Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Generation, representation and flow of phase information in structure determination: recent developments in and around *SHARP* 2.0

G. Bricogne,[a]* C. Vonrhein,[a] C. Flensburg,[a] M. Schiltz[a,b] and W. Paciorek[a]

[a]Global Phasing Ltd, Sheraton House, Castle Park, Cambridge CB3 0AX, England, and [b]Université Paris XI, LURE, Bâtiment 209d, F-91898 Orsay CEDEX, France

Correspondence e-mail: gb10@globalphasing.com

Received 15 April 2003
Accepted 8 August 2003

The methods for treating experimental data in the isomorphous replacement and anomalous scattering methods of macromolecular phase determination have undergone considerable evolution since their inception 50 years ago. The successive formulations used are reviewed, from the most simplistic viewpoint to the most advanced, including the exploration of some blind alleys. A new treatment is proposed and demonstrated for the improved encoding and subsequent exploitation of phase information in the complex plane. It is concluded that there is still considerable scope for further improvements in the statistical analysis of phase information, which touch upon numerous fundamental issues related to data processing and experimental design.

1. Introduction

It is almost exactly 50 years ago that the potential of isomorphous replacement (Green *et al.*, 1954) and anomalous scattering (Bijvoet, 1954) to provide experimental phase information for macromolecules was identified. Since then, considerable progress has been made in realising that potential through developments in instrumentation (synchrotron radiation, area detectors), experimental protocols (crystal freezing, SeMet MAD, halide soaks) and computational methodology (solution of large heavy-atom substructures, maximum-likelihood refinement and phasing, density modification).

This paper gives a fairly informal historical survey of the successive treatments devised to extract optimal phase information from given experimental data and presents recent developments related to the encoding and further use of that phase information in the complex plane. Finally, directions for further developments are indicated.

2. Phase information from small amplitude differences

Phase information is derived from a comparison of several related sets of amplitude measurements and from the modelling of the differences between them in terms of a collection of 'heavy atoms' (*i.e.* additional or anomalous scatterers) whose number is considerably smaller than the number of available measurements.

In the ideal situation where no errors of any kind are present, consistency relations between the structure-factor contributions F_j^H from the heavy atoms, the available amplitude measurements $|F_j^{PH}|^{\rm obs}$ and the phased structure factor F^P

© 2003 International Union of Crystallography
Printed in Denmark – all rights reserved

for the macromolecule are expressed by a set of equations for each unique reflection **h**,

$$k(j, \mathbf{h})|F^P(\mathbf{h}) + F^H(j, \mathbf{h})| = |F^{PH}(j, \mathbf{h})|^{\rm obs}. \quad (1)$$

Here, j is a generic label which encodes book-keeping information about various isomorphous compounds, distinct crystals of these compounds, different X-ray wavelengths, successive time batches and the identity (+ or −) of members of a Bijvoet or Friedel pair, while the scale factor $k(j, \mathbf{h})$ relates the scale for observation j at **h** to the absolute scale. If an observation is available for the 'native' macromolecule, free from heavy atoms, it is customary to label it as $j = 0$ (say) and to put $F^H(0, \mathbf{h}) = 0$ for all **h**: the equations then determine the phases associated with the observed native amplitudes $|F^{PH}(0, \mathbf{h})|$. These equations are traditionally displayed in the form of Harker's construction (Harker, 1956) and it is well known from the geometry of circles that a pair (isomorphous or anomalous) of measurements typically gives a twofold ambiguous solution for F^P, while three or more give a unique solution.

Equations (1) involve two sets of quantities: firstly, the collection of parameters **p** involved in calculating the $F^H(j, \mathbf{h})$ and $k(j, \mathbf{h})$, which may be called global parameters since their influence is felt throughout reciprocal space (*i.e.* for all unique reflections **h**), and secondly, the collection of 'native' structure factors $\{F^P(\mathbf{h})\}$, which may be called local parameters since they each belong to a single reflection **h**. The very expression 'experimental phasing' highlights the extent to which the small collection of global parameters **p** are perceived as determining (perhaps with some residual twofold ambiguity) the much larger collection of phases contained in the local parameters $\{F^P(\mathbf{h})\}$ by the condition that together they should exactly 'explain' all data $\{|F^{PH}(j, \mathbf{h})|^{\rm obs}\}$ *via* (1).

3. Treatment of errors: a first glance

In a real situation, several categories of error will come and spoil the simplicity of equations (1). Firstly, the contributions $F^H(j, \mathbf{h})$ will be in error because the parameters describing the atoms in the current model do not have ideal values and possibly because the heavy-atom model for compound j is incomplete. Next, the contribution $F^P(\mathbf{h})$ from the macromolecule will not be the same for all j, as it will be affected by non-isomorphism between crystals, even in the case of an experiment involving only one crystal, because of the effects of radiation damage. Finally, the scaling parameters determining the relative scale factors $k(j, \mathbf{h})$ will not be exactly known and the observed amplitudes $|F^{PH}(j, \mathbf{h})|^{\rm obs}$ themselves will be affected by measurement errors of various origins. As a result, in general no collection of complex numbers $\{F^P(\mathbf{h})\}$ will exist such that equations (1) are satisfied. Blow & Crick (1959) were the first to propose a statistical treatment of this problem, arguing that if for fixed values of parameters **p** an arbitrary point $F^P(\mathbf{h})$ is chosen for each **h** and substituted into the left-hand side (LHS) of (1), then all discrepancies between the LHS and right-hand side (RHS) of (1) can be 'explained' by invoking the sources of error listed above. There are many such explanations or ways of 'apportioning blame' between the various types of error once **p** and $\{F^P(\mathbf{h})\}$ have been specified. Each of these explanations will have a different degree of plausibility, measured by the ability of the probability model for all sources of error affecting the LHS of (1) to account for the actual observations in its RHS within their own experimental accuracy. Such a measure of plausibility is called a *likelihood*: it will be considered in more detail and with more rigour later, but the term will be used in the meantime as shorthand for a hitherto unspecified 'measure of plausibility'.

4. The impasse of 'phase estimates'

At first sight, therefore, experimental phasing in the presence of errors seems to lead to a large-scale optimization problem in which a likelihood criterion should be maximized with respect to **p** and all the $F^P(\mathbf{h})$ values simultaneously. Such a problem, however, is quite uninviting: its objective function can be hopelessly multimodal, since it can be bimodal with respect to each $F^P(\mathbf{h})$ for given **p**. Furthermore, the parameter-to-observation ratio is unfavourable since there are two parameters $\{\Re[F^P(\mathbf{h})]$ and $\Im[F^P(\mathbf{h})]\}$ per reflection **h** in addition to the global parameters **p**. This optimization approach is therefore unworkable in a general case and the question arises of how to treat the local parameters $\{F^P(\mathbf{h})\}$ if they cannot be treated as refinable parameters.

Historically, the problem first arose with centric projection data for myoglobin (Dickerson *et al.*, 1960, 1961), for which two phase values are allowed but where many reflections **h** showed only one plausible $F^P(\mathbf{h})$ value by virtue of a simple 'no-crossover' argument which, it should be noted for later reference, is a probabilistic argument leading to a quasi-deterministic conclusion. In such cases, these sole plausible values could be called 'estimates' of $F^P(\mathbf{h})$ and substituted as known constants in (1). Sufficiently many such cases were available to allow the relatively few global parameters **p** to be refined by least squares (Hart, 1961) against the associated isomorphous differences with a comfortable observations-to-parameter ratio.

A major blind alley was entered when this protocol was extended to acentric reflections (where the phases now have unrestricted values) by setting up the least-squares refinement of global parameters **p** *via* equations (1) involving similar 'estimates' of acentric $F^P(\mathbf{h})$ values. The perils inherent in this approach were realised at the time and it was pointed out that a satisfactory refinement method for SIR along these lines may never be found because the intrinsic bimodality of the phase indications gave rise to two exactly equivalent candidates for the choice of each $F^P(\mathbf{h})$. When these indications were bimodal but not equivalent, the problem persisted and two main schools of thought developed: one in favour of the mode ('most probable phase') and another in favour of the centroid phase ('best phase') according to Blow & Crick (1959). In both cases, it was necessary to use only reflections with relatively high figures of merit (to avoid choosing a single mode out of two nearly equivalent ones in the first case and to

avoid using a centroid phase which is locally the least rather than most plausible phase in the second), so that there was little practical difference between the two. Successful computer programs were produced to implement this 'phased refinement' procedure (Dickerson *et al.*, 1968) and, as pointed out by Dodson (1976), made irresistible the temptation of refining global parameters **p** against initial estimates of $F^P(\mathbf{h})$, which would then be 'updated' once refinement had produced 'better' values for these parameters, the rationale being that better parameters **p** ought to give better estimates of $F^P(\mathbf{h})$, which in turn should give even better parameters **p** and so on. It seemed reasonable to expect that iteration of this procedure until self-consistency was reached would produce convergence to optimal parameter values. This hope had to be abandoned when Blow & Matthews (1973) noticed that the procedure led to serious bias problems and could even be unstable. Their analysis could be paraphrased by saying that choosing $F^P(\mathbf{h})$ estimates which were the most favourable to the current values of parameters **p**, together with restricting the data consulted during the refinement of **p** to well phased reflections only, amounted to 'gerrymandering', as a considerable number of degrees of freedom (N phases for N acentric reflections) were being continuously adjusted to absorb as much as possible the inconsistencies between model (LHS) and observations (RHS) in (1). When parameters refined in this way were subsequently used to phase all available data, the resulting electron-density maps were corrupted by prominent artefacts (typically large negative holes with associated series-termination ripples at heavy-atom sites), bearing witness to the fact that considerable systematic errors remained in the refined parameter values which were somehow being legitimized and hence made uncorrectable by the naïve procedure used.

5. Breaking the deadlock: Bayesian data analysis with marginalization

Blow and Matthews's alarm call resulted in a variety of defensive measures being taken against bias in phased refinement. Their own recommendation was a 'separation of powers' whereby the subset of parameters associated with each heavy-atom compound should only be refined against $F^P(\mathbf{h})$ estimates obtained from data for other compounds as unrelated as possible to that compound. Other measures tried to avoid the use of phase information altogether, such as the FHLE method (Kartha, 1965; Dodson *et al.*, 1975; Blundell & Johnson, 1976) and the origin-removed Patterson-correlation function of Terwilliger & Eisenberg (1983), at the cost of preventing multiple derivatives from assisting each other's refinement through the generation of phase information.

In examining the modern solution to this conundrum, it is worthwhile going back to the original treatment of the centric case in Dickerson *et al.* (1960, 1961) and reinterpreting the use of an estimate for unambiguous $F^P(\mathbf{h})$ values as a *single-point integration* rather than the substitution of a definite value. The correct generalization to the acentric case [and to the centric cases where the alternative value of $F^P(\mathbf{h})$ cannot be ruled out] is therefore an integration over the continuum of possible values of $F^P(\mathbf{h})$, each weighted by its likelihood according to the error model mentioned earlier, rather than the substitution of a single 'acentric $F^P(\mathbf{h})$ estimate' into equations (1). This integration brings the same benefits in the general case as the use of estimates did in the special case of amenable centric reflections, namely the elimination of all local parameters from the refinement and hence the dramatic improvement in the observations-to-parameter ratio in the refinement of the global parameters **p** alone compared with that for the simultaneous refinement of both local and global parameters considered at the beginning of §4.

In the terminology of modern Bayesian statistical methods (see, for instance, the excellent introduction by Sivia, 1996), the local parameters $\{F^P(\mathbf{h})\}$ are a typical instance of so-called *nuisance parameters*, *i.e.* of quantities which must be introduced in order to relate the parameters of interest **p** to the observations in the RHS of (1), but which are in themselves of no intrinsic interest. The process of integrating over nuisance parameters to get rid of any direct dependence on them of a likelihood involving **p** is called *marginalization* with respect to these parameters.

It may seem paradoxical to call the $F^P(\mathbf{h})$ values by the unflattering name of 'nuisance parameters' and declare them to be 'of no intrinsic interest', when they are in fact the desired end-product of the whole analysis! The crux of the argument is that for the purpose of determining the optimal values of the global parameters **p** (which in turn give rise to the final phase information), the $F^P(\mathbf{h})$ values do have to be integrated out so as to fully represent, within the refinement process, the degree to which they remain ambiguous at any given stage. The use of 'estimates' for these local parameters, by trying to summarize the probability distribution of each of them by a single value, was clearly doomed to engender fatal degrees of bias.

To conclude this retrospective sketch of the evolution of ideas in phase determination, it is also worthwhile noting that one of the early remedies proposed against phase-mediated bias is related to marginalization, albeit to a greater degree than necessary. The method of Terwilliger and Eisenberg does indeed consist of integrating out the phase difference between $F^P(\mathbf{h})$ and $F^H(j, \mathbf{h})$ with a uniform probability distribution in the expression for the expected squared isomorphous difference between the native and compound j and then refining the global parameters **p** against those squared differences. A similar treatment can be applied to anomalous differences. When several compounds are available, this approach leads to *separate* marginalizations with respect to all such phase differences, whereas they all involve the same phase for $F^P(\mathbf{h})$. The pattern of interaction between the various compounds which arises from a more careful marginalization with respect to the phase of $F^P(\mathbf{h})$ *simultaneously* was derived analytically by Bricogne (1991*a*,*b*) and showed that the benefits of the interactions between different compounds during parameter refinement could be recovered while avoiding the pitfalls of the old 'phased refinement'. Together with the treatment of non-isomorphism through the model of Luzzati (1952) proposed by Read (1991) and the implementation by Otwi-

nowski (1991) of numerical integration over the native phase in a least-squares program previously used for phased refinement, this analysis forms the basis of the modern maximum-likelihood approach to experimental phasing embodied in *SHARP* (de La Fortelle & Bricogne, 1997), which will now be described in more detail.

6. The mechanics of likelihood functions and marginalization

Returning to the situation of §3, any given physically reasonable values of **p** and $\{F^P(\mathbf{h})\}$ may be considered as an 'explanation' of the data in the RHS of (1) to the extent to which the various sources of errors listed there can account for the discrepancies between the two sides of equations (1). The question is to determine to what extent those data contain 'evidence' that privileges some of these explanations over others. This viewpoint is the standard setting for applying the Bayesian 'calculus of evidence' enshrined in the notion of likelihood and in Bayes's theorem, for which the reader is again referred to Sivia (1996).

The first step is to build a probabilistic model of all the relevant sources of error in the form of the joint probability distribution of all complex quantities of the form

$$F^{PH}(j,\mathbf{h})^{\rm calc} = k(j,\mathbf{h})[F^P(j,\mathbf{h}) + F^H(j,\mathbf{h})] \qquad (2)$$

for given values of the global parameters **p** and local parameters $\{F^P(\mathbf{h})\}$ under the effect of all these errors. Here, the quantities $F^P(\mathbf{h})$ denote the contribution of the macromolecule in the various compounds j, whose distribution in terms of a common underlying $F^P_\star(\mathbf{h})$ (the unknown 'trial native structure factor') is obtained by using Luzzati's model (Luzzati, 1952) separately for each j.

Specifying this joint distribution will call upon new classes of global parameters, which will be denoted collectively by **q**, describing for instance the incompleteness and imperfection of the current heavy-atom models, the non-isomorphism between different crystals or the effects of radiation damage on each crystal. Since the only observations available are structure-factor amplitudes, this joint distribution of complex structure factors must be converted into a joint distribution of measurable amplitudes $|F^{PH}(j,\mathbf{h},\mathbf{p})^{\rm calc}|$ by integration over all the associated phases. Once this has been performed, substituting the observed values $\{|F^{PH}(j,\mathbf{h})^{\rm obs}|\}$ of those amplitudes for the arguments $|F^{PH}(j,\mathbf{h})^{\rm calc}|$ of that joint distribution of amplitudes will produce the likelihood

$$\lambda = \lambda[\mathbf{p}, F^P_\star(\mathbf{h}), \mathbf{q}|\{|F^{PH}(j,\mathbf{h})|^{\rm obs}\}] \qquad (3)$$

for the 'explanation' or hypothesis described by **p** and $F^P_\star(\mathbf{h})$ under an error model with parameters **q** in the light of the available data. Technically speaking, the transition from the joint distribution of measurable amplitudes to the likelihood function is not a simple substitution, but requires an extra integration over the experimental error model for the observations. Finally, according to the analysis in §5, all local parameters $F^P_\star(\mathbf{h})$ must now be considered as nuisance parameters and integrated out to yield the likelihood function best suited for refining the global parameters **p** and **q** against the data,

$$\Lambda = \Lambda[\mathbf{p}, \mathbf{q}|\{|F^{PH}(j,\mathbf{h})|^{\rm obs}\}]. \qquad (4)$$

Once the optimal values **p*** and **q*** have been obtained by maximization of Λ, the likelihood function (3) calculated for $\mathbf{p}=\mathbf{p}^*$ and $\mathbf{q}=\mathbf{q}^*$ as a function of the $F^P_\star(\mathbf{h})$ gives the final form of the experimental phase information extracted from the data by means of the heavy-atom model and error model (for more details, see de La Fortelle & Bricogne, 1997). The centroid of that distribution may then be used in the calculation of electron-density maps in the usual way. More details will be given in Flensburg *et al.* (2003).

In practice, various approximations are made in *SHARP* to render the construction of the likelihood criteria more tractable. The error model used in building the joint distribution of complex structure factors (2) assumes that the effects of all sources of non-isomorphism are uncorrelated between different reflections **h**, so that the likelihoods are products of factors for the various reflections and can be handled through log-likelihoods which are additive over reflections. The current version of *SHARP* also assumes independence of non-isomorphism between the different values of j for each given **h**, an assumption which is plainly unjustified in some cases. This results in the further simplification that the integrations over the phases of the complex structure factors $F^{PH}(j,\mathbf{h},\mathbf{p})^{\rm calc}$ can be performed separately for each j. Similarly, the integration over the observational error distribution for each data item $|F^{PH}(j,\mathbf{h})^{\rm obs}|$ is carried out separately for each j and **h**, thus assuming uncorrelated measurement errors for all data items. None of these approximations is essential. Expressions for general likelihood functions capable of accommodating arbitrary patterns of covariance between the various sources of error in the $F^{PH}(j,\mathbf{h},\mathbf{p})^{\rm calc}$ values have been published by Bricogne (2000). These new functions are multivariate generalizations of the Rice likelihood which has played a fundamental role in all developments to date and they will underpin future developments aimed at going beyond the present approximations.

7. Recent developments in *SHARP* 2.0

Since the first release of *SHARP* in 1996, the distinguishing features of the program have been the full two-dimensional integration of the likelihood function over the complex local parameter $F^P_\star(\mathbf{h})$ (the 'trial native structure factor') and the use of a full Hessian matrix **H** of partial derivatives along with the gradient vector **g** in the maximization of the log-likelihood $L = \log\Lambda$. The integration over $F^P_\star(\mathbf{h})$ must therefore be carried out in such a way as to yield accurate values not only for the values of L, but also for its first- and second-order derivatives with respect to all the global parameters **p** and **q** on which it depends. This is a computationally demanding process, requiring on the order of 100 or more integration points. This made version 1 of *SHARP* a slow program, which tended to be used only as a weapon of last resort on difficult problems where all other programs had failed to produce any

Table 1
Speed-up factor for *SHARP* 2.0 single-processor code for various tasks.

Task	Sites	Batches	Old (min)	New (min)	Speed up
14 test jobs			6873	525.4	13
Bubble	8	1	40	3.3	12
Cyanase	40	4	4417	168.6	26
KPHMT†	160	2	1478	24.1	61

† One function, gradient and Hessian evaluation.

Table 2
Further speed-up factors from OpenMP parallelized code on various platforms.

Architecture	CPUs	Bubble†, time (s)	Cyanase†, time (s)	IF3-C‡, time (s)	Speed up
ES45 1 GHz	1	28.7	496.1	359.1	
	2	15.4	257.6	185.9	1.91
ES40 500 MHz	1	57.1	1013.7	717.7	
	2	30.7	528.6	373.6	1.90
	3	21.9	365.4	260.7	2.71
	4	17.9	319.1	221.8	3.20
AMD M2000+	1	22.9	413.7	278.4	
	2	12.8	229.1	145.8	1.84
PII 333 MHz	1	153.7	2714.7	1934.4	
	2	80.8	1422.9	998.2	1.92

† One function, gradient and Hessian evaluation. ‡ Complete job.

useful results. Considerable effort has since been expended to rewrite the code almost entirely so as to gain speed without sacrificing accuracy. Full details will be published elsewhere (Flensburg *et al.*, 2003), but Tables 1 and 2 give an idea of the respective speed gains achieved for the single-processor code and for a parallel version of the code using OpenMP threads. In the case of KPHMT, for instance, the parallel version of *SHARP* 2.0 now runs over 200 times faster on a four-processor machine than *SHARP* 1.4.0 did on a single processor of the same machine. It also produces significantly better results.

8. Representation and transfer of phase information: beyond ABCDs

The experimental phase information or, more precisely, the two-dimensional structure-factor information generated in *SHARP* is embodied in the posterior probability density $P^{\rm post}(F^P_\star)$ for each $F^P_\star(\mathbf{h})$, which according to Bayes's theorem is proportional to the likelihood (3) for optimal parameter values $(p^\star, q^\star)$ if it is assumed that the maximum of λ at that point is infinitely sharp,

$$P^{\rm post}[F^P_\star(\mathbf{h})] \propto \lambda[\mathbf{p}^\star, F^P_\star(\mathbf{h}), \mathbf{q}^\star, \{|F^{PH}(j,\mathbf{h})|^{\rm obs}\}]. \qquad (5)$$

This information is ordinarily not used as such, but is summarized to various degrees for various purposes. For map calculation the 'best' Fourier coefficient of Blow & Crick (1959) is still in universal use, while for phase combination the *ABCD* coefficients of Hendrickson & Lattman (1970) are the established standard. Both entities are a legacy of the Blow and Crick treatment of errors in the context of the MIR method, in the sense that they refer to a native 'phase circle' centred at the origin of the complex plane with a fixed error-free radius and to a phase which is a polar angle defined from that same origin. These definitions clearly need revising to accommodate the two-dimensional nature of the structure-factor probability information contained in $P^{\rm post}(F^P_\star)$.

The mildest extension of the traditional Blow and Crick picture would be to replace the error-free radius for the native structure factor by a sharply peaked distribution for a native amplitude referred to the origin. This would preserve the key feature of the *ABCD* representation, namely that the two-dimensional probability density for the distribution of $F^P_\star$ should be a *direct product* of an amplitude-dependent part and a phase-dependent part,

$$P[F\exp(i\varphi)] \propto P_{\rm rad}(F)P_{\rm ang}(\varphi). \qquad (6)$$

Unfortunately, this is not the case. Under the current approximation where the various sources of non-isomorphism are assumed to be independent, the posterior probability density for each $F^P_\star(\mathbf{h})$ according to (3) is a product of radially symmetric Rice distributions centred at $-F^H(j,\mathbf{h})$ for each j (equation 23 in de La Fortelle & Bricogne, 1997). In many instances, it can be shown that this distribution is concentrated near an 'optimal circle' which does not coincide with the native circle even if the latter is present. More specifically, this circle is centred at $-F^{\rm off}(\mathbf{h})$, a weighted average of the various $-F^H(j,\mathbf{h})$, which is called the 'offset' in the sequel, and its radius is the expectation value of $|F^P_\star(\mathbf{h}) + F^{\rm off}(\mathbf{h})|$ under the distribution $P^{\rm post}(F^P_\star)$. It is optimal in the sense that it minimizes the cross-talk between radial and angular information and is therefore the best circle around which to approximate the true two-dimensional distribution $P^{\rm post}(F^P_\star)$ in direct-product form (6). For this purpose, a radial integration of the two-dimensional distribution is carried out along radii emanating from the offset. The logarithms of the resulting marginal probabilities are then Fourier analysed with respect to the polar angle (also referred to the offset) to produce *ABCD* coefficients encoding the phase information around the optimal circle. A description of that circle must accompany the *ABCD* coefficients to allow the regeneration of the initial two-dimensional distribution up to the approximation inherent in the direct-product form (6) of its reconstruction.

An enriched model must therefore be defined involving eight parameters: two coordinates for the offset $F^{\rm off}(\mathbf{h})$, defining the centre of the optimal circle, the radius of that circle, a standard deviation describing the dispersion of the radial distribution of $|F^P_\star(\mathbf{h}) + F^{\rm off}(\mathbf{h})|$ along that radius and the four *ABCD* coefficients encoding the angular dependence of the marginal probability obtained by integrating the two-dimensional distribution along radii of the circle. Further details on the implementation of this enriched model in *SHARP* 2.0, including the definition of figures of merit for two-dimensional distributions, will be given in Flensburg *et al.* (2003).

This eight-parameter representation of two-parameter structure-factor distributions offers the possibility of transfering a more faithful summary of experimental phase information to subsequent steps of structure determination, provided these steps themselves are able to handle it.

9. Application to an extended density-modification protocol

It has become customary to process 'raw' experimental phase information in order to improve it before computing an electron-density map for visual inspection and interpretation. This is especially necessary in SIR or SAD situations, where that raw information remains highly bimodal.

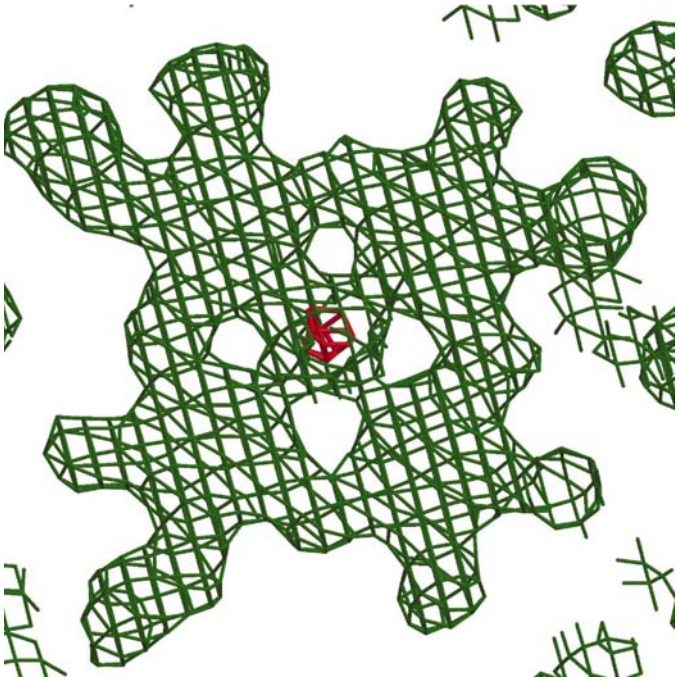

Figure 1
Haem and Fe atom in Mb: *SHARP* + *DM*.

The most common form of post-processing consists of phase improvement and extension through density modification, as exemplified by programs such as *DM* (Cowtan, 1994) and *SOLOMON* (Abrahams & Leslie, 1996). However, the underlying protocols are still based on 'paradigms' inherited from the MIR era: not only do they use only *ABCD* coefficients to represent phase information (see §8), but they implicitly assume that the electron-density map to be modified is that of the 'native' macromolecule, only exceptionally containing heavy atoms for which no special treatment is provided. The real-space properties imposed upon this electron-density map during density modification are based on this viewpoint, which is clearly not well suited to modern methods such as MAD or SAD where heavy atoms are systematically present in the macromolecule, whereas density-modification procedures are tuned to structures containing light atoms only.

The eight-parameter representation of two-dimensional structure-factor distributions offers a natural solution to this problem, which has been implemented in the density-modification step (based on *SOLOMON*) in the current versions of *SUSHI* and *autoSHARP* (Vonrhein, Blanc *et al.*, 2003).

The main feature of this new treatment is the handling of the offset, which roughly speaking corresponds to a sort of average heavy-atom structure taken over all compounds (further details will be given in Flensburg *et al.*, 2003 and Vonrhein, Schiltz *et al.*, 2003). It is taken out to compute the structure factors to which the density-modification procedure is applied, so that the latter operates only on electron density for light atoms; it is then re-applied to ensure that phase combination in *SIGMAA* (Read, 1986) takes place around the

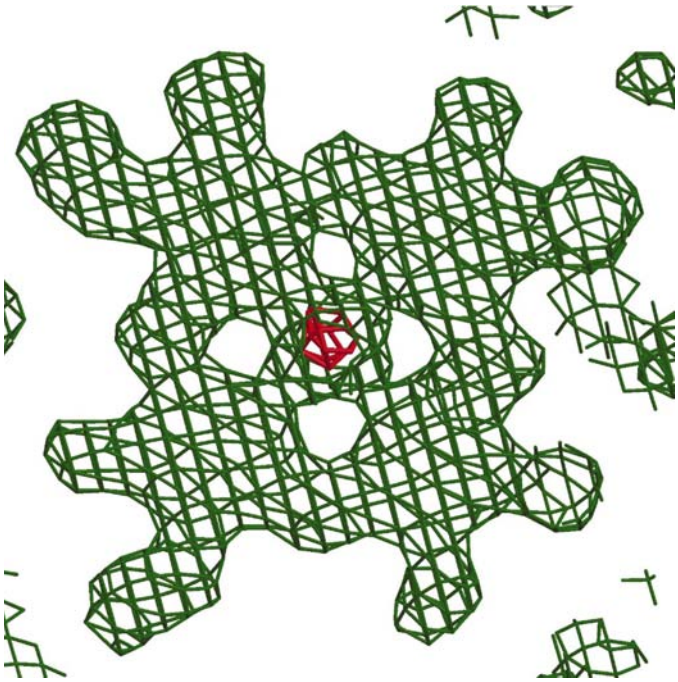

Figure 2
Haem and Fe atom in Mb: *SHARP* + *SOLOMON*, standard *ABCD*s.

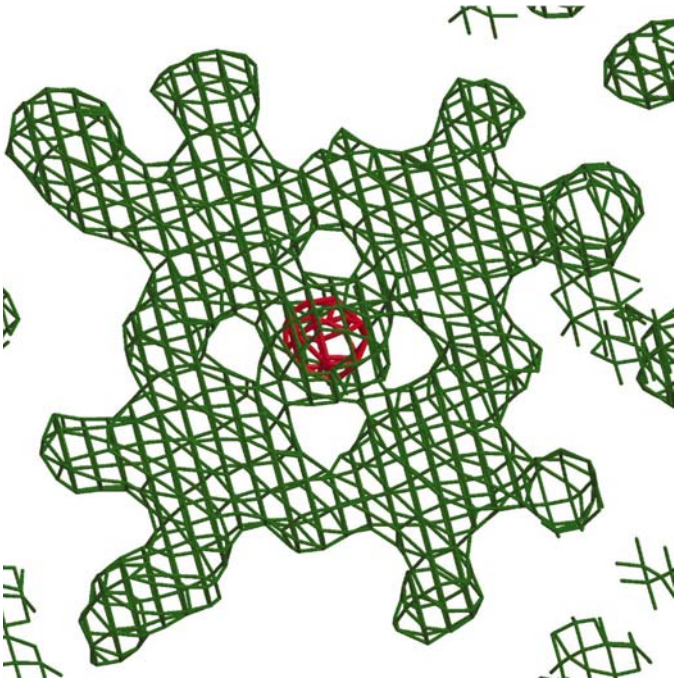

Figure 3
Haem and Fe atom in Mb: *SHARP* + *SOLOMON*, *ABCD*s with offsets.

optimal circle, where the direct-product assumption of the Hendrickson–Lattman model is best fulfilled. In this way, the heavy atoms do not interfere with the density-modification process nor suffer as a result of it.

As a first test, SAD data for *P*6 myoglobin (Mb for short), collected in-house at Cu $K\alpha$ wavelength to 1.8 Å resolution, were used in *SHARP* 2.0 to produce a refined heavy-atom model for the single Fe atom and two-dimensional probability distributions, which were then encoded into the eight-parameter model described in §8. To resolve the twofold ambiguity of each acentric phase, this information was then used to carry out density modification with *SOLOMON* or with *DM* without special treatment of the offset (*i.e.* leaving the heavy atoms in the maps subjected to density modification) and with *SOLOMON* with proper treatment of the offset (*i.e.* taking out the heavy atoms before density modification and putting them back after). In all cases, the combination of phase information was carried out on the optimal circle. The electron density for the haem and the Fe atom is shown in Figs. 1, 2 and 3. With *ABCD*s only, the *DM* map gives a peak height of 13.3σ and the *SOLOMON* map 13.8σ. With *ABCD*s and offsets, *SOLOMON* gives a peak height of 21.7σ. The correct peak height, inferred from a map generated from the refined model for Mb with bulk-solvent correction, is 19.8σ. The extended density-modification protocol using the offset information together with the *ABCD*s therefore produces better density around the heavy atom than does the use of *ABCD*s alone. This example may seem rather academic, but the same extended protocol was responsible for the considerable improvement in the density for the extracellular domain of the LDL receptor (reported by Rudenko *et al.*, 2003) near the 12 tungstophosphate clusters used to phase that structure (see Fig. 6 in that paper).

As a second test, the ToxD data set distributed with *CCP*4 4.2.2 as an example of a well behaved MIR phase determination with *MLPHARE* was used similarly to compare *SHARP* and *MLPHARE*, *SOLOMON* and *DM* with *ABCD*s only and *SOLOMON* with *ABCD*s supplemented with offset information. When *MLPHARE* is used as a phasing program, both *DM* and *SOLOMON* use standard *ABCD*s. When *SHARP* is used for phasing, *DM* uses *ABCD*s without offsets, while *SOLOMON* uses *ABCD*s with their associated offset information. The results are summarized in Figs. 4 and 5 and show that the best results are obtained across the whole resolution range with *SHARP* + *SOLOMON* using the offset information.

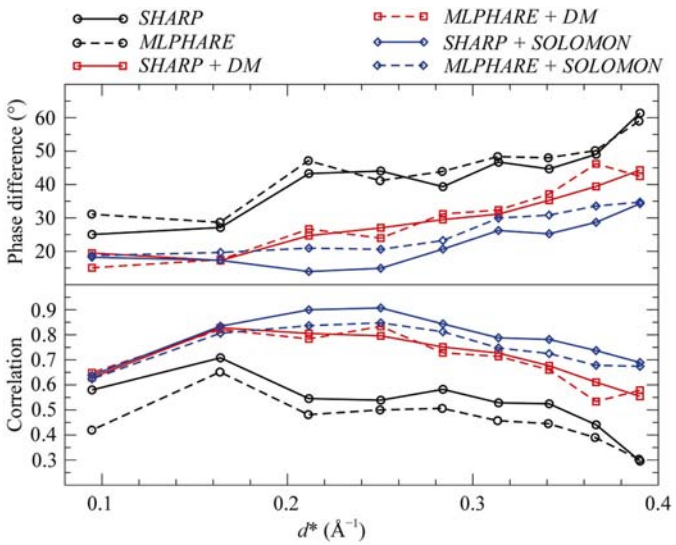

Figure 4
Phased correlation coefficient and weighted mean absolute phase errors.

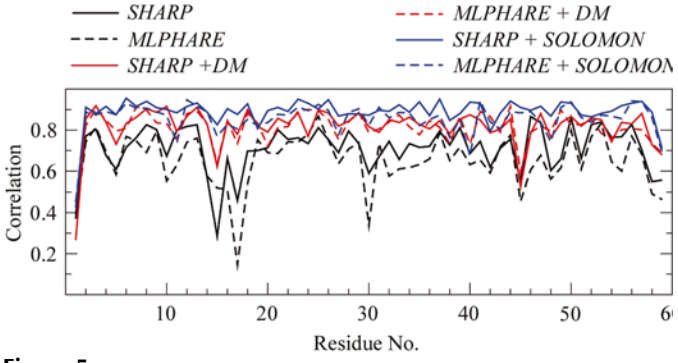

Figure 5
Real-space correlation coefficients along the polypeptide chain.

10. Outlook

The quick survey of phasing methodology given here should by now have made it clear that our ability to extract experimental phase information from isomorphous replacement and anomalous scattering data had been limited by our ability to identify the correct statistical framework within which to treat the problem. The potential 'phasing signals' are given by small differences between data sets, which are affected by numerous and sometimes highly correlated sources of error, some acting on complex contributions to overall structure factors and others on measurements of structure-factor amplitudes through diffraction intensities.

In order to progress beyond the current 'state of the art', as represented for instance by the present capabilities of *SHARP* 2.0, a number of further advances are necessary.

(i) We need to build and exploit better models for sources of error or uncertainty on complex structure factors, such as heavy-atom model errors, non-isomorphism in the macromolecule and the effects of radiation damage on both heavy atoms and macromolecule, including all possible correlations between the different copies of these errors affecting the various compounds available, and to carry out the integrations over all relevant phases while representing all these correlations: a task for which the multivariate likelihood functions derived in Bricogne (2000) should prove most valuable.

(ii) We also need to build and exploit better error models for the observations themselves and especially on the correction factors and instrumental factors applied to the raw measurements during data processing. The statistical structure of these measurement errors will play a crucial role in future developments. On the one hand, it will depend greatly on the

precise strategy according to which the data were collected; on the other hand, it will need to be taken into account in the integration operation through which the data are incorporated into joint distributions of amplitudes to compute likelihood functions (equation 3 in §6) and will thus directly influence the generation of phase information from these data. Through these dual roles, the observational error model will provide the natural channel through which to design optimal experiments so as to maximize the expected signal-to-noise ratio of a given type of phasing signal.

There remains a considerable amount of work to be performed in formulating and implementing of the necessary statistical methodology, so that we can look forward to numerous reconvenings of this *CCP*4 Study Weekend on Experimental Phasing for years to come.

We wish to thank Dr Hans Parge (Agouron, San Diego, California) for making his in-house SAD data on myoglobin available to us, the members of the Global Phasing Consortium for financial support and much scientific feedback and our colleagues at Global Phasing (Drs Pietro Roversi, Eric Blanc, Richard Morris and Gwyndaf Evans) for numerous helpful discussions. We also wish to acknowledge partial financial support for this work from European Commission Grant No. QLRT-CT-2000-00398 within the AUTOSTRUCT project. Finally, we wish to thank both referees, whose comments greatly helped to improve the manuscript.

References

Abrahams, J. P. & Leslie, A. G. W. (1996). *Acta Cryst.* D**52**, 30–42.
Bijvoet, J. M. (1954). *Nature (London)*, **173**, 888–891.
Blow, D. M. & Crick, F. H. C. (1959). *Acta Cryst.* **12**, 794–802.
Blow, D. M. & Matthews, B. W. (1973). *Acta Cryst.* A**29**, 56–62.
Blundell, T. L. & Johnson, L. N. (1976). *Protein Crystallography*. New York: Academic Press.
Bricogne, G. (1991*a*). *Crystallographic Computing 5*, edited by D. Moras, A. D. Podjarny & J. C. Thierry, pp. 257–297. Oxford: Clarendon Press.
Bricogne, G. (1991*b*). *Proceedings of the CCP4 Study Weekend. Isomorphous Replacement and Anomalous Scattering*, edited by W. Wolf, P. R. Evans & A. G. W. Leslie, pp. 60–68. Warrington: Daresbury Laboratory.
Bricogne, G. (2000). *Proceedings of the Workshop on Advanced Special Functions and Applications, Melfi (PZ), Italy, 9–12 May 1999*, edited by D. Cocolicchio, G. Dattoli & H. M. Srivastava, pp. 315–321. Rome: Aracne Editrice.
Cowtan, K. (1994). *Jnt CCP4/ESF–EACBM Newsl. Protein Crystallogr.* **31**, 34–38.
Dickerson, R. E., Kendrew, J. C. & Strandberg, B. E. (1960). *Symposium on Computer Methods and the Phase Problem*, p. 84. Glasgow: Pergamon Press.
Dickerson, R. E., Kendrew, J. C. & Strandberg, B. E. (1961). *Computing Methods and the Phase Problem in X-Ray Crystal Analysis*, edited by R. Pepinsky, J. M. Robertson & J. C. Speakman, pp. 236–251. Oxford: Pergamon Press.
Dickerson, R. E., Weinzierl, J. E. & Palmer, R. A. (1968). *Acta Cryst.* B**24**, 997–1003.
Dodson, E. J. (1976). *Crystallographic Computing Techniques*, edited by F. R. Ahmed, pp. 259–268. Copenhagen: Munksgaard.
Dodson, E. J., Evans, P. R. & French, S. (1975). *Anomalous Scattering*, edited by S. Ramaseshan & S. C. Abrahams, pp. 423–436. Copenhagen: Munksgaard.
Flensburg, C., Schiltz, M., Paciorek, W., Vonrhein, C. & Bricogne, G. (2003). In preparation.
Green, D. W., Ingram, V. M. & Perutz, M. F. (1954). *Proc. R. Soc. London Ser. A*, **225**, 287–307.
Harker, D. (1956). *Acta Cryst.* **9**, 1–9.
Hart, R. G. (1961). In *The Crystal Structure of Myoglobin: Phase Determination to a Resolution of 2 Å by the Method of Isomorphous Replacement* [Dickerson, R. E. Kendrew, J. C. & Strandberg, B. E. (1961), *Acta Cryst.* **14**, 1188–1195], pp. 1194–1195.
Hendrickson, W. A. & Lattman, E. E. (1970). *Acta Cryst.* B**26**, 136–143.
Kartha, G. (1965). *Acta Cryst.* **19**, 883–885.
La Fortelle, E. de & Bricogne, G. (1997). *Methods Enzymol.* **276**, 472–494.
Luzzati, V. (1952). *Acta Cryst.* **5**, 802–810.
Otwinowski, Z. (1991). *Proceedings of the CCP4 Study Weekend. Isomorphous Replacement and Anomalous Scattering*, edited by W. Wolf, P. R. Evans & A. G. W. Leslie, pp. 80–85. Warrington: Daresbury Laboratory.
Read, R. J. (1986). *Acta Cryst.* A**42**, 140–149.
Read, R. J. (1991). *Proceedings of the CCP4 Study Weekend. Isomorphous Replacement and Anomalous Scattering*, edited by W. Wolf, P. R. Evans & A. G. W. Leslie, pp. 69–79. Warrington: Daresbury Laboratory.
Rudenko, G., Henry, L., Vonrhein, C., Bricogne, G. & Deisenhofer, J. (2003). *Acta Cryst.* D**59**, 1978–1986.
Sivia, D. S. (1996). *Data Analysis. A Bayesian Tutorial.* Oxford: Clarendon Press.
Terwilliger, T. C. & Eisenberg, D. (1983). *Acta Cryst.* A**39**, 813–817.
Vonrhein, C., Blanc, E., Roversi, P. & Bricogne, G. (2003*a*). In preparation.
Vonrhein, C., Schiltz, M., Flensburg, C., Paciorek, W. & Bricogne, G. (2003*b*). In preparation.