is remarkable too that Σ_H is well obtained, in the whole resolution range, for all levels of LOI. Less satisfactory are the results at low resolution, for low and medium LOI, on the lack of closure and thus also on the phasing power. However, for high LOI, the agreement is remarkable over the whole resolution range. Furthermore, two points should be noted. Firstly, the quality of the Σ_H estimate is an *a posteriori* justification of the use of Wilson statistics (even though it must be kept in mind that other cases could give much less satisfactory results). Secondly, the relative importance of the lack of closure, from high to low LOI, is very well appreciated and the agreement is greatest when the knowledge of this term is of most practical importance, namely for high LOI. Therefore, the lack of closure, up to now only estimated at the refinement step of the heavy-atom parameters, can now be obtained prior to the determination of any site of substitution. It will thus be possible to use it as a fixed quantity for a proper weighting of such refinement. This is thus one possible solution to the weighting problem as explained by Bricogne (1991). These considerations await further practical tests.

We discuss here only one example of potentially high practical importance. This method, when used on the mercury derivative of the BRP (see Table 3), gives a nearly null value for B_N (thus for the level of LOI) and, consequently, a very high value of the theoretical phasing power as given by (25). If this result is valid, the discrepancy between the high theoretical and much lower observed values of the phasing power $W(s)$ must be interpreted as due to defects in the heavy-atom model. In particular, equation (24) for the r.m.s of the lack of closure must be modified to

$$(\overline{\varepsilon^2})^{1/2} = (\Sigma_{exp} + \Sigma_N + \theta\Sigma_H)^{1/2}, \tag{26}$$

where θ represents the fraction of the heavy-atom power incorrectly accounted for by the present model. It should be kept in mind that θ may correspond to either a missing or an excess fraction. This θ value thus appears as a potential quantitative criterion to decide whether or not continued improvement of the heavy-atom model is justified. In the most general situation, this is a function not only of the resolution $2s = |\mathbf{h}|$ but of $\mathbf{h}$ as a vector. If we consider only the radial dependence, $\theta(s)$ can be obtained from (24), (25) and (26) as

$$\theta(s) = [1/W_{obs}^2(s)] - [1/W_{theo}^2(s)]. \tag{27}$$

Clearly, only significant discrepancies between observed and theoretical phasing power should be considered for this test. It is instructive to grasp the influence of θ on W_{obs}. From (27), one immediately obtains

$$W_{obs} = W_{theo}/(1 + \theta W_{theo}^2)^{1/2}. \tag{28}$$

In the particular case of the BRP mercury derivative, the values $W_{obs} \simeq 2.9$ and $W_{theo} \simeq 10$ were obtained at 10 Å resolution. From the previous equations, a value of θ as low as 11% is sufficient to explain the difference. In this particular case, where the heavy atom was not present as a mere cation but was complexed by an organic molecule (*p*-chloromercuribenzenesulfonate, PCMBS), it may be that much of the θ value could be accounted for by the light atoms ignored in the heavy-atom model. Incidentally, a minor peak from a residual map, 5.5 Å from the major site, was not interpreted as a minor site but as the sulfonate moiety of the PCMBS molecule and modelled by a single S atom. Therefore, the light atoms ignored are the three O atoms of the sulfonate moiety and the six C atoms of the aromatic ring. Interestingly, these missing atoms yield the estimate $\theta \simeq 5\%$, whose magnitude fits well with the upper value of 11%.

More work is necessary to fully assess these considerations. This could be done by obtaining all minor sites of the BRP mercury derivative (if any) and refining all parameters by refining the whole structure [protein plus the bound PCMBS molecule(s)] against the derivative data. There is no doubt that such a heavy-atom model would be ideal, since it is unattainable in practice when solving a structure, otherwise the multiple-isomorphous-replacement phases would directly correspond to the best phases. Clearly, the estimate of the ideal 'phasing power' obtained from this ideal heavy-atom model will inevitably be higher than the phasing power actually obtained at the solution step. The test would therefore amount to verification that this higher value is close to the theoretical one obtained *a priori*. This should obviously hold in the whole resolution range to be significant, not only at a given resolution. Paper II (Dumas, 1994) examines the quantitative relationship between the phasing power and the figure of merit quantifying the quality of the phase estimate.

8. Concluding remarks

A statistical study of the heavy-atom problem has been performed. This has permitted the acquisition of a great deal of information with no other knowledge than that of the two experimental sets of measured intensities (and their σ's) from a native and derivative crystal pair. These concern:

(1) The definition of an efficient criterion for the scaling of derivative to native data. The latter amounts to

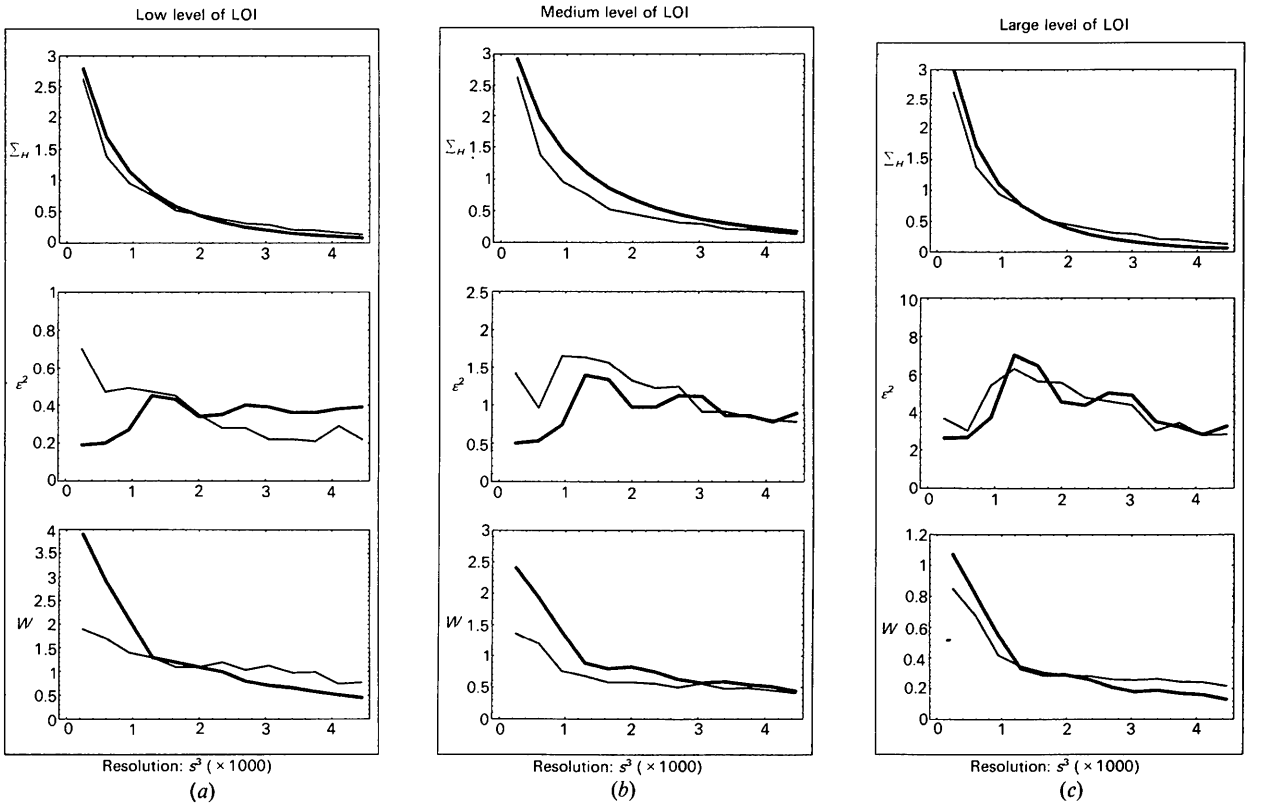

Fig. 3. Comparison with their exact values (thin lines) of the values obtained *a priori* (thick lines) for Σ_H, the square of the lack of closure ($\overline{\varepsilon^2}$) and the phasing power (W). The comparison has been made for three levels of LOI: (*a*) low; (*b*) medium; (*c*) high. The horizontal axis is labelled in $s^3 = (\sin\theta/\lambda)^3$ units in order to yield an equal density of reflections along the axis (resolution range 12-3 Å). The vertical axes for the square of the lack of closure and for Σ_H are labelled in arbitrary units (but are comparable in all cases).

evaluating the power of the component of the derivative signal uncorrelated with the native signal. This power not only corresponds to Σ_H, the genuine power of the heavy-atom signal, but to its sum with any other sources of variance between native and derivative signals. In particular, when LOI affects the derivative signal, this criterion determines correctly the part of the LOI signal that is negatively correlated with the native signal from its true noise component.

(2) An *a priori* estimation of the level of substitution by the calculation of global absolute occupancy and thermal parameters for the heavy atom(s).

(3) An *a priori* estimation of the influence of the noise due to LOI in terms of a 'LOI factor' analogous to a Debye–Waller coefficient. As an important consequence, one obtains a quantitative estimate of the dependence *versus* resolution of the signal-to-noise ratio, *i.e.* an early indication of the lack of closure and of the phasing power. Several theoretical results were carefully tested. They concern the scaling procedure and the determination of the level of substitution and of LOI. However, the results, of potentially great practical importance for heavy-atom-parameter refinement, require more work to be assessed.

The present considerations have been explicitly used to write a Fortran program, *LOCHVAT*. It is fully compatible with the *CCP*4 package and is available on request.

I thank D. Moras for his constant support during this work. I am grateful to all my colleagues for their kind collaboration in making their data available to me. I thank G. Webster, D. Logan and D. Harris for correction of the English of the successive versions of the manuscript. I express my warmest thanks to R. Ripp for his invaluable help in using TEX and to A. Urzhumtsev for a careful reading of the manuscript and many valuable comments. I also acknowledge the very constructive remarks of one referee.

APPENDIX
Calculation of $\langle F_{PH}^2 \rangle$ and $\langle F_{PH} \rangle$

All properties of the Bessel functions used below can be found in any textbook on the topic (for example Nikiforov & Ouvarov, 1976).

Calculation of $\langle F_{PH}^2 \rangle$

By definition,

$$\langle F_{PH}^2 \rangle = \int_0^\infty F_{PH}^2 P(F_{PH}) \mathrm{d}F_{PH} \tag{29}$$

$$\langle F_{PH}^2 \rangle = (\Sigma_H^3/8F_P^4) \exp(-F_P^2/\Sigma_H) \times \int_0^\infty X^3 \exp(-\alpha X^2) I_0(X) \mathrm{d}X, \tag{30}$$

with $\alpha = \Sigma_H/4F_P^2$. By using the series expansion of $I_0(X)$ and inverting the order of summation and integration (by virtue of the absolute convergence of the resulting series), one obtains

$$\langle F_{PH}^2 \rangle = (\Sigma_H^3/8F_P^4) \exp(-F_P^2/\Sigma_H) \times \sum_0^\infty [1/4^k (k!)^2] \times \int_0^\infty X^{2(k+1)+1} \exp(-\alpha X^2) \mathrm{d}X. \tag{31}$$

The integral appearing in the infinite series is equal to $(k+1)!/2\alpha^{k+2}$ (Gradshteyn & Ryzhik, 1980, p. 337), which gives

$$\sum_0^\infty [1/4^k (k!)^2] \int_0^\infty X^{2(k+1)+1} \exp(-\alpha X^2) \mathrm{d}X = (1/2\alpha^2) \sum_0^\infty [(k+1)/k!](1/4\alpha)^k. \tag{32}$$

The series obtained can be explicitly summed by considering the function

$$f(z) = z \exp z = \sum_0^\infty z^{k+1}/k!, \tag{33}$$

whose derivative leads to the sought-after result with $z = 1/4\alpha$:

$$f'(z) = (z+1) \exp z = \sum_0^\infty [(k+1)/k!] z^k \tag{34}$$

and the result given in (10) is obtained after a few manipulations.

Calculation of $\langle F_{PH} \rangle$

Analogously to the previous case, one is led to

$$\langle F_{PH} \rangle = (\Sigma_H^2/2F_P^2) \exp(-F_P^2/\Sigma_H) \times \int_0^\infty X^2 \exp(-\alpha X^2) I_0(X) \mathrm{d}X. \tag{35}$$

Instead of expanding $I_0(x)$, it is preferable to replace it by $J_0(iX)$, with $i^2 = -1$ and $J_0(z)$ being the Bessel function of the first kind of order 0. This leads to (Nikiforov & Ouvarov, 1976, p. 235)

$$\int_0^\infty X^2 \exp(-\alpha X^2) J_0(iX) \mathrm{d}X = (\pi/16\alpha^3)^{1/2} \exp(1/4\alpha) F\left(-\tfrac{1}{2}, 1; -1/4\alpha\right), \tag{36}$$

$F(-\frac{1}{2}, 1; -1/4, \alpha)$ being the confluent hypergeometric

function. Finally, one obtains

$$\langle F_{PH} \rangle = \tfrac{1}{2}(\pi\Sigma_H)^{1/2} F\left(-\tfrac{1}{2}, 1; -Y\right), \qquad (37)$$

which is the result given in (12) with $Y = 1/4\alpha = F_P^2/\Sigma_H$.

References

BLUNDELL, T. L. & JOHNSON, L. N. (1976). *Protein Crystallography*. London: Academic Press.

BRICOGNE, G. (1991) *Crystallographic Computing* 5, edited by D. MORAS, A. D. PODJARNY & J. C. THIERRY, pp. 257-297. IUCr/Oxford Univ. Press.

COMARMOND, M. B., GIEGÉ, R., THIERRY, J. C. & MORAS, D. (1986). *Acta Cryst.* B**42**, 272-280.

CRICK, F. H. C. & MAGDOFF, B. S. (1956) *Acta Cryst.* **9**, 901-908.

DUMAS, P. (1986). PhD thesis, Univ. Louis Pasteur, Strasbourg, France.

DUMAS, P. (1991). *Crystallographic Computing* 5, edited by D. MORAS, A. D. PODJARNY & J. C. THIERRY, p. 479. IUCr/Oxford Univ. Press.

DUMAS, P. (1994). *Acta Cryst.* **50**, 537–546.

DUMAS, P., BERGDOLL, M., CAGNON, C. & MASSON, J. M. (1994). *EMBO J.* **13**. In the press.

GRADSHTEYN, I. S. & RYZHIK, I. M. (1980). *Tables of Integrals, Series, and Products*. London: Academic Press.

International Tables for X-ray Crystallography (1974). Vol. IV. Birmingham: Kynoch Press. (Present distributor Kluwer Academic Publishers, Dordrecht.)

KRAUT, J., SIEKER, L. C., HIGH, D. F. & FREER, S. T. (1962). *Proc. Natl Acad. Sci. USA*, **48**, 1417-1424.

LUZZATI, V. (1952). *Acta Cryst.* **5**, 802-810.

NIKIFOROV, A. & OUVAROV, V. (1976). *Eléments de la Théorie des Fonctions Spéciales*, pp. 206-217. Moscow: Editions MIR.

NIXON, P. E. & NORTH, A. C. T. (1976). *Acta Cryst.* A**32**, 325-333.

PERUTZ, M. F. (1956). *Acta Cryst.* **9**, 867-873.

PRESS, W. H., FLANNERY, B. P., TEUKOLSKY, S. A. & VETTERLING, W. T. (1986). *Numerical Recipes: the Art of Scientific Computing*, pp. 522-523. Cambridge Univ. Press.

READ, R. J. (1986). *Acta Cryst.* A**42**, 140-149.

READ, R. J. (1990). *Acta Cryst.* A**46**, 900-912.

READ, R. J. (1991). *Crystallographic Computing* 5, edited by D. MORAS, A. D. PODJARNY & J. C. THIERRY, pp. 158-167. IUCr/Oxford Univ. Press.

RONDEAU, J. M., SAMANA, J. P. SAMANA, B., BARTH, P., MORAS, D. & BIELLMANN, J. F. (1987). *J. Mol. Biol.* **195**, 945-948.

ROSSMANN, M. G. (1960). *Acta Cryst.* **13**, 221-226.

SAMAMA, J. P., DELARUE, M., MOUREY, L., CHOAY, J. & MORAS, D. (1989). *J. Mol. Biol.* **210**, 877-879.

SHMUELI, U. & WEISS, G. H. (1987). *Acta Cryst.* A**43**, 93-98.

SIM, G. A. (1959). *Acta Cryst.* **12**, 813-815.

TICKLE, I. J. (1991). *Isomorphous Replacement and Anomalous Scattering*, compiled by W. WOLF, P. R. EVANS & A. G. W. LESLIE, pp. 87-95. Proceedings of the *CCP*4 study weekend, 25-26 January 1991. SERC Daresbury Laboratory, Warrington, England.

WESTHOF, E., DUMAS, P. & MORAS, D. (1985). *J. Mol. Biol.* **184**, 119-145.

WILSON, A. J. C. (1942). *Nature (London)*, **150**, 151-152.

WILSON, A. J. C. (1949). *Acta Cryst.* **2**, 318-321.

Acta Cryst. (1994). A**50**, 537–546

The Heavy-Atom Problem: a Statistical Analysis. II. Consequences of the *A Priori* Knowledge of the Noise and Heavy-Atom Powers and use of a Correlation Function for Heavy-Atom-Site Determination

BY PHILIPPE DUMAS

UPR de Biologie Structurale, Institut de Biologie Moléculaire et Cellulaire, 15 rue René Descartes, 67084 Strasbourg CEDEX, France

(*Received* 22 *September* 1993; *accepted* 7 *February* 1994)

Abstract

A preceding paper reported how to obtain *a priori* quantitative information on the lack of isomorphism (LOI), considered as noise corrupting the heavy-atom signal in a derivative data set. This related paper initially examines how additional *a priori* information can be drawn from the knowledge of the level of LOI. First, a corrected estimate of the coefficients necessary for a difference Patterson synthesis is derived. An estimate of their accuracy is also obtained. Then, individual and, independently, shell-averaged figures of merit that can be expressed in terms of the phasing power obtained in the preceding paper are determined. These afford an early estimate of the probable phase error on the heavy-atom structure factor. In a second and independent part of the paper, a correlation/translation function is proposed for the localization of the heavy-atom site(s). The results, bearing on both test and real cases, show that this method can be helpful in many situations.

1. Introduction

In a preceding paper (Dumas, 1994), from now on referred to as I, it was shown that a great deal of information can be obtained about the LOI corrupting a derivative data set before any heavy-atom sites are determined. This second paper is first devoted to drawing useful consequences from this knowledge, with regard first to re-estimating the best coefficients for a difference Patterson synthesis. All notation used in the paper is consistent with that used in I or is defined when necessary.