Acta Cryst. (1995). D**51**, 110–117

Correlation Function Method in Protein Crystallography

BY D. ALEXEEV

Edinburgh Centre for Molecular Recognition and The Department of Biochemistry, The University of Edinburgh, George Square, Edinburgh, EH8 9XD, Scotland

(*Received* 10 *July* 1993; *accepted* 23 *June* 1994)

Abstract

A weighted correlation function as a method for computing electron-density maps is proposed to reduce the errors of the Fourier syntheses performed on inaccurate and/or incomplete data. The formulae are revised for the difference Patterson vector search, for multiple isomorphous replacement (MIR) and single isomorphous replacement (SIR) syntheses and for the difference Fourier synthesis. The examples show that the correlation-function approach has the potential to provide more reliable results than those obtained by conventional Fourier syntheses.

Introduction

The Fourier transformation is the mathematical representation of the nature of the diffraction process. The individual point scatterers form a regular grid over the unit cell which comprises the electron-density map $\rho(\mathbf{r}_i)$ of an object. The structure factor of the ith scattering unit $\rho(\mathbf{r})$ is $\mathbf{f}(\mathbf{k},\mathbf{r}_i) = \rho(\mathbf{r}_i)\exp(i\mathbf{k}\mathbf{r}_i)$, $\mathbf{r}_i$ being its position in the real space and $\mathbf{k}$ being the scattering vector in reciprocal space. The sum of the complex structure factors $\mathbf{f}(\mathbf{k},\mathbf{r}_i)$ of all individual scatterers of the map makes the total structure factor of a whole object $\mathbf{F}(\mathbf{k})$,

$$\mathbf{F}(\mathbf{k}) = \sum \mathbf{f}(\mathbf{k},\mathbf{r}_i),$$

where

$$\mathbf{f}(\mathbf{k}, \mathbf{r}_i) = \rho(\mathbf{r}_i)\exp(i\mathbf{k}\mathbf{r}_i).$$

The basic property of the diffraction is that the average product of waves $\mathbf{f}(\mathbf{k},\mathbf{r}_{j1})$ and $\mathbf{f}(\mathbf{k},\mathbf{r}_{j2})$ diffracted by any pair of individual scatterers $\rho(\mathbf{r}_{j1})$ and $\rho(\mathbf{r}_{j2})$ is always zero over the reciprocal space except when $\mathbf{r}_{j1} = \mathbf{r}_{j2}$,

$$\begin{aligned}(1/N)\sum \mathbf{f}(\mathbf{k},\mathbf{r}_{j1})\mathbf{f}^*(\mathbf{k},\mathbf{r}_{j2}) &= \langle \mathbf{f}(\mathbf{k},\mathbf{r}_{j1})\mathbf{f}^*(\mathbf{k},\mathbf{r}_{j2})\rangle \\ &= \rho(\mathbf{r}_{j1})\rho(\mathbf{r}_{j2})\langle \exp[i\mathbf{k}(\mathbf{r}_{j1} - \mathbf{r}_{j2})]\rangle \\ &= 0,\end{aligned}$$

where $\mathbf{f}^*(\mathbf{k},\mathbf{r}_{j1})$ is the complex conjugate of $\mathbf{f}(\mathbf{k},\mathbf{r}_{j1})$; the summation is over all $N \rightarrow \infty$ reflections $hkl \neq (0,0,0)$ in reciprocal space. The cross-term $\langle \exp[i\mathbf{k}(\mathbf{r}_{j1} - \mathbf{r}_{j2})]\rangle$ comprises the mean value of a periodic function and it is always zero in the ideal case at infinite resolution. This means that the individual components $\mathbf{f}(\mathbf{k},\mathbf{r}_{j1})$ and $\mathbf{f}(\mathbf{k},\mathbf{r}_{j2})$ of a complex diffraction pattern $\mathbf{F}(\mathbf{k})$ are orthogonal to each other and the electron density $\rho(\mathbf{r}_{j1})$ can be returned provided both the amplitudes and the phases of the complex structure factor $\mathbf{F}(\mathbf{k})$ are known and the resolution is high enough to ensure that the cross-terms are zero,

$$\begin{aligned}\rho(\mathbf{r}_{j1}) &= \langle \mathbf{F}(\mathbf{k})\exp(-i\mathbf{k}\mathbf{r}_{j1})\rangle \\ &= \rho(\mathbf{r}_{j2})\langle \exp[-i\mathbf{k}(\mathbf{r}_{j1} - \mathbf{r}_{j2})]\rangle + \rho(\mathbf{r}_{j1}) = \rho(\mathbf{r}_{j1}),\end{aligned}$$

where $\langle\ \rangle$ means averaging over the reciprocal space $hkl \neq (0,0,0)$ as above. Now the conventional Fourier synthesis can be rewritten in the form,

$$\rho(\mathbf{r}_i) = \langle \mathbf{F}(\mathbf{k})\exp(-i\mathbf{k}\mathbf{r}_i)\rangle = \langle \mathbf{F}(\mathbf{k})\mathbf{t}^*(\mathbf{k},\mathbf{r}_i)\rangle, \quad (1)$$

The term $\mathbf{t}(\mathbf{k},\mathbf{r}_i) = \exp(i\mathbf{k}\mathbf{r}_i) = 1\exp(i\mathbf{k}\mathbf{r}_i)$ can be understood as the structure factor of a unit scatterer. Hence the sense of the Fourier transformation is to probe the net diffraction pattern with a trial wave $\mathbf{t}$ diffracted by a unit scatterer. The probe $\mathbf{t}(\mathbf{k},\mathbf{r}_i)$ specifically extracts its mate $\mathbf{f}(\mathbf{k},\mathbf{r}_i)$, corresponding to the same position $\mathbf{r}_i$ in the unit cell, from the net pattern $\mathbf{F}(\mathbf{k})$ as $\langle \mathbf{f}(\mathbf{k},\mathbf{r}_i)\mathbf{t}^*(\mathbf{k},\mathbf{r}_i)\rangle = \rho(\mathbf{r}_i)$ and the cross-terms $\mathbf{r}_i \neq \mathbf{r}_j$ are zero: $\langle \mathbf{f}(\mathbf{k},\mathbf{r}_j)\mathbf{t}^*(\mathbf{k},\mathbf{r}_i)\rangle = 0$. The probe is successively applied to the map, point by point, and returns the electron density of a molecule.

The equation $\mathbf{t}(\mathbf{k},\mathbf{r}_i) = \exp(i\mathbf{k}\mathbf{r}_i)$ corresponds directly to the space group $P1$. For the higher symmetry space groups $\mathbf{t}(\mathbf{k},\mathbf{r}_i)$ becomes the sum of the unit structure factors over M symmetry-related positions in the unit cell,

$$\mathbf{t}(\mathbf{k},\mathbf{r}_i) = \sum_m \mathbf{t}(\mathbf{k},\mathbf{r}_{im}), \quad m = 1 \rightarrow M.$$

The Fourier synthesis (1) works perfectly for an ideal case where the resolution is infinite and structure-factor amplitudes and phases are precise. Real diffraction data sets are collected to a limited resolution with imprecise amplitudes and phase determination can produce significant errors and leads to distortions in the electron-density map. The errors can alter the value of the product $\langle \mathbf{f}(\mathbf{k},\mathbf{r}_i)\mathbf{t}^*(\mathbf{k},\mathbf{r}_i)\rangle$ and hence the true value $\rho(\mathbf{r}_i)$ of the

electron density at $\mathbf{r}_i$ is not returned correctly by the Fourier synthesis. The problem of electron-density evaluation at a position $\mathbf{r}_i$ can now be reformulated in more general terms. Looked at another way, a wave $\mathbf{f}(\mathbf{k},\mathbf{r}_i)$ needs to be extracted from a set of noisy experimental observations $\mathbf{F}(\mathbf{k})$. Various mathematical approaches are available by which the Fourier transformation (1) is the simplest product function. However for problems like this, the correlation function is known to be the most appropriate (Samuels, 1989). The violation of the *a priori* assumptions essential for the Fourier transformation results in map errors which might be avoided by the correlation function where these assumptions are no longer necessary. Below we show how the weighted correlation-function formulae can be applied to some conventional Fourier-based map syntheses used in protein crystallography. We also test its performance for a simple model case and for a real heavy-atom search using experimental data. For error-free data available to infinite resolution both approaches produce identical results but the correlation function as the more rigorous approach might provide the more precise answer for a real case which contains experimental errors and limitations.

Fourier synthesis and the correlation function

For simplicity we shall omit the index $\mathbf{k}$ and assume that the average is carried out over all reflections excluding F_{000}. We also omit the index $\mathbf{r}_i$ so that $\mathbf{t}(\mathbf{k},\mathbf{r}_i)$ will be designated $\mathbf{t}$ or $\mathbf{t}_i$. The scalars will be written in italics (the structure-factor amplitude will be F) and the complex numbers in bold (the complex structure factor is $\mathbf{F}$). Capital letters will stand for the total diffraction pattern ($\mathbf{F}$ or F) and lower case will correspond to the trial waves ($\mathbf{t}$ or t). The Fourier synthesis (1) can then be rewritten in a short form,

$$\rho(\mathbf{r}_i) = \langle \mathbf{F}(\mathbf{k})\mathbf{t}^*(\mathbf{k},\mathbf{r}_i)\rangle \text{ becomes } \rho = \langle \mathbf{F}\mathbf{t}^*\rangle, \qquad (2)$$

with the averaging over all $hkl \neq (0,0,0)$ as above. In terms of the correlation function the electron density is evaluated as,

$$\rho = \text{coef}(\langle \mathbf{F}\mathbf{t}^*\rangle - \langle \mathbf{F}\rangle\langle \mathbf{t}^*\rangle)/[\sigma(\mathbf{F})\sigma(\mathbf{t})],$$

where coef is a normalization coefficient,

$$\sigma(\mathbf{F})^2 = \langle \mathbf{F}\mathbf{F}^*\rangle - \langle \mathbf{F}\rangle\langle \mathbf{F}^*\rangle,$$

and

$$\sigma(\mathbf{t})^2 = \langle \mathbf{t}\mathbf{t}^*\rangle - \langle \mathbf{t}\rangle\langle \mathbf{t}^*\rangle.$$

If the diffraction data set is complete to infinite resolution and error free then the mean values $\langle \mathbf{F}\rangle$ and $\langle \mathbf{t}\rangle$ of the vectors $\mathbf{F}$ and $\mathbf{t}$ must be zero, and the squared standard deviations of $\mathbf{F}$ and $\mathbf{t}$ are,

$$\sigma(\mathbf{F})^2 \rightarrow \langle F^2\rangle = I_P,$$

I_P = mean protein diffraction intensity,

$$\sigma(\mathbf{t})^2 \rightarrow \langle \mathbf{t}^2\rangle = I_t,$$

I_t = mean trial wave intensity.

Although $\mathbf{F}$ and $\mathbf{t}$ are complex numbers, because Friedel's law implies the summing of complex numbers with the corresponding complex conjugates, the mean values $\langle \mathbf{F}\rangle$ and $\langle \mathbf{t}\rangle$ here are real and $\langle \mathbf{t}\rangle = \langle \mathbf{t}^*\rangle$.

The correlation function becomes identical to the Fourier synthesis (2) for perfect data if the normalization coefficient coef equals $(I_P I_t)^{1/2}$,

$$\rho = (I_P I_t)^{1/2}(\langle \mathbf{F}\mathbf{t}^*\rangle - \langle \mathbf{F}\rangle\langle \mathbf{t}^*\rangle)/[\sigma(\mathbf{F})\sigma(\mathbf{t})] \rightarrow \langle \mathbf{F}\mathbf{t}^*\rangle. \qquad (3)$$

For the case of complete, error-free data I_P and I_t are the theoretical precise values, I_t being constant and different for special and general positions. For a limited resolution I_t is practically constant for a general position and grows if a special position is approached. We calculate I_t for each $\mathbf{t}$ as an average over a complete set of hkl's that is theoretically generated up to the resolution experimentally available.

The accuracy of experimentally derived complex structure factors $\mathbf{F}$ (especially the phase component) varies significantly and a weighting scheme is always used to suppress unreliable terms. The weights w can also be introduced into the correlation function. The sum and the average values in (3) can be treated as weighted-sum and weighted-average values and the weighted correlation function can be written as,

$$\rho = (I_P I_t)^{1/2}(\langle \mathbf{F}\mathbf{t}^*\rangle_w - \langle \mathbf{F}\rangle_w\langle \mathbf{t}^*\rangle_w)/[\sigma_w(\mathbf{F})\sigma_w(\mathbf{t})], \qquad (4)$$

where,

$\langle \mathbf{F}\mathbf{t}^*\rangle_w = \sum w\mathbf{F}\mathbf{t}^*/\sum w$ conventional Fourier synthesis (w = FOM),

$$\sigma_w(\mathbf{F})^2 = \langle \mathbf{F}\mathbf{F}^*\rangle_w - \langle \mathbf{F}\rangle_w\langle \mathbf{F}^*\rangle_w$$

and

$$\sigma_w(\mathbf{t})^2 = \langle \mathbf{t}\mathbf{t}^*\rangle_w - \langle \mathbf{t}\rangle_w\langle \mathbf{t}^*\rangle_w.$$

Note that the weighted $\sigma_w(\mathbf{F})$ and $\sigma_w(\mathbf{t})$ are different from the unweighted normalization coefficients $(I_P)^{1/2}$ and $(I_t)^{1/2}$. (4) uses the correlation function to perform the same task as the product function used by the Fourier synthesis, but in a more rigourous manner. Functions (2), (3) and (4) are identical for error-free data available to an infinite resolution. If the resolution is limited, the data are incomplete (a common problem with protein data collection) and/or the phases determined are not accurate then the *a priori* assumptions of the Fourier transformation might not hold and the map will be distorted. The correlation function does not require any assumptions and might compensate for these errors. Mathematically the correlation map is not identical

to the Fourier map because of the different weighting scheme. The real weighted mean values $\langle \mathbf{F} \rangle$ and $\langle \mathbf{t} \rangle$ can differ from their theoretical zero value. The weighted $\sigma(\mathbf{F})$ depends on the reflection phase reliabilities (the weights) available from a particular experiment. The value of $\sigma(\mathbf{t})$ and $\langle \mathbf{t} \rangle$ becomes dependent on the position in the unit cell.

The Fourier synthesis procedure assumes that the values $\langle \mathbf{F} \rangle$, $\langle \mathbf{t} \rangle$, are always zero and $\sigma(\mathbf{F})$ and $\sigma(\mathbf{t})$ are constant throughout as it should be for the error-free case. The correlation function allows for their possible variations.

So far we have assumed that an estimate of $\mathbf{F}$ is provided which includes both amplitudes and phases. The correlation formula can also be applied where only partial information about $\mathbf{F}$'s is available [single isomorphous replacement (SIR) synthesis without anomalous data is an example]. The weighted correlation can be calculated between a known fraction of the complex $\mathbf{F}$ and a similar fraction of the trial wave $\mathbf{t}$. For various cases these fractions vary. Some of these applications are considered below.

Error comparison

Let us compare the errors of the Fourier synthesis and of the correlation function. Assume that the phases are inaccurate, the data are available to a limited resolution and the structure-factor amplitudes are precise. This is a fair approximation to the most frequently encountered situation in practice. The weights (usually figures of merit) reflect the reliability of phase determination. The averaging $\langle\ \rangle$ below will mean weighted averaging as in (4) but the weighting index $\langle\ \rangle_w$ is omitted for clarity. The weighted Fourier synthesis provides an estimate of $\rho_{\text{true}}(\mathbf{r}_i)$,

$$\begin{aligned}\rho_{\text{Four}}(\mathbf{r}_i) &= \langle \mathbf{F}\mathbf{t}_i^* \rangle = \langle (\mathbf{F}_{-i} + \mathbf{f}_i)\mathbf{t}_i^* \rangle \\ &= \langle \mathbf{F}_{-i}\mathbf{t}_i^* \rangle + \langle \mathbf{f}_i\mathbf{t}_i^* \rangle \qquad (5) \\ &= \langle \mathbf{F}_{-i}\mathbf{t}_i^* \rangle + \rho_{\text{true}}(\mathbf{r}_i)\langle \mathbf{t}_i\mathbf{t}_i^* \rangle \rightarrow \rho_{\text{true}}(\mathbf{r}_i),\end{aligned}$$

where $\mathbf{F}_{-i}$ is the diffraction from the map lacking density at grid point i. The Fourier approximation works well if the value of the parasitic cross-term $\langle \mathbf{F}_{-i}\mathbf{t}_i^* \rangle$ is small and the mean weighted intensity $\langle \mathbf{t}_i\mathbf{t}_i^* \rangle$ of each trial wave t_i is close to unity, which is ideally true but can differ in practice. The correlation function (4) gives,

$$\begin{aligned}\rho_{\text{cor}}(\mathbf{r}_i) = (I_P I_{ti})^{1/2}(\langle \mathbf{F}_{-i}\mathbf{t}_i^* \rangle - \langle \mathbf{F}_{-i} \rangle\langle \mathbf{t}_i^* \rangle)/ \\ [\sigma(\mathbf{F})\sigma(\mathbf{t}_i)] + \rho_{\text{true}}(\mathbf{r}_i)\sigma(\mathbf{t}_i)(I_P I_{ti})^{1/2}/ \qquad (6) \\ \sigma(\mathbf{F}) \rightarrow \rho_{\text{true}}(\mathbf{r}_i).\end{aligned}$$

For the error-free case and infinite resolution both (5) and (6) approach the true value [see definitions of $\sigma(\mathbf{F})$ and $\sigma(\mathbf{t})$ above]. The Fourier parasitic cross-term $\langle \mathbf{F}_{-i}\mathbf{t}_i^* \rangle$ can sometimes be large due to the fact that both mean values $\langle \mathbf{F}_{-i} \rangle$ and $\langle \mathbf{t}_i^* \rangle$ might not be zero. The correlation formula compensates for this. The true density $\rho_{\text{true}}(\mathbf{r}_i)$ in (5) and (6) is multiplied by $\langle \mathbf{t}_i\mathbf{t}_i^* \rangle$ and $[\text{const}(i)\sigma(\mathbf{t}_i)(I_{ti})^{1/2}]$, respectively. These coefficients might also vary for different positions in the map and should be constant to return $\rho_{\text{true}}(\mathbf{r}_i)$ correctly. As well as for the cross-term of (5) and (6) we believe that the r.m.s. deviation $\sigma(\mathbf{t}_i)$ of the trial-wave intensity from its mean value [coefficient in (6)] is more likely to be constant than the mean value $\langle \mathbf{t}_i\mathbf{t}_i^* \rangle$ itself [coefficient in (5)].

The difference is illustrated by the one-dimensional model calculations presented in Fig. 1. The original electron density comprises two overlapping peaks and is represented by the solid line. The peak amplitudes are 0.5 and 1.0. The structure-factor amplitudes and phases are calculated and a random phase error is introduced in the range $\pm 60°$. The Fourier synthesis (dotted line) and the correlation function (the dashed line) are calculated and scaled to the original curve so that their r.m.s. deviations from the error-free data are minimal. The overall shape of the peaks is reproduced reasonably well by both methods. The correlation function is twice as accurate as the Fourier synthesis, the mean random deviations from the original being 0.057 and 0.112, respectively. The correlation function also reveals the relative peak heights better than the Fourier synthesis. This example shows that the correlation map might be a worthwhile improvement over the conventional Fourier synthesis.

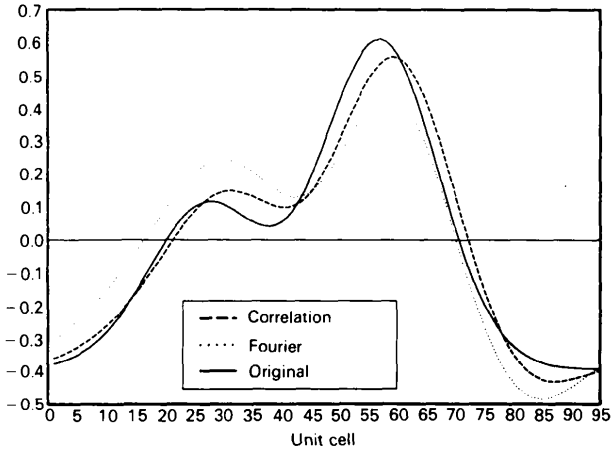

Fig. 1. One-dimensional model calculations. The original periodic density is represented by two partially overlapping Gaussian peaks in the unit cell of 100 grids (horizontal axis). The periodic diffraction pattern is calculated for the first 50 reflections. The random phase error $\Delta\alpha$ is introduced in the range $\pm 60°$, $\Delta\alpha = 60°2[0.5\text{–rnd}(1)]$, and the weights $w = 1 - |\Delta\alpha/180°|$ are assigned to the reflections. The weighted Fourier synthesis and correlation function are calculated.

MIR synthesis

The MIR synthesis represents the case where the weights (figures of merit) can range from 0 to 1. The weighting scheme is determined by the reliability of the phase estimations. The number of low-weighted reflections can be large and in consequence the synthesis is effectively performed on a distorted set of structure factors. If a strong reflection is weighted down then the correct structure-factor amplitude value is reduced, *i.e.* a systematic error is introduced. The traditional Blow & Crick figure of merit does not remove this systematic error but minimizes the damage arising from the phase error by distorting the amplitudes (Blow & Crick, 1959). The weighted correlation function corrects for this error and allows the experimental errors in phase determination to be treated as random ones by the definition. The formula for MIR can be rewritten as,

$$\rho = \langle \mathbf{m}F\exp(-i\mathbf{kr})\rangle = \langle \mathbf{m}F\mathbf{t}^*\rangle = \langle m\mathbf{F}\mathbf{t}^*\rangle.$$

Here $\mathbf{F} = F\exp(i\alpha_{\text{best}})$ is the best experimental estimate for the complex structure factor, $\mathbf{m} = m\exp(i\alpha_{\text{best}})$ is the conventional merit factor as determined by the phase probability distribution $P(\alpha)$ in the MIR experiment,

$$\mathbf{m} = m\exp(i\alpha_{\text{best}}) = \int P(\alpha)\exp(i\alpha)\mathrm{d}(\alpha)/\int P(\alpha)\mathrm{d}(\alpha)$$

The figure of merit m represents the weight $w = m$ reflecting the reliability of phase estimate α_{best}. The replacement for the weighted correlation function is carried out exactly as in (4). The weights $w = m$ are separated from the structure-factor estimates so that the systematic distortions inherent to the weighted Fourier map are eliminated. The result should not be much different from the conventional Fourier synthesis as the phases used are essentially the same but as the actual difference between an interpretable and uninterpretable map is rather small then even a minor improvement should prove useful.

SIR synthesis

If only one isomorphous derivative without anomalous data is available then the SIR synthesis is conventionally used and can be written as,

$$\rho(\mathbf{r}) = \langle \mathbf{m}F\exp(-i\mathbf{kr})\rangle,$$

where $\mathbf{m}$ is the same merit factor as in the MIR synthesis above. The phase probability distribution, however, is symmetrical about the phase of the heavy atom α_H and has two most probable phases α_{prob} as only $\cos(\alpha_{\text{prob}} - \alpha_H)$ is available experimentally. In the absence of experimental errors, the phase probability distribution is sharp and the SIR formulae can be written as,

$$\begin{aligned}\rho &= \langle \cos(\alpha_{\text{prob}} - \alpha_H)F\exp(i\alpha_H)\mathbf{t}^*\rangle \\ &= \langle F\cos(\alpha_{\text{prob}} - \alpha_H)t\cos(\alpha_t - \alpha_H)\rangle \\ &= \langle F_{\mathbf{H}}t_{\mathbf{H}}\rangle, \end{aligned} \tag{7}$$

where α_{prob} is either of the two possible, most probable, phases. The formula comprises a product of the projections $F_{\mathbf{H}}$ and $t_{\mathbf{H}}$ of the vectors $\mathbf{F}$ and $\mathbf{t}$ on the vector $\mathbf{H}$, the structure factor of the heavy atoms. The product can be replaced by the correlation function analogous to the MIR case. The difference is that the correlation is calculated between the projections rather than between complex structure factors themselves. In the presence of experimental errors, the projection $F_{\mathbf{H}}$ can be estimated from the phase probability distribution. Hence the weight w (the figure of merit m) will be modified accordingly. The phase-probability distribution is centred around α_H and the best projection $F_{\mathbf{H}\text{best}}$ is the integral over the whole circle or over either half of it,

$$F_{\mathbf{H}\text{best}} = F\int P(\alpha)\cos(\alpha - \alpha_H)\mathrm{d}(\alpha)/\int P(\alpha)\mathrm{d}(\alpha),$$

where

$$2\pi > \alpha \geq 0 \text{ or } \alpha_H + \pi > \alpha \geq \alpha_H$$
$$\text{or } \alpha_H > \alpha \geq \alpha_H - \pi.$$

Either of two possible solutions of the SIR problem can be defined which are analogous to the MIR case,

$$m\exp(i\alpha_{\text{best}}) = \int P(\alpha)\exp(i\alpha)\mathrm{d}(\alpha)/\int P(\alpha)\mathrm{d}(\alpha), \tag{8}$$

where integration is carried out over $\alpha_H + \pi > \alpha \geq \alpha_H$ for one of them and over $\alpha_H > \alpha \geq \alpha_H - \pi$ for the other. Then the weight w reflects the reliability of the projection and the merit factor (weight w) can be determined analogous to MIR above them,

$$w[F\cos(\alpha_{\text{best}} - \alpha_H)] = wF_{\mathbf{H}} = \mathbf{F}_{\mathbf{H}\text{best}},$$

or from

$$\begin{aligned} &w\cos(\alpha_{\text{best}} - \alpha_H) \\ &\quad = \int P(\alpha)\cos(\alpha - \alpha_H)\mathrm{d}(\alpha)/\int P(\alpha)\mathrm{d}(\alpha). \end{aligned} \tag{9}$$

The definition [(8)–(9)] allows the estimate of a true projection value to be separated from its reliability (weight). If $\mathbf{F}$ is perpendicular to $\mathbf{H}$ then the projection value is zero and the weight of the correlation function term can be 1, if these particular measurements are accurate enough. This term is as valid as a term with a high conventional figure of merit.

The correlation formula for SIR becomes,

$$\begin{aligned}\rho = (I_P I_t)^{1/2}&(\langle F_{\mathbf{H}}\, t_{\mathbf{H}}\rangle_w \\ &- \langle F_{\mathbf{H}}\rangle_w\langle t_{\mathbf{H}}\rangle_w)/[\sigma_w(F_{\mathbf{H}})\sigma_w(t_{\mathbf{H}})]. \end{aligned} \tag{10}$$

If there is a single heavy atom in the unit cell and a trial unit hits its position then $\mathbf{t} = \mathbf{H}$ and $t_{\mathbf{H}} = 1$ for

all reflections. In the error-free infinite-resolution case for space group $P1$, the conventional SIR synthesis (7) becomes,

$$\rho = \langle F_t \rangle = \langle F \cos(\alpha_{best} - \alpha_t) \rangle = \langle \mathbf{F}\mathbf{t}^* \rangle$$
$$= \langle \mathbf{F} \exp(-i\mathbf{kr}) \rangle,$$

where F_t is the projection of $\mathbf{F}$ on the direction of $\mathbf{t}$ and α_t is the phase of $\mathbf{t}$. Two possible α_{best} are centred around α_t and the value of $\cos(\alpha_{best} - \alpha_t)$ is the same for both of them. Thus, the results does not depend on a choice between them. Either phase is as suitable as if the true phase is known. The SIR synthesis is equivalent to the conventional Fourier synthesis which returns the undisturbed electron density of the protein 'under' the heavy atom. This density is zero if the replacement is isomorphous. It is actually the only position in the unit cell where the electron density is returned exactly by the error-free SIR synthesis. An image of the heavy atom develops as experimental errors bring the figure of merit down and systematically distort the mean $\langle F_t \rangle$. This error associated with the traces of heavy-atom structure is absent in the weighted correlation map (10) as the experimentally determined projections F_t are separated from their weights (merit factors) by the correlation formula.

Difference Patterson synthesis and heavy-atom search

The heavy-atom search in Patterson space has its equivalent in the reciprocal space (see, Argos & Rossmann, 1976). The sum of difference Patterson values on the Harker sections corresponding to a position $\mathbf{r}$ in real space can be rewritten in reciprocal space as,

$$\Sigma(F_{ph} - F_p)^2(t^2 - 1), \qquad (11)$$

where F_{ph} and F_p are the derivative and native protein structure-factor amplitudes. This expression comprises a product of the squared differences $dP = (F_{ph} - F_p)^2$ and origin-removed normalized trial intensity $t^2 - 1$. All symmetry-related positions in the unit cell are included in the calculation of $\mathbf{t}$ and the trial intensity is normalized to unity over infinite reciprocal space. For the ideal error-free case the mean trial intensity $\langle t^2 \rangle$ should be unity so the second term in (11) oscillates around zero and the formula represents an intensity correlation. In general the transformations from reciprocal into real space and *vice versa* are amplitude correlations. Intensity correlations are used for the heavy-atom search for mostly historical reasons and for clarity of interpretation. We propose to return to amplitudes for this particular case also. The experimental differences $dF = |F_{ph} - F_p|$ are correlated to the trial amplitudes t rather than intensities,

$$\rho = (I_P I_t)^{1/2}(\langle dFt \rangle_w - \langle dF \rangle_w \langle t \rangle_w)/[\sigma_w(dF)\sigma_w(t)]. \qquad (12)$$

The result of the calculation is a correlation map that shows heavy-atom positions directly. The correlation map is two times less sensitive to the experimental errors because the correlation function involves structure factors rather than their squares ($\delta I/I = 2\delta F/F$).

In the absence of experimental errors, the maximum possible correlation value is about 50% for the acentric reflections and 100% for centric. If the real correlation-map maxima are close to these target values then there are no more heavy atoms. If there is more than one heavy atom then the first atom is assigned to the highest peak in the map. The search is repeated again except that now t represents the diffraction from the atom already located together with another trial unit running through the map to give the secondary correlation map. The mean value of the second map is equal to the maximum value of the first map and the position of the secondary-map maximum reveals the next heavy atom. The procedure is repeated until the correlation map represents a random noise. The secondary correlation map is the reciprocal-space analogue of the Patterson self- plus cross-vector search.

Fig. 2 shows an example of the primary correlation map calculated for the real data. The single heavy-atom positions are clearly visible both on the Patterson difference synthesis and on the correlation map. The peak-to-noise ratio for the Patterson map is about 3 and for the correlation map it is about 6. The secondary correlation map reveals another four minor sites which are not recognisable in the double difference Fourier synthesis (not shown).

Difference Fourier synthesis

The formula for the difference Fourier synthesis of the unknown fraction B of the whole structure AB based on the phases from the known model A also consists of the product of two terms,

$$\langle (F_{AB} - F_A) \exp(i\alpha_A) \exp(-i\mathbf{kr}) \rangle$$
$$= \langle (F_{AB} - F_A) \exp(i\alpha_A)\mathbf{t}^* \rangle$$
$$= \langle (F_{AB} + F_A) t \cos(\alpha_t - \alpha_A) \rangle = \langle dFt_A \rangle$$
$$\simeq \langle F_B \cos(\alpha_B - \alpha_A) t \cos(\alpha_t - \alpha_A) \rangle. \qquad (13)$$

The first term is the experimentally available difference $dF = (F_{AB} - F_A)$. If $F_B << F_A$ then $dF \simeq F_B \cos(\alpha_B - \alpha_A)$ and the difference represents the projection of the unknown $\mathbf{F}_B$ onto the direction of $\mathbf{F}_A$. The Sim weights (Sim, 1959) reflect the reliability of this approximation. The second term comprises

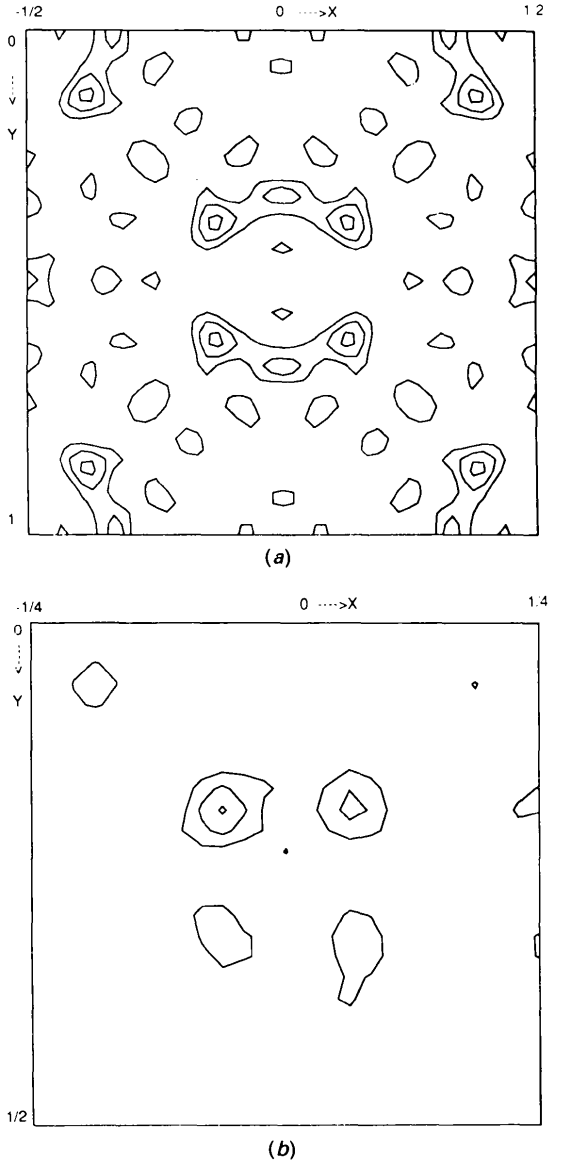

Fig. 2. The heavy-atom search for the crystals of human tissue factor, space group $P4_12_12$, $a = b = 45.2$, $c = 231.5$ Å (Harlos *et al.*, 1994). (*a*) The difference Patterson, Harker section $z = 1/2$. Eight symmetry-related equivalent peaks correspond to a single heavy atom in the asymmetric unit. The contour levels are 1, 2 and 3 r.m.s. (*b*) The correlation map section through the major heavy-atom position. The scale is twice the scale of the difference Patterson map so that the Patterson peaks overlap the corresponding peaks of the real-space correlation map. The strongest peak (correlation 21%, overall map r.m.s. 3.4%) corresponds to the true heavy-atom position, the contour levels being 2, 4 and 6 r.m.s. The weaker peaks are the ghosts symmetry related to the major peak. They disappear in the secondary correlation map. The secondary correlation map is flat, r.m.s. being 1.2% compared with 3.4% of the primary correlation map. It reveals four minor binding sites with occupancies 0.2–0.3 (not shown) which bring the correlation up to 24.6% and make the map perfectly flat (no big peaks, r.m.s. 0.8%). The signal/noise ratio for the correlation map is twice as good as that for the difference Patterson map.

the analogous projection t_A of the trial unit structure factor **t** onto the same vector $\mathbf{F}_A$. Thus, the difference synthesis can be treated as the correlation between projections of $\mathbf{F}_B$ and **t** onto the same vector $\mathbf{F}_A$. The replacement for the weighted correlation function is carried out as above with the Sim or other weights reflecting phase probabilities of the partial structure (Read, 1986),

$$\rho = (I_P I_t)^{1/2}[\langle \mathrm{d}F t_A\rangle_w - \langle \mathrm{d}F\rangle_w\langle t_A\rangle_w]/[\sigma_w(\mathrm{d}F)\sigma_w(t_A)]. \quad (14)$$

Once again, if there are no experimental errors (all weights are unity) and $F_B << F_A$ then the difference Fourier map does not reveal an image of the model F_A. The errors and weights, if present, are phased as $\mathbf{F}_A$ and introduces traces of the model *A* into the map. The correlation map by analogy to the SIR synthesis should not contain an image of the known model in contrast to the difference Fourier map.

Atomicity and symmetry averaging

If the trial wave **t** is replaced with an average atomic structure factor, then the correlation function will produce the probability of finding an atom at the specified position. If a protein is assumed to be composed of identical atoms, then the atomic scattering factors can be used in syntheses in the same way as they are used in the structure-factor calculations.

The correlation function becomes similar to the translation function with a single atom as a search fragment and, further, the Fourier synthesis itself can also be regarded as a translation function with a single scattering point as a primitive search fragment.

Non-crystallographic symmetry can be incorporated directly into a correlation synthesis once the borders of the symmetry-related regions are established. The calculation of a trial wave **t** is carried out according to whether a position **r** is subject to the local symmetry operations or not. The number of locally equivalent positions in the unit cell varies with **r** but the correlation stays between -1 and 1. This provides a rigorous reciprocal-space analogue for real-space map averaging procedures.

The correlation function and fast Fourier transformation

To be able to use the fast Fourier transformation algorithm (FFT) the correlation formula has to be rewritten in the form of standard Fourier transform-

ations but one term in the correlation formula is unsuitable for this. It is the mean weighted trial intensity $\langle \mathbf{tt}^* \rangle$ which is part of $\sigma_w(t)$ in (4). For the space group $P1$ it is unity irrespective of weights and data completeness. For other space groups it can vary and the FFT program itself has to be altered to calculate it. For our first qualitative calculations we have used a parallel computer and calculated all transformations directly without using an FFT algorithm. This has allowed the necessary flexibility at the present state of development of the correlation method, at the expense of considerable computer time. A faster program for a conventional computer will be developed later.

Discussion

The conventional Fourier formalism and the proposed correlation approach converge as the quality of the experimental data improves and the methods become absolutely identical for the ideal case of error-free data available to infinite resolution. The advantage of the correlation function is that it treats the systematic errors inherent in the conventional Fourier technique in a way that converts them into random error such that the resulting map is not distorted other than randomly.

The introduction of weighting into a Fourier synthesis is similar to the changing of the measurements themselves so that the weighted Fourier synthesis effectively uses a data set that is different from an original one and corresponds to some different systematically distorted structure. The noise level of the distorted map becomes less because the unreliable terms are suppressed. Thus, a map is cleaned up at the expense of its overall accuracy to the true map. If some of the reflections are not measured (or the weight is zero) then they are intrinsically assigned zero value and the Fourier synthesis returns an electron-density map that really corresponds to $F = 0$ for these reflections. If many strong reflections are not observed then the Fourier-synthesis map becomes significantly distorted. The weights are mixed with the experimentally determined structure-factor estimates and any weighting scheme introduces a map error which is generally avoided by the correlation approach.

The correlation function allows separation of the weights from the measurements themselves and removes the distortions. The weighting decreases the effective overall resolution of the map without biasing the statistical properties of the data. The lack of some of the measurements only reduces the effective number of observations and increases the statistical noise level without biasing the mean result. The weak and zero reflections are as important for the correlation function as the strongest ones in contrast to the Fourier synthesis where they can be omitted.

The suggested way of map calculation provides a universal approach to various methods of structure determination, as universal as Fourier transformation itself. Few examples are described in the present paper but other applications are equally straightforward. Computationally the same program can perform various syntheses, heavy-atom searches, non-crystallographic symmetry averaging, *etc.* To apply the program to a particular case one has to realise what information about complex structure factors is available from a particular experiment and how accurate that information is. In the MIR experiment, the full complex structure factor is determined. For the SIR (without anomalous data) and the difference Fourier synthesis only a projection of a complex structure factor onto a certain direction is available. Then the appropriate parameters are calculated for the trial wave and correlated with the experimental values. The correlation is higher if full structure factors are available. It is weaker if only a fraction of the full structure factor is known. The extreme case is when only structure-factor amplitudes are measured and no phase information is available. Then the correlation can be constructed between absolute values F and t. The situation is similar to the heavy-atom search: if no symmetry is present, then t is constant throughout and nothing can be done. If the space group is higher than $P1$ then t varies and the correlation might reveal a position of the molecule in the unit cell and molecular envelope if the solvent content is high.

The structure factors calculated from a correlation map by direct transformation can differ from the initial experimental structure factors that were used in correlation-map evaluation. But they should coincide for the ideal error-free data. The difference indicates the errors of the map. It might be possible to combine the recalculated structure factors with the initial experimental data, build the next map and repeat the process several times until it converges. This potential for map improvement looks promising but still has to be analysed in more detail.

The weighted correlation function uses the same data as the conventional Fourier but in a more rigorous way. If the errors are high then both methods fail. If they are low then both methods work well. However, the correlation method might allow interpretaion of maps produced from lower quality data. The extent of the improvement depends upon the specific experiment. It is obviously more sensible to collect better data than to spend time trying to improve maps computed with poor data. Unfortunately, it is often impossible and any improvement, however small, must therefore be considered welcome. The present paper is more an intro-

duction of the correlation approach rather than an extensive analysis of its merits and drawbacks which will be carried out elsewhere separately for each particular implementation.

I thank Dr W. Boys for permission to use the data on the human tissue factor, Drs T. Wess, M. Z. Papiz, I. Polikarpov and L. Sawyer for useful discussions and Dr P. Adams for help with parallel computing. The Edinburgh Parallel Computing Centre is thanked for providing the computer facilities and the SERC for financial support.

References

ARGOS, P. & ROSSMANN, M. G. (1976). *Acta Cryst.* B**32**, 2975–2979.

BLOW, D. M. & CRICK, F. H. C. (1959). *Acta Cryst.* **12**, 794–803.

HARLOS, K., MARTIN, D. M. A., O'BRIEN, D. P., JONES, E. Y., STUART, D. I., POLIKARPOV, I., MILLER, A., TUDDENHAM, E. G. D. & BOYS, C. W. G. (1994). *Nature (London)*, **370**, 662–666.

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

SAMUELS, M. L. (1989). *Statistics for the Life Sciences*, pp. 434–492. San Francisco: Dellen Publishing Company.

SIM, G. A. (1959). *Acta Cryst.* A**25**, 813–817.

WOOLFSON, M. M. (1970). *X-ray Crystallography*, pp. 84–87. Cambridge Univ. Press.