# Acknowledging Errors: Advanced Molecular Replacement with Phaser

Airlie J. McCoy

Department of Haematology, University of Cambridge, Cambridge Institute for Medical Research, Wellcome Trust/MRC Building, Hills Road, Cambridge CB2 0XY, UK

Correspondence e-mail: ajm201@cam.ac.uk

**Abstract**

Molecular replacement is a method for solving the crystallographic phase problem using an atomic model for the target structure. State-of-the-art methods have moved the field significantly from when it was first envisaged as a method for solving cases of high homology and completeness between a model and target structure. Improvements brought about by application of maximum likelihood statistics mean that various errors in the model and pathologies in the data can be accounted for, so that cases hitherto thought to be intractable are standardly solvable. As a result, molecular replacement phasing now accounts for the lion's share of structures deposited in the Protein Data Bank. However, there will always be cases at the fringes of solvability. I discuss here the approaches that will help tackle challenging molecular replacement cases.

**Keywords:** molecular replacement, maximum likelihood, LLGI

**Running head:** Molecular replacement with Phaser

## 1 Introduction

As originally conceived [1–3], the aim of molecular replacement (MR [4]; Note 1) was to correctly orient and place a model that had high homology to the target and represented the bulk of the scattering, for the purpose of phasing. It has since been generalized to cases of targets being modelled by any number of components with any homology to the target, and each component representing any fraction of the scattering in an asymmetric unit [5]. The central problem of MR is to identify the correct placement (where *placement* refers to the three orientation angles and the three translation coordinates) of all model components in the asymmetric unit, with the hope that the resulting phases will be good enough to see novel features of the target structure and for iterative cycles of model building and refinement to commence [6].

MR consists of two aspects: a search procedure, for sampling orientations and translations of the model(s) in the crystal asymmetric unit; and a scoring function, for determining the (best) match of the structure factors calculated from the oriented and positioned model(s) to the observed structure factors, and hence the correct placement of the components. If the model is good and the data extend to high resolution and are free of pathologies, MR is successful with any of the implemented search strategies and scoring functions (X-PLOR [7], CNS [8], AMoRe [9], MOLREP [10], EPMR [11], Qs [12], SOMoRe [13], COMO [14], and Phaser [15]), each with their own strengths [16].

When it works, the speed and automation of MR rivals that of the direct methods used for small molecule crystallography, but it has a dark side. Because it is a search procedure, the success or failure of the method depends on the signal-to-noise of the correct placement, which depends on the quality of the model and data. Quick when it works with the first model and dataset input, it can be prohibitively slow if it does not, leading to an ever-increasing drain of computational resources. Paradoxically, successful MR strategies include knowing when to stop searching and attempt other structure solution methods.

With the extension of the Protein Data Bank ([17] PDB) to cover much of fold space, the chances are good that there will be a structure already in the PDB with the same fold as the target

protein [18]. Despite this, it is still common for MR models to have very low or even barely detectable sequence identity with the target (Note 2). Statistically, this is not a surprise, given the uncountable number of ways proteins can diverge in sequence from one another. It is also natural that researchers choose to crystallize proteins only when they require novel structural information.

Although much smaller, the database of nucleic acid and nucleic acid-protein complexes also offers a wealth of opportunity for MR phasing, partly because nucleic acid helices can adopt similar conformations with drastically different sequences, and because it is now recognised that there are nucleic acid structural building blocks [19].

## 2 Protocols

The aim of this review is not to provide a set of proscriptive protocols for MR. I assume that the reader is familiar with basic MR theory and practice. When MR is non-trivial, no two pathways to structure solution will be identical. Apart from the crystal-specific differences, there is the constantly changing background of instrumentation and software. Therefore, I aim to describe approaches to difficult cases that can be flexibly adapted to the problem at hand.

## 3 Overview

This review is directed at maximum likelihood MR (MLMR), and specifically the use of the LLGI (Log-Likelihood Gain on Intensity) target introduced this year [20]. MLMR scoring methods are superior in discriminating correctly from incorrectly placed models than Patterson methods [21]. LLGI adds the ability to account for experimental error in the data to the well-established ability of MLMR to account for errors in the models. It removes biases in MLMR targets formulated in terms of structure factor amplitudes, where the very poorly measured reflections are not appropriately down-weighted. LLGI has the correct asymptotic behaviour for data with infinite experimental error: these data have no contribution to the total LLGI. LLGI abolishes the need for the conversion of intensity data to amplitudes (usually performed with the French and Wilson method [22]) before MR.

Most of the problems with MR arise when there is a need to place a large number of components in the asymmetric unit, particularly if there is also low structural homology between models and targets. These situations may be engendered by the choice of crystallization target, for example, a macromolecular complex for which the structures for individual components, in isolation, are known, but not the complex in its entirety; or it may come about because the crystal happens to grow with many copies of the macromolecule in the asymmetric unit; or it may arise because the crystallographer chooses to attempt MR with small, generic, structural elements. Large errors are intrinsic to these problems, which is why MLMR targets are well suited to tackling them.

It is also possible for cases that seemed likely to be trivial at the outset to turn out to be fiendishly difficult, due to particular pathologies. MR is increasingly being attempted with crystals that are inherently twinned, show highly anisotropic diffraction and/or have translational non-crystallographic symmetry. MLMR approaches account for the intensity modulations arising from anisotropy and translational non-crystallographic symmetry, and the use of the LLGI target correctly weights the weak data with high error that are intrinsic to these data.

The most critical difference between MLMR approaches and Patterson approaches to MR is that MLMR is optimized when both the mean of the distribution *and the standard deviation of the distribution* are closest to the real values used to generate the data. The standard deviation is a fully-fledged parameter, and can be refined along with the mean in minimization (optimization) algorithms. If the errors are low, optimizing the parameters contributing to the standard deviation will make little difference to the outcome of MR. However, successful MLMR in borderline cases is not simply about good estimates of structure factors; it is also about good estimates of the *errors in the structure factors* (Figure 1). When the errors are high it is important to understand the sources of error so that they can be reduced and/or correct estimates optimally incorporated in the likelihood functions, so that the LLGI is maximized.

The contribution to the total LLGI from any individual reflection depends on the variables $E_{calc}$, $\sigma_A$, $I_{obs}$ and $\sigma_{Iobs}$ [20]. The total LLGI is the sum of the reflection LLGI values. These principles are the basis for the discussion of optimization of the signal in MLMR (Figure 2).

## 4 Methods

MLMR targets and target specific search strategies for MLMR are implemented in Phaser [15] (and previously BEAST [21]). Phaser is distributed through the CCP4 [23] and phenix [24] software suites under license. The software can be run from the command line, from python scripts, or through the ccp4i [25] or phenix interfaces [26]. Phaser is used in MR pipelines including MrBump [27] from the CCP4 suite, MRage [28] from the phenix suite and the WS-MR SBGrid [29]. It is also the basis of the anisotropy server [30]. Phaser has been incorporated into the development of *ab initio* phasing via MR in Arcimboldo [31] and Ample [32]. Many of the methods discussed here are relevant to all versions of Phaser, but some require Phaser-2.7.12 and above.

### *4.1 Target Function*

Phaser's LLGI target is the log of the likelihood of the MLMR hypothesis minus the log of the likelihood of the null hypothesis, where the hypothesis is formulated in terms of intensities [20]. The MLMR hypothesis is the current orientation and placement (translation function) or just orientation (rotation function) of the search component, within the background of the orientation and placement of other components under consideration. The null hypothesis is the Wilson distribution [33] of intensities, arising from a random distribution of isotropically scattering atoms in the asymmetric unit.

### *4.2 Search Strategies*

Phaser implements automated search strategies for finding multiple components. In the default search strategy, data are corrected for anisotropy and translational non-crystallographic symmetry; rotation and translation functions run with automated selection of potentially correct orientations and translations; packing checks performed; the partial solutions rigid body refined; and these steps iterated over the number of components. The resolution is optimized for speed. Steps are automatically repeated with altered parameters if the first set of parameters fail to yield a solution. The default search strategy is likely to find a solution if at all possible, but is also highly configurable (see documentation for details [34])

### *4.2.1 Fast Fourier Transform*

Conceptually, the full MR search space is 6N dimensional, i.e. the three rotational dimensions plus the three translational dimensions multiplied by "N", the number of models to be placed. An exhaustive 6N dimensional search becomes infeasible even with N in the low single digits, a problem that has spurred sparse sampling approaches [35]. These include the standard divide-and-conquer method of splitting the rotation and translation into two 3D searches (a rotation function and a translation function), but also include genetic algorithms [11] and *Monte Carlo* methods [12]. An advantage of performing the search as separate 3D rotation and 3D translation functions is that suitable target functions can be calculated by Fast Fourier Transform (FFT) [36]. The drawback of full MLMR targets is that they cannot be calculated by FFT, but targets for the rotation and translation functions that are suitable for FFT can be derived using the insights gained from the full MLMR treatment [37,38].

### *4.2.2 Sequential Addition*

One of the greatest strengths of the MLMR targets is that whenever a component is placed, the variances representing the remaining uncertainty in explaining the structure factors are reduced, thus increasing the signal-to-noise of the search for the next component. Thus, the natural way to build an MR solution of multiple components using MLMR is by sequential addition (Note 3). Phaser's default search strategy is to run consecutive rotation and translation functions, iterating the two 3D searches over the number of models N, and using the structure factors calculated from the known (already placed) components to leverage the search for the next component.

### *4.2.3 Peak selection*

Since the Phaser search strategy is to iterate the rotation and translation function searches over the number of components, it is necessary to set the selection criteria for the rotation and translation function partial solutions so that the correct rotation(s)/translation(s) are included in the list of rotation(s)/translation(s) carried through to the next step or iteration. The process is to sort the rotation(s)/translation(s) at the rotation/translation step in LLGI order and to select the highest. Selection criteria rely on there being at least some signal at the partial solution stage, so that the correct rotation(s)/translation(s) will be sorted towards the top of the list. The ideal place to prune the list is just below the correct rotation(s)/translation(s), but of course this is not

known. Rather, (by default) solutions are selected if they have a LLGI that is over 75% of the difference between the top and the mean of the search. This has the advantage that if the signal is high, then only the single top (correct) solution will be carried through, but many will be carried through if the signal is low. Other selection criteria are possible (see documentation for details [34]).

## 5 Identifying MR solutions

Since MR is a search procedure, the correct solution is identified by the signal-to-noise of the correct placement. The correct placement is obvious when a single point in the 6N dimensional search has a high LLGI value that is clearly discriminated from all others. As previously discussed, this point is not normally found with a 6N dimensional search. Instead, search methods rely on intermediate steps systematically eliminating regions of search space as they home in on the correct placement. The correct placement is found as long as intermediate steps do not eliminate it from the search space along the way. It is not necessary for each step in the search procedure to have high signal-to-noise in and of itself.

If the signal in the rotation or translation function is low, the default peak selection criteria may be eliminating the correct orientation or position from the search space. Since the signal from the rotation function is generally lower than that of the translation function, an obvious first parameter to change is the number of rotation function peaks being carried through to the translation function. The default *FAST* search strategy in Phaser automatically reserves a second tranche of rotation function peaks to pass to the translation function if the first (upper) tranche fails to yield a placement with a translation function Z-score (TFZ; see documentation for details [34]) over 8. If the signal in the translation function is also low, it may also be necessary to change the number of translation function peaks being carried through at intermediate stages of the search (see documentation for details [34]).

### *5.1 Success*

The LLGI is a direct measure of the probability of a placement being correct. There is also a direct relationship between the absolute value of the LLGI and its discrimination from the noise;

the higher the LLGI for the correct placement, the higher its TFZ. LLGI values for a model of the whole asymmetric unit greater than 60 generally have a TFZ of 8 and almost definitely represent a true solution [39]. The LLGI can be lower (50; TFZ ~7) and still indicate the correct placement of the first molecule in polar space groups, where there is one less degree of freedom. Clear discrimination from the noise is an excellent secondary indicator that a solution is correct.

Many (different and wrong) placements with an LLGI over 60 (TFZ over 8) indicate some unexpected pathology in the data that breaks the assumptions of the likelihood hypothesis. Common pathologies include twinning (possibly complicated by pseudo-symmetry) and errors in the space group determination.

### *5.2 Failure*

Assuming that a MR solution exists using the models provided, if the correct placements of the components, as determined by superposition after structure solution, is not indicated by a LLGI clearly discriminated from the noise, then the MR with that set of components will never be conclusive, whatever search strategy is used. If many MR trials (Note 4) do not produce a peak in the LLGI, then the crystallographer is justified in considering MR to have failed. However, putative model structures that may be somewhat superimposed on the target after structure solution but whose placement is not indicated by a signal in the LLGI will have very high phase errors, so that, had the placement been identified prior to structure solution, taking the MR solution forward to refinement would be extremely problematic.

### *5.3 Enrichment*

If there are a small number of solutions with LLGI approaching 60 (TFZ ~8), it is likely that one of these represents the true placement. Approximately half of the solutions with an LLGI around 30 (TFZ ~5.5) are correct [39]. If the list of potential placements is small, then it is likely that the signal-to-noise of these possible solutions is also relatively high, and that the low likelihood is not due to pathology. The solution list is *enriched*, even though MR is not conclusive.

It may be possible to distinguish the correct MR solution in an enriched list by taking each potential solution through to refinement. This is the approach taken in Balbes [40]. Rosetta

software [41] incorporates a wide convergence radius refinement method using approaches from *ab initio* modelling, and a pipeline for examining an enriched solution list from Phaser is implemented in phenix.mr_rosetta [42].

Anomalous data (e.g. from S-SAD, Se-SAD or a heavy metal soak) may also help to find the correct MR solution in an enriched list, even if the anomalous substructure has not been determined. In the MRPM (MR parameter matrix) search, the putative MR solutions from an enriched list are used to phase an anomalous difference Fourier and the MR solutions scored with respect to the peak heights in the resulting map [43]. If an anomalous substructure has been determined independently of a MR solution, but the resulting phases do not yield interpretable electron density, then the phases may still be good enough to identify the correct MR placement, simply by calculating the phase correlation between the experimental and (putative) MR phases. If both experimental and MR phase information is available, then phase combination will help bootstrap structure solution (Note 5).

### *5.4 Persistence*

If MR is failing, the crystallographer will have many MR trials from which to draw additional information. If the correct solution is somewhere at the top of the LLGI list in a number of trials, then considering the results of many trials in totality can identify the correct solution by the *persistence* of a solution across trials. This process was first introduced to MR in the context of looking across multiple rotation functions for the correct orientation [44], and later for translations [45] using AMoRe. The identification of similar placements can be done in real space, by clustering rotations and translations, or, for translations, in reciprocal space, by looking for phase correlations. If done in real space, it is advisable to pre-align all homologous model structures so that high-scoring orientations and translations from different models can be compared easily.

### *5.5 R-value*

When all components of the asymmetric unit have been placed, it is usual to calculate other scores to test their validity, particularly the R-value ($\Sigma||F_{obs}|-|F_{calc}||)/\Sigma|F_{obs}|$). The theoretical R-value for a random distribution of atoms, *i.e.* a maximally incorrect solution, is 0.586 [46]. In

practise, wrong solutions have R-values a few points lower because the absolute scale of $F_{obs}$ is not known, and $|F_{obs}|$ is scaled to $|F_{calc}|$. However, the R-value need not be lower than 0.586 immediately after MR. A notion that the R-value should be low after MR is equivalent to a notion that the R-value should show a signal for MR, and therefore would make a good target function, an idea abandoned with the introduction of the correlation coefficient [47,48] even before the introduction of MLMR. The R-value takes no account of the errors in $F_{obs}$ and $F_{calc}$, and is only a useful indicator when the phase error is small. Most models, even when correctly placed by MR, will still have very considerable errors. However, the R-value does give an indication of how straightforward will be the progression into model building and refinement. At values less than 0.40, the R-value becomes a reliable indicator of good phases. MR solutions giving high R-values will require advanced techniques to refine. For a detailed discussion of refinement, see the chapter by XXX in this volume.

### *5.6 Termination*

If the composition of the asymmetric unit is uncertain, it can be difficult to know when all components have been placed and the MR search can be terminated. The termination problem is usually solved when the signal-to-noise for adding components, which should increase with each additional component added, suddenly disappears and/or there ceases to be space for additional components in the asymmetric unit. A necessary but not sufficient condition for an MR solution is that the components form a connected lattice. In the end though, the MR search is only definitively terminated after the structure has been refined and passed validation tests.

## 6 Models

The important criteria for MR searches have been the subject of rules-of-thumb about sequence identity between model and target, editing of the model, and required size of the model [18,49–52]. Only some of these have been systematically studied (e.g. editing of the model [53]). The properties of the LLGI clearly indicate the veracity or otherwise of these rules-of-thumb for MLMR. Contrary to these traditions [18,49,50], there aren't generally applicable cut-offs in sequence identity or root-mean-square deviation (rmsd) of a model to the target for successful MR. The sequence identity *per se* is irrelevant, except in that it allows homologues to be

identified and an initial estimate to be made of the rmsd. Exactly how low the rmsd between the model and the target needs to be for success depends on the other parameters, particularly the model completeness (fraction of the scattering) and the number of reflections. High rmsd can be compensated by high model completeness. Low completeness can be compensated by low rmsd between model and target, and a large number of reflections (Figure 2).

Of course, the rmsd of the model to the target cannot be predicted reliably before MR. Strategies described here are designed to minimize the rmsd prior to MR.

### *6.1 Model improvement*

Because the rmsd of the model to the target cannot be predicted reliably before MR, the best model of all the alternatives cannot be chosen reliably either. Having the best possible starting model will make subsequent refinement and rebuilding much easier, so it is well worth spending some time evaluating a variety of alternative models, especially in difficult cases. MR models can be derived from different template structures in the PDB, processed in different ways by pruning, remodelling, or collected into ensembles, with or without trimming to a conserved core structure. In testing many models, it can be very helpful to use MR pipeline software such as MRage [28], which compares and combines results from many models in parallel.

#### *6.1.1 Modelling*

Techniques developed for *ab initio* modelling of protein structures have come of age for improving structures for MR. Application of chemical force fields can improve the structure to the point where the rmsd is low enough to find the solution [54]. Modelling specifically for MR is implemented in Ample [32]. For a detailed discussion of *ab initio* methods, see the chapter by DiMaio in this volume.

#### *6.1.2 Normal Mode Analysis*

Conformational change in proteins is particularly problematic for MR. The crystallographer may expect conformational change - even be hoping to probe it - from previous studies of the macromolecule or macromolecular complex. Conformational change in proteins has been shown to be modelled by normal-mode analysis of the elastic network model [55–59]. One or more

normal modes may contribute to a given conformational change [58,59]. Perturbations along normal modes were first used successfully to find MR solutions with *AMoRe* [60]. Neither the normal modes that model the conformational change nor the perturbation distance along the modes are known in advance; multiple perturbed models need to be generated with different normal mode combinations and perturbation distances. By chance, one of the conformations generated may have a lower rmsd to the target structure than the original model, and hence yield a signal in MR. However, it is necessary to sample hundreds or even thousands of possible perturbations in order to sample conformational space finely enough to generate a good model. Normal mode perturbation of protein structures in rmsd increments can be performed with Phaser (see documentation for details [34]).

### *6.1.3 Conformational Sampling*

Some families of proteins have been intensively studied and are present in the Protein Data Bank in many different conformations. Kinases are a prominent class of protein for which the structure with the highest sequence identity in the PDB is not likely to be the one with the lowest rmsd to the target. Kinases undergo a conformational change upon NTP binding, but the changes are not well-modelled as a simple change in disposition of domains [61]. There are thousands of kinase structures in the PDB (including serine/threonine, tyrosine, histidine, receptor and non-receptor types), and these represent many different kinase conformations. Although not all are unique, the conformational sampling they represent can be used to solve MR problems by trying all kinase structures regardless of sequence identity.

### *6.1.4 Wide search*

*Wide Search MR* (WS-MR) [29] is the extension of the database of search models to the entire PDB. The CPU intensive search becomes tractable through the use of national supercomputer grids. The approach allows optimization of the MR search model by brute force: it does not rely on sequence identity to identify models. As a consequence, MR solutions can be found with very low sequence identity and/or sequence coverage. As implemented through the SBGrid [62], the LLG and TFZ scores from Phaser are initially used to filter possible solutions, and then the structures that generate these solutions clustered by fold to find folds that persist in the solution list (Note 6).

### *6.1.5 Ensembling*

A reduction in model errors can be achieved using an ensemble of superimposed structures that are similar. The result of the *ensembling* [21] procedure is a set of $E_{calc}$, which are used *in lieu* of structure factors from a single model. The errors in the ensemble $E_{calc}$ are lower than those of each model individually. The assumption is that some parts of the structures will be systematically closer to the target than others. The scattering from these sections will be reinforced, while regions that differ will be down-weighted. If all the structures were weighted identically, ensemble $E_{calc}$ would be equivalent to summing structure factors from the components and dividing by the number of components, or equivalently, taking the N superimposed structures, each with the fractional occupancy $^1/_N$. A more sophisticated approach is to weight the structures according to the expected rmsd to the target structure. An ensemble of structures has been shown to be particularly effective when there are a number of low sequence identity models available for the target structure [21,63].

### *6.1.6 Bulk solvent*

Ordered atoms are only part of the scattering matter present in the crystal. Also present are disordered atoms in the bulk solvent. Solvent corrections to the structure factors were originally developed for refinement [8,64] where they clearly improve map quality. A mask-based solvent correction has been successfully applied to fast translation searches in MR [65] with *AMoRe* [47] and with *CNS* [8]. Phaser has the option of applying a mask bulk solvent correction throughout MR. The model structure factors are calculated for structure factor interpolation, by placing the model in a large *P*1 cell with less contrast to the surrounding solvent as compared to the default calculation, which places the model *in vacuo*. The mask-based bulk solvent correction reduces the error in the calculated structure factors at low resolution and hence increases the $\sigma_A$ at low resolution. Although not applied by default, the bulk solvent correction in Phaser can rescue failed MR in some cases, particularly those with data to only low resolution (see documentation for details [34]).

### *6.2 Model Completeness*

There is a penalty to the LLGI associated with reducing the size of the model. However, reducing the size of a model can be advantageous if the atoms in the wrong (relative) positions can be removed prior to the search.

#### *6.2.1 Pruning*

The longest standing method for removing atoms in the wrong (relative) positions is pruning the amino acid side chains of the model. Amino acids that are not conserved between model and target should be trimmed back to a common core. In the simplest analysis, this is the $C_\beta$ atom (polyalanine), but more complex analysis can add one or two atoms further along the amino acid side chain where there are conserved atoms between common rotamers of spatially equivalent model and target amino acids. Pruning of the model has been shown to be decisive in the solution of MR problems [53]. This can be particularly powerful when using ensemble models, as the loops that differ among members of the ensemble can be trimmed to leave just the conserved core. Pruning can be performed with CHAINSAW [66] from the CCP4 suite or *phenix.sculptor* [67] from the phenix suite.

#### *6.2.2 Domain analysis*

Protein domains are variously defined, for example in terms of sequence motifs, functional elements or evolutionary modules. For the purposes of MR, a domain is a structural element within which the atoms are fixed relative to one another between model and target, and hence are suitable sub-structures for MR. There are often changes in the disposition of domains in multi-domain proteins sometimes related to function, but also simply due to flexibility and crystal packing forces. When the protein (or a homologue) has been solved in two or more conformers, the structurally invariant regions can easily be identified [68–71]. It is more challenging to identify domains for MR when the structure of only one conformer has been solved. In simple cases, visual inspection may be sufficient to identify potential rigid domains. Various automated approaches have been taken including considerations of surface area [72] , molecular dynamics simulations [69], TLS group analysis [73], and normal-mode analysis [74], amongst others. Phaser implements the SCEDS procedure [71].

If a model is split into N domains, the search for that component becomes 6N dimensional to allow each set of atoms with correct (relative) positions to be optimally positioned. The signal for the correct placement of all the domains may not be discriminated from noise until the final component is placed. In difficult cases, it is therefore advisable to search for all components in one run of Phaser, allowing the software to build a complete solution component by component, and optimizing the signal from each component as it progresses.

### *6.2.3 Oligomers*

If the protein or proteins in the crystal are known to form oligomers, either hetero-oligomers or homo-oligomers, then searching with models with the target's oligomeric arrangement will increase the signal. Dimers, trimers, tetramers and hexamers with point group symmetry are able to crystallize with one unit (which may itself be made up of a protein assembly) in the asymmetric unit of the crystal in space groups with the same two-, three-, four- or six-fold point group symmetry. Fibres, which are infinite chains of proteins with screw symmetry, must crystallize so that the crystal screw symmetry generates the infinite chain. Searching with an oligomer with more scattering matter than present in one asymmetric unit, where the oligomer is placed on a special position with respect to the crystal symmetry, is supported in Phaser.

### *6.2.4 Brute searches*

In difficult cases, the full MLMR targets can be calculated point by point on rotational and translational grids [21], rather than using the likelihood enhanced fast rotation/translation functions and FFT [37,38]. This is termed a "brute" search. Since the full likelihood functions are slow to compute, brute searches are most useful when the search space can be restricted to a particular set of angles/coordinates near a particular placement. Such a scenario occurs when searching for a multi-domain, flexible protein for which a model of the entire target exists. It is often possible to place the large domain(s) but not the small domain(s). The approximate placement of the small domain(s) can be inferred from the placement of the large domain(s). Performing a brute rotation/translation searches restricting the orientation/position to angles/coordinates within a few tens of degrees/Ångstroms of the position relative to the (large) placed domain(s) often finds the correct placement of small domain(s) with high signal-to-noise, using the power of MLMR. In practice, it is usually sufficient to carry out a brute-force limited

search of orientations combined with a fast translation search over the entire volume, because the signal is much stronger for the translation search than the rotation search. Obtaining a solution consistent with connectivity between the domains increases confidence in the correctness of that solution. The brute search method can be thought of as a wide-convergence-radius rigid-body minimization.

### *6.2.5 Fragments*

If the number of reflections is high then it becomes feasible to use very small but accurate (low rmsd) search fragments for MR. Elements of secondary structure can prove useful generic models. Helices are particularly suitable as they are very regular over lengths of several turns; beta sheets have twists that distort the disposition of atoms within a short stretch of amino acids. This approach is particularly effective in solving coiled-coil structures [75], where MR with Phaser often fails to dock the sequence (amino acid or base) onto the helix, probably due to the strong helical modulations of the diffraction pattern. That small accurate fragments can be used to solve MR problems when whole accurate models are not available is the basis for Ample [32], Arcimboldo [31], Arcimboldo-Borges [76] and Arcimboldo-Shredder [77].

### *6.2.6 Search B-factor*

Model components differ not only in the rmsd to the target, and model completeness, but also the relative B-factor. The components with low B-factors are generally found first in any search. The high B-factor components can be very hard to place, because these contribute less to scattering at high resolution than other components. The relative B-factors of components are not known before structure solution, but if later components in a search are proving difficult to locate, high B-factors should be suspected. This is particularly likely if one copy of a component has been found, and therefore shown to be a good model. In Phaser, the average B-factor of all ensembles (and members of an ensemble) is, in effect, set to the Wilson B-factor. Thus, average B-factor differences do not affect the ensemble structure factors, but Phaser has the option to explicitly add a relative B-factor to the search for a component to down-weight the structure factors at high resolution appropriately, and hence increase signal (see documentation for details [34]).

### *6.3 Model errors*

Model errors are important parameters in the likelihood targets. Correct estimation will improve signal-to-noise in borderline cases. The model error, $\sigma_A$, is computed from the estimated rmsd of the coordinates between model and target and the fraction scattering that it represents.

The LLGI for the placement of a component should be positive, and should increase as components are added. If it is negative or decreasing, it means that the parameters of the likelihood function are predicting the data worse than would a collection of random atoms. The errors are underestimated, too optimistic about how well the model can predict the data: the completeness is being over estimated and/or the rmsd of the coordinates is being underestimated.

#### *6.3.1 Sequence Identity*

Although not known exactly until after structure solution, the rmsd can be estimated from the sequence identity [78] or more accurately by also taking into account the size of the protein [39]. Optimization of the estimated rmsd can be the difference between success and failure in MR trials with low signal [39].

#### *6.3.2 Composition*

The fraction scattering of a given ensemble is calculated in Phaser from the atomic composition of the input ensemble and the total atomic composition of the asymmetric unit, usually entered as protein and/or nucleic acid sequence and number of copies. The asymmetric unit composition is thus an important parameter in MLMR. Increasing the composition of the asymmetric unit will decrease the fraction of the scattering accounted for by each component.

If the composition of the asymmetric unit is uncertain, then so too will be the fraction scattering of each component. If the asymmetric unit is assumed to have less scattering than actually present, then $\sigma_A$ will be over estimated, and vice versa. The LLGI will be optimized when the composition is correct. In difficult cases, it will be necessary to perform MR not only altering the number of search components but also the composition.

### *6.3.3 Conformational Change*

When modelling conformational change, the rmsd used to estimate the $\sigma_A$ should be close to the rmsd expected to apply after successful structure solution, not the higher value expected between model and target before modelling the conformational change. If conformational change is being modelled by normal mode perturbations, then the rmsd between perturbations will give an estimation of the upper limit for rmsd of the best model to the target. Phaser generates normal-mode perturbed structures by rmsd increments for this purpose (see documentation for details [34]).

### *6.3.4 Atomic B-factors*

Although the *overall* scale of the B-factors of the model coordinates does not affect MR with Phaser (see section 6.2.6), differences in B-factors *between* atoms in a model affect the relative contribution of the scattering of each atom to the calculated structure factors at different resolutions; scattering from regions of high B-factor are down-weighted at high resolution. The atomic B-factors should be set proportional to the expected positional error squared. Modelling expected coordinate errors along the polypeptide chain as B-factors, usually lowest in the core and highest on the protein surface, have been shown to dramatically improve the utility of homology models for MR [79].

### *6.3.5 VRMS Refinement*

Phaser refines the coordinate errors (VRMS [39]) for each component in conjunction with the rotation and translation of the model. The VRMS will often refine to a lower value than the input rmsd for a correct solution. If VRMS values of a solution refine to a significantly different value than input, then repeating the search with the refined VRMS input from the start should increase the signal-to-noise of the rotation and translation functions.

## *6.4 Model Case Study: Antibodies*

The approaches to optimizing a model are well illustrated by the long-standing MR problem of how to solve Fab antibody structures [80], with or without their protein antigens. The elbow angle, the angle between the variable (Fv) and constant (Fc) domains of the antibody, is highly

variable [81]. If the data are high enough resolution (*i.e.* there are a large number of reflections) then Fab placement will be possible by splitting the Fab into Fv and Fc domains and searching for these consecutively, even using Fabs with low sequence identity to the target. Pruning the Fv and Fc to the core conserved with the target is always advisable. Because of the flexibility at the elbow angle, the B-factors of one of the domains may be high, causing problems for the MR. If the Fv domain is well ordered (due to binding to its well-ordered protein antigen), and hence is easily located by MR, then a partly disordered Fc may be found by increasing the B-factor in the search for Fc or by local brute search. If the data are not so numerous, then a good signal will only be obtained searching with the whole Fab and with the elbow angle of the Fab correctly modelled. The only correlation between elbow angle and sequence is via the subtype of light chain (κ or λ) [81]. The correct antibody hinge angle may be found amongst those Fab structures already in the PDB, or novel conformations may need to be generated with normal mode perturbations. If the data are even poorer, then the signal can be further improved by modelling the Fv. Modelling approaches for Fv domains regularly achieve an rmsd of 1Å or better [82]. Searches may be necessary using a range of rmsd values, or an ensemble of Fv models may be useful. Placing the Fv and Fc domains correctly in the asymmetric unit can bootstrap the placement by MR of the protein antigen, or indeed phasing by other methods. This is a secondary benefit of the use of Fabs as chaperones to aid the structural determination of otherwise 'uncrystallizable' target proteins, a method that has become relatively standard [83].

## 7 Data

Guidance about good data collection strategies becomes particularly relevant in difficult MR cases. For a detailed discussion of data collection strategies, see the chapter by Dauter in this volume. Some problems arising from fundamentally bad data collection simply cannot be resolved by data processing and will be fatal to MR. The following discussion assumes that the data are correctly indexed, are free of overlaps and overloads, and that the $\sigma_{Iobs}$ associated with an $I_{obs}$ encapsulates the measurement error reasonably accurately.

Like model preparation, data preparation for MR has also been the subject of rules-of-thumb regarding the resolution of the data, the need for completeness of the low-resolution data, and so

forth [18,49–52]. Again, the properties of the LLGI clearly indicate the veracity or otherwise of these rules-of-thumb for MLMR. If the data have no pathology, then, for a particular model, the LLGI depends only on the number of reflections, not the resolution of the data, or the completeness of the data in resolution shells (Figure 2). This runs contrary to experiences with Patterson methods, where the completeness of the low-resolution data is critical to the success of MR [52,84] and where high-resolution data are not essential [85].

### *7.1 High-resolution data and high rmsd*

Although the number of reflections is a key factor in determining the LLGI, reflections with a resolution higher than 1.8 times the rmsd of the model have $\sigma_A$ values so small that they contribute insignificantly to the total LLGI. Estimates of the rmsd for MR show that for sequence identities of 15%, the rmsd is estimated as 1.5Å for small models and up to 2.5Å for large (1500 residue) models [39], which implies that data better than 2.7Å for small models, and 4.5Å for large models, will only increase CPU time. However, in cases where MR is not expected to succeed based on the most likely rmsd, success will only be found for models that happen to be somewhat better than expected, so it can help to run trials with optimistic values for the rmsd. An rmsd of only one standard deviation below the expected value (0.2 times the expected value [39]) increases the useful resolution by nearly 40%. If the VRMS is lowered in refinement, it will benefit from the additional data. Deliberately truncating data, for example at 3.5Å, can lose critical signal for marginal cases. Phaser sets the resolution limit optimally for the rmsd input, and changes the resolution limit during the course of MR depending on how much signal is present.

### *7.2 High-resolution data and low rmsd*

Using the same argument as in section 7.1, MR with small accurate fragments will benefit greatly from high-resolution data. On no account should resolution be truncated in the search for helices or other small structural motifs.

### *7.3 Measurement error*

At the diffraction limit of the crystal, the issue becomes measurement error. Since the demonstration that useful information can be extracted from very weak diffraction data in refinement [86,87], and the introduction of pixel counting detectors, data are now frequently integrated beyond traditional resolution limits (e.g. merged $I_{obs}/\sigma_{Iobso} > 2$ in the outer shell). The LLGI will down weight the contribution for the poorly measured reflections at the diffraction limit of the crystal. Using LLGI, adding data at the high-resolution limit with high experimental error will not bias the MLMR target in the way that amplitude-based likelihood targets do, and at the same time allow all well measured reflections, regardless of overall the $I_{obs}/\sigma_{Iobs}$ in their resolution shell, to contribute to structure solution. With the use of LLGI, it should not be necessary to vary the high-resolution limit for MR in an attempt to get a solution, unless there is some pathology in the data at high resolution (*e.g.* an ice ring near the resolution limit). However, in the extreme of integrating data well beyond any reasonable diffraction limit, e.g. 2.0Å ($I_{obs}/\sigma_{Iobs} = 2$) data integrated to 1.0Å, the integration and scaling programs may do a poorer job of estimating the intensities and standard deviations, and some degree of restraint should be exercised.

### *7.4 Low resolution data*

If MR is failing and the data are poor, then improving the data should be a priority. The higher the resolution of the data, the more options for attempting MR with smaller, more accurate fragments. Low-resolution data below 15Å is disproportionately affected by the poorly modelled diffraction from the solvent, and so has lower $\sigma_A$ values than do data around 6Å. Mid-resolution data thus give more signal per reflection than do low resolution data. Unlike Patterson based MR, high completeness at low resolution is not particularly valuable for MLMR.

### *7.5 Completeness*

Because the LLGI is dependent on the number of reflections, it is obvious that collecting complete data will maximize the number of reflections to the diffraction limit of the crystal. Missing data affects the map resolution: the electron density is convoluted with the Fourier

Transform of the mask of the missing data. Randomly missing data lower the effective resolution of the map isotropically. If data are systematically missing in a wedge, then the effective resolution in the plane perpendicular to the wedge will be lower. MR orientation and position parameters will be less accurate in the direction where the effective resolution is lowest.

### *7.6 Intensities*

Structure factor amplitudes are normally generated from intensities by the French and Wilson [22] *truncate* procedure. In some structure solution pipelines, data are subjected to the truncate procedure by default, and all subsequent steps are performed with these amplitudes, however this transformation introduces serious biases in the likelihood targets. The LLGI targets abrogate the need for any transformation to amplitudes during MR, and it is important to input the data to Phaser in terms of intensities rather than amplitudes [20].

### *7.7 Alternative Datasets*

If data are generally poor, it is advisable to forward a number of differently processed datasets of merged intensities for MR trials. This is a good strategy in the presence of radiation damage, where it is often not clear where to cut the data with dose to balance merging R-values against multiplicity and completeness. Differently processed or merged data sets can be used to test the *persistence* of a solution (see section 5.4).

### *7.8 Space Group*

Patterson based likelihood targets are less effective for higher symmetry space groups, due to the presence of inter-molecular vectors in the Patterson calculated from the data. As the symmetry increases, more and more inter-molecular vectors crowd the observed Patterson, and the signal is reduced. For MLMR, higher symmetry also increases difficulty in structure solution, because greater uncertainty in adding up structure factor contributions from symmetry-related molecules with unknown relative phase increases the variance of the rotation likelihood target. The space group has no equivalent effect on the difficulty of the translation step in MLMR.

### *7.9 Alternative Space Groups*

Enantiomorphic space groups cannot be distinguished in the data processing stage, only by structure solution. Space groups that only differ by screw symmetry can be distinguished by the presence of systematic absences, but if the axial data are weak or missing, the assignment of screw axes is not certain, and again, the correct space group can only be distinguished by structure solution. Clear identification of space group amongst a list of alternatives is a good secondary indicator of the validity of a solution.

### *7.10 Aniosotropy*

There are often differences in long-range order in different directions in reciprocal space. MLMR relies on comparing structure factors computed from a model isotropically scattering atoms with the observed data. If the implicit assumption of isotropic scattering is wrong, MLMR will not score the placements correctly and structure solution will fail. The anisotropy parameters are refined by fitting the structure factor intensity to the Wilson distribution to correct the data for anisotropy and allowing structure solution as for isotropic data. The anisotropy correction is applied to both the data and the experimental errors in the data. The anisotropic correction factors calculated by fitting the data to the Wilson distribution will not be as good as those that can be calculated once the atomic model is known. Anisotropically corrected structure factors used for MR should not be passed to refinement programs.

## 8 General Non-crystallographic Symmetry

There is nothing particularly special about the presence of general non-crystallographic symmetry in determining the solvability of the problem by MR, as compared to any other problem with multiple components in the asymmetric unit.

The Matthews coefficient [88], originally established from a study of protein content in protein crystals, has been reinvestigated for crystals of nucleic acid-protein complexes and nucleic acid alone [89,90]. The most likely number of macromolecules in the asymmetric unit is the number that gives the most likely solvent content. When the most likely number of macromolecules is

one or two, it is well determined, but as the number of macromolecules increases so too does the uncertainty.

Clues to the crystal composition can be gleaned from sources other than the crystal data in hand. The number of copies in the asymmetric unit may be informed by the oligomeric state of the complex in solution, combined with the presence or absence of pure rotational symmetry operators in the space group. Light scattering, native gel electrophoresis, ultracentrifugation, and electron microscopy can indicate oligomeric state. However, differences between the buffers in which these experiments are performed (such as salt and pH), physical forces, and possible proteolysis, mean that these experiments are not necessarily good indicators of the oligomeric state in the crystal.

Information about the NCS can also be gleaned from the self-rotation function (SRF [91]). The SRF is most intuitively specified with three polar angles: the azimuthal angle $\Phi$ and the zenith angle $\Psi$, which specify the direction of the rotation axis; and $\kappa$, the rotation about this axis. When there are multiple copies in the asymmetric unit, the SRF is complicated and generally not interpretable, unless there is rotational symmetry. The $\kappa$ section of the peak in the SRF shows the rotation order=360/$\kappa$, *e.g.* two folds will appear as peaks on the 180° $\kappa$ section. If rotational symmetry of a given order is clearly present, then the number of copies in the asymmetric unit is likely to be a multiple of the rotation order.

If the number of copies of the macromolecule in the asymmetric unit is not well determined, a lack of certainty becomes a significant problem for MR through not knowing when to terminate the search, and not knowing the fraction scattering of the components as the search is progressing.

Although the presence of NCS can increase the difficulty of MR, it has the compensating advantage of enabling NCS averaging after MR, which will remove some of the model bias. This is especially valuable in low-resolution structure determinations.

## 9 Translational Non-crystallographic Symmetry

Translational non-crystallographic symmetry (tNCS) arises when two or more copies of a macromolecule or macromolecular complex are present in the asymmetric unit in the same orientation. The presence of tNCS modulates the diffraction pattern in a way that is problematic for likelihood functions, because, like anisotropy, it violates the implicit assumption behind likelihood targets that the data follow an isotropic Wilson distribution. Macromolecules related by tNCS will have an associated peak in the native Patterson. The magnitude of the Patterson peak is a measure of both how exactly the translation vector models the translation between all atoms in copies of the macromolecule and the strength of the resulting diffraction modulation. Peaks in the native Patterson more than 20% of the origin peak are a good indicator of macromolecules being present in approximately the same orientation (up to 10° rotation for an average size protein), and for the modulation being a significant hindrance to the likelihood targets.

An important aspect of accounting for tNCS with likelihood is the modelling of the errors. The tNCS is characterized not only with a vector, but also with parameters describing the deviation from simple translations of identical coordinates between the tNCS copies. The naive, non-likelihood approach, of modelling the tNCS as a simple translation of one structure by the tNCS vector, is inadequate for structure solution in the majority of crystallographic problems with tNCS. The likelihood correction to the tNCS is performed by refining expected intensity factors for each reflection, derived from the tNCS model of tNCS vector(s) and errors. The expected intensity factors are then used in the likelihood functions as usual, and in many cases structure solution becomes straightforward [92].

When tNCS is present and can be characterized and the intensity modulations accounted for, it can be considered an advantage for MR, because there are fewer independent copies in the asymmetric unit to place versus the same asymmetric unit contents without tNCS. On the other hand, tNCS reduces the power of NCS averaging to improve phase quality [93].

### *9.1 tNCS Order*

Frequently, tNCS associates *NMOL* macromolecules in the asymmetric unit in a series of vectors that are multiples of 1, 2, 3... (*NMOL*–1) times a basic translation vector (*TVEC*), with *NMOL*× *TVEC* being a unit cell translation, possibly along a unit cell diagonal. In this case the tNCS represents a pseudo-cell and is known as commensurate modulation. The integer *NMOL* is the order of the tNCS. Trying to find the related set of vectors by inspection is complicated by the Patterson symmetry and cell translations. The series will not generally have all peaks the same height. Lower peaks in the vector series represent relative rotations between vector-related molecules that are larger, and may even be missed by the default 20% origin cut-off. Phaser finds the order of tNCS and the translation vector in cases of commensurate modulation by Fourier analysis of the Patterson.

### *9.2 Pairs of molecules*

If there is a single peak in the native Patterson, it represents macromolecules clustered into two groups (*NMOL*=2) related by a single tNCS vector. When accounting for pairs of macromolecules, Phaser is not restricted to pseudo-cells/commensurate modulation. The tNCS vector can be in a general position. In these cases, Phaser can refine not only an rmsd between tNCS related copies but also a specific relative orientation between the macromolecules in the two groups. Starting from the Patterson translation vector, an estimate for the rmsd between copies, and a small number of initial rotational perturbations, the parameters are refined against the Wilson distribution to optimize the expected intensity factors for use in the LLGI.

### *9.3 Complex tNCS*

If there are many macromolecules in the asymmetric unit but they are not all related by tNCS, or there are sub-groups of macromolecules related by different tNCS vectors, then the modulations of the expected intensities due to the tNCS will be much less significant than for commensurate modulation or for pairs of macromolecules. In these cases it is possible that structure solution will be achieved without any tNCS correction factors being applied. Indeed, searching exclusively for tNCS-related multiples when some molecules are not related by tNCS will cause structure solution to fail.

If ignoring tNCS fails to give a solution, then the solution must be approached step-wise. Consider the highest native Patterson peak first, apply the associated epsilon correction factors, and locate all the molecules with this tNCS. Fix these components, and then take the second independent native Patterson peak, apply the correction factors associated with it, and locate the second set of molecules. Finally, turn tNCS correction off to find any orphan molecules.

### *9.4 Helices*

Crystals of nucleic acid, particularly DNA duplexes, and from α-helical coiled-coils, can show clear helical modulation of the diffraction pattern, and have correspondingly large Patterson peaks due to the helical repeats. The direction of the helices can be inferred from the large Patterson peaks alone. If helical features generate Patterson peaks above the 20% threshold, it will be necessary to turn off the automatic tNCS correction in Phaser.

## 10 Twinning

In general, MR works well with twinned data. The errors in the calculated structure factors need to be only slightly lower than would be needed for untwinned data from the same crystal form. Twinning may not even be suspected [94]. For a more detailed discussion of twinning, see the chapter by Thompson in this volume.

### *10.1 Merohedral*

With hemihedrally twinned data, Phaser should produce two sets of solutions that are equivalent under the twin operator, although they may not be on the same origin. However, if the twin fraction α is even slightly less than 0.5, Phaser may only give one solution. For more than two twin domains Phaser may (or may not) produce more than one solution, related by the twin law(s). To test for the twinning with a particular twin operator, twin related solution(s) can be generated manually and the LLGI calculated to compare with the original solution.

Twinning is detected with a range of tests [95]. Twinning tests that rely on structure factor intensity statistics work poorly in the presence of anisotropy and tNCS, but if the anisotropic and tNCS intensity modulations are corrected as described above, these tests become reliable [96].

Phaser reports p-values that will suggest whether twinning is present after removing the systematic intensity modulation effects [92].

The main problem with merohedral twinning in the context of MR occurs when perfect twinning causes the space group to be misidentified and the data are merged in a higher symmetry than the true symmetry. MR will then either fail outright, give a partial solution, or the R-value of what appears to be a full solution may stall during refinement, with the electron density showing breaks and spurious features that cannot be corrected by model building.

It can be difficult to detect perfect twinning masquerading as crystallographic symmetry, unless the asymmetric unit volume is too small to contain even a single copy of the macromolecule. If the data are merged in too high symmetry, the twinning tests that depend on twin laws, which compare reflections that are equivalent according to a possible twin law, cannot be performed. Only the tests for twinning that consider intensity statistics, such as the moment test in Phaser, will still indicate that twinning is present.

If twinning is indicated by the intensity statistics, and MR/refinement fails, then the true symmetry is probably lower. However, any or all of the symmetry operators could correspond with the twin operator(s). Phaser reports all the subgroups of the current space group, any of which could be the true space group in the presence of twinning. Especially in higher symmetry space groups, the number of subgroups can be very considerable, as screw symmetries also need to be considered. These can be systematically investigated by merging the data in all the lower symmetry point groups. However, if the twinning is perfect, the data can simply be expanded to lower symmetry without it being necessary to re-merge the data. MR pipelines can run numerous jobs simultaneously [28].

If the MR model is good, then solving the structure in *P*1 and using the symmetry of the resulting structure to determine the true space group can bypass the expansion to all subgroups. The calculated structure factors from the MR model in *P*1 are tested to see if they obey higher symmetry [97–99].

### *10.2 Reticular Merohedral*

In reticular merohedry, the reciprocal lattices of the twin domains superimpose exactly, but for only a fraction of the reflections. A characteristic warning sign is a pattern of "bizarre" apparent systematic absences, which are not consistent with any space group [100,101]. The problem can be one of unit cell and/or space group determination because overlapping lattices may be interpreted as a single lattice. Overlapping lattices can be interpreted as a large unit cell, or make a centred space group appear primitive. Indexing twin-related lattices as one will cause MR to fail. If the strongest twin component can be indexed and integrated independently of the others, and enough of these reflections are unaffected by twinning, MR should be possible using the unaffected reflections alone. Data may be augmented by adding the intensities of the common reflections divided by the number of twin contributors [100].

### *10.3 Pseudo-Merohedral*

Pseudo-merohedrally twinned data are equivalent to merohedral twins for the purpose of MR. The difficulty with pseudo-merohedral twins is in the data integration step. If the difference in unit cell dimensions is very small, and the reflections are very close to one another on the detector, then the aim of integration is to mask the twinned reflections into a single reflection so that reflections of the same index are integrated as one, so as to, in effect, force the data to be merohedrally twinned.

### *10.4 Static Disorder*

Twinning is just one of the crystal pathologies of crystal disorder. On the other end of the continuum is statistical disorder, where the mosaic blocks are small compared to the coherence length of the X-rays. MR with statistical disorder is likely to produce several solutions with high signal-to-noise with severe packing clashes. Refinement will involve setting the occupancy of overlapping components in the asymmetric unit to appropriate values.

## 11 Packing

Explicit checks for the presence of overlap amongst and between the crystallographically and non-crystallographically related components in the unit cell are powerful additional criteria for the selection of MR solutions. The problem with these overlap tests, also known as packing tests, is that any errors in the MR model will mean that the model will not fill the same molecular volume as the true structure it represents, and so there are errors in the packing tests that cannot be properly accounted for.

A measure of the packing is given by the FFT-calculated overlap function [102], which quantitates the total volume of the unit cell that is occupied. This is a continuous function, and has been used to weight the translation function score in proportion to the total volume occupied, with the effect of (potentially) reordering the translation function peaks in MOLREP [103] and AMoRe [104]. The overlap function becomes less useful as the number of components in the asymmetric unit increases. If there is only one component in the asymmetric unit, then any reduction in the total volume occupied can be fully attributed to overlap between crystallographically related copies of that component. If there are many components, then the reduction in the total volume occupied may be entirely due to overlap of one component, or some overlap of them all. The latter should be accommodated, but the former should not. The two cases can be distinguished by counting atomic (or atomically representative) contacts, and this is the basis of the packing analysis in Phaser. Solutions are excluded if the pairwise overlap between two components is more than a given percentage. The Phaser packing test is therefore pass/fail, rather than a continuous function, and does not reorder the translation function peaks.

### *11.1 Trace of Coordinates*

It is prohibitively slow to include all atoms in the analysis unless the model has less than about 1000 atoms. Instead, "trace" atoms represent the volume of the components. These can be Cα atoms for protein and a selection of phosphate backbone and base atoms in nucleic acid. Alternatively, the trace atoms can be abstracted to a set of points filling the molecular volume, for example to points on a hexagonal grid within the van der Waals volume of the protein. The default trace used to represent a set of coordinates adjusts to the size of the macromolecule so

that the volume is represented by a maximum of 1000 trace points (see documentation for details [34]).

### *11.2 Trace of Maps*

Electron density maps can be used to define a model for MR in Phaser, using similar input to that for a coordinate-based definition of ensembles. Putative solutions from electron density maps are tested for packing in Phaser by filling the Wang volume [105] with a hexagonal grid of points and proceeding as for the packing of atomic models.

### *11.3 Explicit Trace*

By default, the trace of an ensemble used for packing is derived from the ensemble coordinates or Wang volume. However, it is possible to input the trace to be used so that it is defined independently of the coordinates or electron density input for calculating structure factors for the likelihood targets. This can be useful if searching with small fragments, where it is possible to exclude a larger volume around the search fragment in the packing tests points (see documentation for details [34]).

### *11.4 High TFZ solutions*

Solutions that have high LLGI, indicating that a placement is correct, but which fail the packing test, need to be investigated more closely. A second copy of a component may be placed on top of an identical, previously placed component if the component has a B-factor significantly lower than the Wilson B-factor: the second copy attempts to model the missing scattering. Significant overlap may also be caused by the presence of static disorder. Solutions with minor overlaps may be excluded because the allowed percentage for overlap is too strict given the accuracy of the model. Although the solution may be accepted by being more accommodating of overlap, ideally the model should be edited to remove atoms that are not shared between the model and the target, which will also increase the LLGI. Solutions with high TFZ scores that do not pack are saved to a separate output solution file for subsequent analysis (see documentation for details [34]) or undergo likelihood-guided pruning (see section 11.5).

### *11.5 Likelihood-guided pruning*

The fundamental problem of the packing test, that of ignoring packing clashes between atoms in the model that are outside the true molecular envelope, can be addressed with likelihood guided pruning. This method identifies atoms in the components placed in the asymmetric unit that do not contribute to the LLGI, and hence are not present in the structure. Blocks of atoms, in groups large enough to cause a significant change in the LLGI, are removed in sections along the polypeptide chain and the LLGI calculated. If the LLGI goes up by a statistically significant amount as a result, then the corresponding block of atoms is removed from the packing test. This can rescue solutions with high likelihood that fail the initial packing test.

### *11.6 Packing during translation function*

A high LLGI solution that does not pack influences the results of the translation function if it is the top-most peak from the translation function, since, (in the default selection criteria), the top peak is taken as the reference for the cut-off LLGI value for acceptance. If the LLGI of this top peak is much higher than any others, then it may be the case that no other solutions are output from the translation function, causing structure solution to fail in the subsequent packing test due to the loss of other candidate solutions. To avoid this case, a packing test is performed on the top solution *during* the translation function and the top peak is discarded if the overlap of any component is more than 50% of the volume. Alternative placements due to static disorder will likely be lost in this process.

## 12 Electron microscopy maps

Improvements in detectors and reconstruction software now allow atomic resolution electron microscopy (EM) imaging [106]. With high-resolution images from electron microscopy now available, it is possible to bring X-ray crystallography and electron microscopy together in two ways. Structures solved by X-ray crystallography can be docked into the high resolution EM maps, in a process analogous in many ways to MR, but this is not the subject of this review. Secondly, the electron microscopy images can be used as models for MR. This is possible even if the electron microscopy imaging has not (yet) yielded an atomic resolution structure. Since the

model used in the likelihood targets is represented by the calculated structure factors, it is trivial to replace the structure factors calculated from a model with the observed structure factors from EM. The likelihood functions are then deployed without modification.

An important additional consideration when using EM maps as models in MR is that the scale of the electron micrograph may be miscalibrated by several percent [107]. Miscalibration will at the very least add noise to the MR search, and will often prevent structure solution. The MR search should be done with the scale of the EM map varying +/- 10%.

The resolution of the search using EM as a model is restricted by the resolution of the EM map. Phase extension utilizes NCS averaging (if present) or other density modification processes. It may be necessary to resort to experimental phasing to get high-resolution phase information; however derivative screening and heavy atom location will be greatly facilitated by the phases to low resolution, for example using MR-SAD in Phaser.

A detailed description of the protocol for phasing with EM maps has been published [107].

## 13 Notes

1. The term `molecular replacement' was coined by Michael Rossman [108] for methods that exploit non-crystallographic symmetry for phasing, whether within or between crystal forms. However, it has come to mean the case where an unknown structure is solved with a known structure [109]. Other uses of the technique are now referred to as 'non-crystallographic symmetry averaging' and 'cross-crystal averaging'.
2. Low homology models are detected with multiple sequence alignment methods and have benefitted greatly from whole genome sequencing. For a detailed discussion of sequence database searches, see the chapter by DiMaio in this volume.
3. The natural way to build a solution by Patterson methods is to identify the correct placement of each component independently before assembling the solution. While it is possible to account for partial structures with Patterson translation functions or the correlation coefficient, accounting for partial structure in Patterson rotation functions is much more difficult. Patterson subtraction methods for the rotation function are highly

susceptible to differences in B-factors between the component placed and the component remaining to be found, as well as coordinate differences. With low signal-to-noise for the rotation function, solutions are easily lost

4. How many is "too many"? It depends on the time and computational resources available to the crystallographer, the possibility of better data becoming available, other options for structure solution, and significance of the project.
5. There are several other ways to combine experimental phasing with MR. If experimental phases can be determined (*i.e.* substructures found), then spherically averaged phased translation functions [110] and phased translation functions [111] can be used to dock models into the experimentally determined electron density [10]. If a MR solution is clear, then experimental phases can be extracted even from poor derivatives by using the MR solution to determine the substructure. The MR-SAD [112] (MR-single-wavelength anomalous dispersion) version of this technique is particularly common, and can be performed in Phaser (see documentation for details [34]).
6. Wide Search MR can be used to resolve structure solution in cases when a protein contaminant accidentally crystallizes rather than the protein of interest. MR using models with sequence identity to the intended target will obviously fail [113].

## 14 Conclusions

Just because MR has solved a structure does not mean that refinement will be straightforward. Because of the sensitivity of the LLGI target, MR solutions can be obtained when the phase accuracy is very low. Solutions with low phase accuracy will have model bias, and will struggle to show novel features in the electron density that could move structure solution forward. Even if MR is showing a clear solution, the approaches described here in the context of improving the models prior to MR, can also be used as an additional step between MR and refinement.

Advanced MR strategies will, almost by definition, remain non-automated. However, methods continue to be developed at the boundaries of MR and the comments here will be superseded as advances are made. It is the responsibility of crystallographic software developers to maintain

good communication about advanced techniques, so that out-dated approaches do not remain part of the crystallographic folklore.

**Acknowledgements**

I thank Isobel Usón for content suggestions and for proposing the title, and Randy Read for critical reading of the manuscript, discussions and the concept for Figure 1. This work was supported by grant BB/L006014/1 from the BBSRC, UK.

## Figure Captions

Figure 1. Deducing $E_{calc}$ and $\sigma_A$ of from a set of $E_{obs}$. The likelihood of $E_{obs}$ given $E_{calc}$ is given by the Rice distribution[15,114]. Twenty-five $E_{obs}$ were randomly generated from a Rice distribution with $E_{calc}$ = 1.3 and a $\sigma_A$ = 0.8. The vertical bars correspond to the $E_{obs}$. The height of each bar represents the probability of $E_{obs}$, given the $E_{calc}$ and $\sigma_A$ of the Rice distribution shown. The log-likelihood is the sum of the log-likelihoods for each $E_{obs}$.

(a) Twenty-five $E_{obs}$ shown with the Rice function that was used to generate them. The centre of the distribution is most heavily populated by the data, and none of the probabilities is very low. The total log-likelihood = -12.5912

(b) Change the $E_{calc}$ of the Rice distribution to 2. The $E_{obs}$ on the low end of the Rice distribution become very improbable, which will reduce the likelihood. Fewer of the data points are now in the peak region. The total log-likelihood =-41.9852

(c) Change the $E_{calc}$ of the Rice distribution to 0.3. The total log-likelihood -29.4902

(d) Change the $\sigma_A$ of the Rice distribution to 0.95. In the heavily populated centre, the probability values go up, but the values in the two tails go down even more, so that the overall value of the likelihood is reduced. The total log-likelihood = -33.1789

(e) Change the $\sigma_A$ of the Rice distribution to 0.3. The probabilities in the tails go up, but the decrease in the heavily-populated peak. The total log-likelihood = -17.2546

(f) Contour plot of the log-likelihood for pairs of $E_{calc}$ and $\sigma_A$. The peak in this distribution (black dot) is close to the $E_{calc}$ and $\sigma_A$ that were used in generating the data (red dot). With the correct $E_{calc}$ the likelihood function will balance out the influence of the sparsely-populated tails and the heavily-populated centre to give the correct $\sigma_A$

Figure 2. Dependence of LLGI on parameters of the data and the model. The total LLGI is the sum of the LLGI for each reflection, so the more reflections the higher the LLGI. The LLGI per reflection depends on $E_{calc}$, $\sigma_A$ and $\sigma_{Iobs}$. The $\sigma_A$ values are estimated from the fraction scattering of the model and the expected rmsd.

**References**


[1] Tollin P. (1969). Determination of the orientation and position of the myoglobin molecule in the crystal of seal myoglobin. J Mol Biol. 45(3):481–90.

[2] Ward KB, Wishner BC, Lattman EE, Love WE. (1975). Structure of deoxyhemoglobin a crystals grown from polyethylene glycol solutions. J Mol Biol. 98(1):161–77.

[3] Schmid MF, Herriott JR, Lattman EE. (1974). The structure of bovine carboxypeptidase B: results at 5·5 Ångström resolution. J Mol Biol. 84(1):97–101.

[4] Rossmann MG, Blow DM. (1962). The detection of sub-units within the crystallographic asymmetric unit. Acta Crystallogr. 15(1):24–31.

[5] McCoy AJ. (2007). Solving structures of protein complexes by molecular replacement with Phaser. Acta Crystallogr D. 63(1):32–41.

[6] Rupp B. (2010). Biomolecular crystallography: principles, practice and applications to structural biology. 808 p.

[7] Axel T Brunger. (1992). X-PLOR: Version 3.1 A System For X-ray Crystallography And NMR. Axel T Brunger, editor. 382 (xvii, 382 pages).

[8] Brünger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS, Read RJ, Rice LM, Simonson T, Warren GL. (1998). Crystallography & NMR System: A New Software Suite for Macromolecular Structure Determination. Acta Crystallogr D. 54(5):905–21.

[9] Navaza J. (2001). Implementation of molecular replacement in AMoRe. Acta Crystallogr D. 57(10):1367–72.

[10] Vagin A, Teplyakov A. (2010). Molecular replacement with MOLREP. Acta Crystallogr D. 66(1):22–5.

[11] Kissinger CR, Gehlhaar DK, Fogel DB. (1999). Rapid automated molecular replacement

by evolutionary search. Acta Crystallogr D. 55(2):484–91.

[12] Glykos NM, Kokkinidis M. (2001). Multidimensional molecular replacement. Acta Crystallogr D. 57(10):1462–73.

[13] Jamrog DC, Zhang Y, Phillips GN. (2003). SOMoRe: a multi-dimensional search and optimization approach to molecular replacement. Acta Crystallogr D. 59(2):304–14.

[14] Jogl G, Tao X, Xu Y, Tong L. (2001). COMO: a program for combined molecular replacement. Acta Crystallogr D. 57(8):1127–34.

[15] McCoy AJ, Grosse-Kunstleve RW, Adams PD, Winn MD, Storoni LC, Read RJ. (2007). Phaser crystallographic software. J Appl Crystallogr. 40(4):658–74.

[16] Toth EA. (2007). Molecular replacement. Methods Mol Biol. 364:121–48.

[17] Berman H, Henrick K, Nakamura H. (2003). Announcing the worldwide Protein Data Bank. Nat Struct Biol. 10(12):980.

[18] Scapin G. (2013). Molecular replacement then and now. Acta Crystallogr D. 69(11):2266–75.

[19] Marcia M, Humphris-Narayanan E, Keating KS, Somarowthu S, Rajashankar K, Pyle AM. (2013). Solving nucleic acid structures by molecular replacement: examples from group II intron studies. Acta Crystallogr D. 69(11):2174–85.

[20] Read RJ, McCoy AJ. (2016). A log-likelihood-gain intensity target for crystallographic phasing that accounts for experimental error. Acta Crystallogr D. 72(3):375–87.

[21] Read RJ. (2001). Pushing the boundaries of molecular replacement with maximum likelihood. Acta Crystallogr D. 57(10):1373–82.

[22] French S, Wilson K. (1978). On the treatment of negative intensity observations. Acta Crystallogr A. 34(4):517–25.

[23] Winn MD, Ballard CC, Cowtan KD, Dodson EJ, Emsley P, Evans PR, Keegan RM, Krissinel EB, Leslie AGW, McCoy A, McNicholas SJ, Murshudov GN, Pannu NS, Potterton EA, Powell HR, Read RJ, Vagin A, Wilson KS. (2011). Overview of the CCP4 suite and current developments. Acta Crystallogr D. 67(4):235–42.

[24] Adams PD, Afonine P V, Bunkóczi G, Chen VB, Davis IW, Echols N, Headd JJ, Hung L-WL-W, Kapral GJ, Grosse-Kunstleve RW, McCoy AJ, Moriarty NW, Oeffner R, Read RJ, Richardson DC, Richardson JS, Terwilliger TC, Zwart PH. (2010). PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr D. 66(2):213–21.

[25] Potterton E, Briggs P, Turkenburg M, Dodson E. (2003). A graphical user interface to the CCP 4 program suite. Acta Crystallogr D. 59(7):1131–7.

[26] Echols N, Grosse-Kunstleve RW, Afonine P V, Bunkóczi G, Chen VB, Headd JJ, McCoy AJ, Moriarty NW, Read RJ, Richardson DC, Richardson JS, Terwilliger TC, Adams PD. (2012). Graphical tools for macromolecular crystallography in PHENIX. J Appl Crystallogr. 45(3):581–6.

[27] Keegan RM, Winn MD. (2008). MrBUMP: an automated pipeline for molecular replacement. Acta Crystallogr D. 64(1):119–24.

[28] Bunkóczi G, Echols N, McCoy AJ, Oeffner RD, Adams PD, Read RJ. (2013). Phaser.MRage: Automated molecular replacement. Acta Crystallogr D. 69(11):2276–86.

[29] Stokes-Rees I, Sliz P. (2010). Protein structure determination by exhaustive search of Protein Data Bank derived databases. Proc Natl Acad Sci U S A. 107(50):21476–81.

[30] Strong M, Sawaya MR, Wang S, Phillips M, Cascio D, Eisenberg D. (2006). Toward the structural genomics of complexes: crystal structure of a PE/PPE protein complex from Mycobacterium tuberculosis. Proc Natl Acad Sci U S A. 103(21):8060–5.

[31] Rodríguez DD, Grosse C, Himmel S, González C, de Ilarduya IM, Becker S, Sheldrick GM, Usón I. (2009). Crystallographic ab initio protein structure solution below atomic

resolution. Nat Methods. 6(9):651–3.

[32] Bibby J, Keegan RM, Mayans O, Winn MD, Rigden DJ. (2012). AMPLE: A cluster-and-truncate approach to solve the crystal structures of small proteins using rapidly computed ab initio models. Acta Crystallogr D. 68(12):1622–31.

[33] Wilson AJC. (1942). Determination of Absolute from Relative X-Ray Intensity Data. Nature. 150(3796):152–152.

[34] McCoy AJ, Read RJ, Bunkóczi G, Oeffner RD. Phaserwiki. http://www.phaser.cimr.cam.ac.uk.

[35] Evans P, McCoy A. (2008). An introduction to molecular replacement. Acta Crystallogr D. 64(1):1–10.

[36] Ten Eyck LF. (1973). Crystallographic fast Fourier transforms. Acta Crystallogr A. 29(2):183–91.

[37] Storoni LC, McCoy AJ, Read RJ. (2004). Likelihood-enhanced fast rotation functions. Acta Crystallogr D. 60(3):432–8.

[38] McCoy AJ, Grosse-Kunstleve RW, Storoni LC, Read RJ. (2005). Likelihood-enhanced fast translation functions. Acta Crystallogr D. 61(4):458–64.

[39] Oeffner RD, Bunkóczi G, McCoy AJ, Read RJ. (2013). Improved estimates of coordinate error for molecular replacement. Acta Crystallogr D. 69(11):2209–15.

[40] Long F, Vagin AA, Young P, Murshudov GN. (2008). BALBES: a molecular-replacement pipeline. Acta Crystallogr D. 64(1):125–32.

[41] Rosetta Commons. https://www.rosettacommons.org/about/pubs.

[42] DiMaio F, Echols N, Headd JJ, Terwilliger TC, Adams PD, Baker D. (2013). Improved low-resolution crystallographic refinement with Phenix and Rosetta. Nat Methods. 10(11):1102–4.

[43] Pedersen BP, Gourdon P, Liu X, Karlsen JL, Nissen P. (2016). Initiating heavy-atom-based phasing by multi-dimensional molecular replacement. Acta Crystallogr D. 72(3):440–5.

[44] Urzhumtseva L, Urzhumtsev A. (2002). COMPANG: automated comparison of orientations. J Appl Crystallogr. 35:644–7.

[45] Buehler A, Urzhumtseva L, Lunin VY, Urzhumtsev A. (2009). Cluster analysis for phasing with molecular replacement: a feasibility study. Acta Crystallogr D. 65(7):644–50.

[46] Phillips DC, Rogers D, Wilson AJC. (1950). Reliability index for centrosymmetric and non-centrosymmetric structures. Acta Crystallogr. 3(5):398–9.

[47] Navaza J. (1994). AMoRe : an automated package for molecular replacement. Acta Crystallogr A. 50(2):157–63.

[48] Fujinaga M, Read RJ. (1987). Experiences with a new translation-function program. J Appl Crystallogr. 20(6):517–21.

[49] Delarue M. (2007). Molecular replacement techniques for high-throughput structure determination. In: Sanderson MR, Skelly J V., editors. Macromolecular Crystallography: Conventional and high-throughput methods. Oxford University Press: Oxford.

[50] Abergel C. (2013). Molecular replacement: tricks and treats. Acta Crystallogr D. 69(11):2167–73.

[51] Turkenburg JP, Dodson EJ. (1996). Modern developments in molecular replacement. Curr Opin Struct Biol. 6(5):604–10.

[52] Dodson E. (2008). The befores and afters of molecular replacement. Acta Crystallogr D. 64(1):17–24.

[53] Schwarzenbacher R, Godzik A, Grzechnik SK, Jaroszewski L. (2004). The importance of

alignment accuracy for molecular replacement. Acta Crystallogr D. 60(7):1229–36.

[54] Qian B, Raman S, Das R, Bradley P, McCoy AJ, Read RJ, Baker D. (2007). High-resolution structure prediction and the crystallographic phase problem. Nature. 450(7167):259–64.

[55] Bahar I, Atilgan AR, Erman B. (1997). Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. 2(3) p. 173–81.

[56] Haliloglu T, Bahar I. (1999). Structure-based analysis of protein dynamics: comparison of theoretical results for hen lysozyme with X-ray diffraction and NMR relaxation data. Proteins. 37(4):654–67.

[57] Tirion M. (1996). Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. Phys Rev Lett. 77(9):1905–8.

[58] Tama F, Sanejouand YH. (2001). Conformational change of proteins arising from normal mode calculations. Protein Eng. 14(1):1–6.

[59] Krebs WG, Alexandrov V, Wilson CA, Echols N, Yu H, Gerstein M. (2002). Normal mode analysis of macromolecular motions in a database framework: developing mode concentration as a useful classifying statistic. Proteins. 48(4):682–95.

[60] Suhre K, Sanejouand YH. (2004). On the potential of normal-mode analysis for solving difficult molecular-replacement problems. Acta Crystallogr D. 60(4):796–9.

[61] Blaszczyk J, Li Y, Yan H, Ji X. (2001). Crystal structure of unligated guanylate kinase from yeast reveals GMP-induced conformational changes. J Mol Biol. 307(1):247–57.

[62] SBGrid Science Portal. https://portal.sbgrid.org/d/apps/wsmr/docs.

[63] Zhou A, Carrell RW, Murphy MP, Wei Z, Yan Y, Stanley PLD, Stein PE, Broughton Pipkin F, Read RJ. (2010). A redox switch in angiotensinogen modulates angiotensin release. Nature. 468(7320):108–11.

[64] Tronrud DE. (1997). TNT refinement package. Methods Enzymol. 277:306–19.

[65] Fokine A, Capitani G, Grütter MG, Urzhumtsev A. (2003). Bulk-solvent correction for fast translation search in molecular replacement: service programs for AMoRe and CNS. J Appl Crystallogr. 36(2):352–5.

[66] Stein N. (2008). CHAINSAW : a program for mutating pdb files used as templates in molecular replacement. J Appl Crystallogr. 41(3):641–3.

[67] Bunkóczi G, Read RJ. (2011). Improvement of molecular-replacement models with Sculptor. Acta Crystallogr D. 67(4):303–12.

[68] Wriggers W, Schulten K. (1997). Protein domain movements: detection of rigid domains and visualization of hinges in comparisons of atomic coordinates. Proteins. 29(1):1–14.

[69] Hayward S, Berendsen HJ. (1998). Systematic analysis of domain motions in proteins from conformational change: new results on citrate synthase and T4 lysozyme. Proteins. 30(2):144–54.

[70] Schneider TR. (2000). Objective comparison of protein structures: error-scaled difference distance matrices. Acta Crystallogr D. 56(6):714–21.

[71] McCoy AJ, Nicholls RA, Schneider TR. (2013). SCEDS: protein fragments for molecular replacement in Phaser. Acta Crystallogr D. 69(11):2216–25.

[72] Wodak SJ, Janin J. (1980). Analytical approximation to the accessible surface area of proteins. Proc Natl Acad Sci U S A. 77(4):1736–40.

[73] Painter J, Merritt EA. (2006). Optimal description of a protein structure in terms of multiple groups undergoing TLS motion. Acta Crystallogr D. 62(4):439–50.

[74] Hinsen K. (1998). Analysis of domain motions by approximate normal mode calculations. Proteins. 33(3):417–29.

[75] Thomas JMH, Keegan RM, Bibby J, Winn MD, Mayans O, Rigden DJ. (2015). Routine

phasing of coiled-coil protein crystal structures with AMPLE. IUCrJ. 2(2):198–206.

[76] Sammito M, Millán C, Rodríguez DD, de Ilarduya IM, Meindl K, De Marino I, Petrillo G, Buey RM, de Pereda JM, Zeth K, Sheldrick GM, Usón I. (2013). Exploiting tertiary structure through local folds for crystallographic phasing. Nat Methods. 10(11):1099–101.

[77] Sammito M, Meindl K, de Ilarduya IM, Millán C, Artola-Recolons C, Hermoso JA, Usón I. (2014). Structure solution with ARCIMBOLDO using fragments derived from distant homology models. FEBS J. 281(18):4029–45.

[78] Chothia C, Lesk AM. (1986). The relation between the divergence of sequence and structure in proteins. EMBO J. 5(4):823–6.

[79] Bunkóczi G, Wallner B, Read RJ. (2015). Local error estimates dramatically improve the utility of homology models for solving crystal structures by molecular replacement. Structure. 23(2):397–406.

[80] Brünger AT. (1993). Structure Determination of Antibodies and Antibody-Antigen Complexes by Molecular Replacement. Immunomethods. 3(3):180–90.

[81] Stanfield RL, Zemla A, Wilson IA, Rupp B. (2006). Antibody elbow angles are influenced by their light chain class. J Mol Biol. 357(5):1566–74.

[82] Almagro JC, Beavers MP, Hernandez-Guzman F, Maier J, Shaulsky J, Butenhof K, Labute P, Thorsteinson N, Kelly K, Teplyakov A, Luo J, Sweet R, Gilliland GL. (2011). Antibody modeling assessment. Proteins. 79(11):3050–66.

[83] Griffin L, Lawson A. (2011). Antibody fragments as tools in crystallography. Clin Exp Immunol. 165(3):285–91.

[84] Tollin P, Rossmann MG. (1966). A description of various rotation function programs. Acta Crystallogr. 21(6):872–6.

[85] Jeffery P. Molecular Replacement Guide.

http://xray0.princeton.edu/~phil/Facility/Guides/MolecularReplacement.html.

[86] Ling H, Boodhoo A, Hazes B, Cummings MD, Armstrong GD, Brunton JL, Read RJ. (1998). Structure of the shiga-like toxin I B-pentamer complexed with an analogue of its receptor Gb3. Biochemistry. 37(7):1777–88.

[87] Karplus PA, Diederichs K. (2012). Linking Crystallographic Model and Data Quality. Science (80- ). 336(6084):1030–3.

[88] Matthews BW. (1968). Solvent content of protein crystals. J Mol Biol. 33(2):491–7.

[89] Kantardjieff KA, Rupp B. (2003). Matthews coefficient probabilities: Improved estimates for unit cell contents of proteins, DNA, and protein-nucleic acid complex crystals. Protein Sci. 12(9):1865–71.

[90] Weichenberger CX, Rupp B. (2014). Ten years of probabilistic estimates of biocrystal solvent content: new insights via nonparametric kernel density estimate. Acta Crystallogr D. 70(6):1579–88.

[91] Sawaya MR. (2007). Characterizing a crystal from an initial native dataset. Methods Mol Biol. 364:95–120.

[92] Read RJ, Adams PD, McCoy AJ. (2013). Intensity statistics in the presence of translational noncrystallographic symmetry. Acta Crystallogr D. 69(2):176–83.

[93] Kleywegt GJ, Read RJ. (1997). Not your average density. Structure. 5(12):1557–69.

[94] Lebedev AA, Vagin AA, Murshudov GN. (2006). Intensity statistics in twinned crystals with examples from the PDB. Acta Crystallogr D. 62(1):83–95.

[95] Yeates TO, Fam BC. (1999). Protein crystals and their evil twins. Structure. 7(2):R25–9.

[96] Sliwiak J, Jaskolski M, Dauter Z, McCoy AJ, Read RJ. (2014). Likelihood-based molecular-replacement solution for a highly pathological crystal with tetartohedral twinning and sevenfold translational noncrystallographic symmetry. Acta Crystallogr D.

70(2):471–80.

[97] Evans P. (2006). Scaling and assessment of data quality. Acta Crystallogr D. 62(1):72–82.

[98] Zwart PH, Grosse-Kunstleve RW, Adams PD. (2005). Xtriage and Fest: automatic assessment of X-ray data and substructure structure factor estimation. CCP4 Newsl Protein Crystallogr. 43:27–35.

[99] Lebedev AA, Isupov MN. (2014). Space-group and origin ambiguity in macromolecular structures with pseudo-symmetry and its treatment with the program Zanuda. Acta Crystallogr D. 70(9):2430–43.

[100] Herbst-Irmer R, Sheldrick GM. (1998). Refinement of Twinned Structures with SHELXL97. Acta Crystallogr B. 54(4):443–9.

[101] Dauter Z. (2003). Twinned crystals and anomalous phasing. Acta Crystallogr D. 59(11):2004–16.

[102] Harada Y, Lifchitz A, Berthou J, Jolles P. (1981). A translation function combining packing and diffraction information: an application to lysozyme (high-temperature form). Acta Crystallogr A. 37(3):398–406.

[103] Vagin A, Teplyakov A. (1997). MOLREP : an Automated Program for Molecular Replacement. J Appl Crystallogr. 30(6):1022–5.

[104] Navaza J, Vernoslova E, IUCr. (1995). On the fast translation functions for molecular replacement. Acta Crystallogr A. 51(4):445–9.

[105] Wang BC. (1985). Resolution of phase ambiguity in macromolecular crystallography. Methods Enzymol. 115:90–112.

[106] Bai X, McMullan G, Scheres SH. (2014). How cryo-EM is revolutionizing structural biology. Trends Biochem Sci. 40(1):49–57.

[107] Jackson RN, McCoy AJ, Terwilliger TC, Read RJ, Wiedenheft B. (2015). X-ray structure

determination using low-resolution electron microscopy maps for molecular replacement. Nat Protoc. 10(9):1275–84.

[108] Rossmann MG. (1972). The molecular replacement method. Rossmann MG, editor.

[109] Rossmann MG. (2001). Molecular replacement – historical background. Acta Crystallogr D. 57(10):1360–6.

[110] Vagin AA, Isupov MN. (2001). Spherically averaged phased translation function and its application to the search for molecules and fragments in electron-density maps. Acta Crystallogr D. 57(10):1451–6.

[111] Colman PM, Fehlhammer H. (1976). The use of rotation and translation functions in the interpretation of low resolution electron density maps. J Mol Biol. 100(3):278–82.

[112] Schuermann JP, Tanner JJ. (2003). MRSAD: using anomalous dispersion from S atoms collected at Cu K α wavelength in molecular-replacement structure determination. Acta Crystallogr D. 59(10):1731–6.

[113] Niedzialkowska E, Gasiorowska O, Handing KB, Majorek KA, Porebski PJ, Shabalin IG, Zasadzinska E, Cymborowski M, Minor W. (2016). Protein purification and crystallization artifacts: The tale usually not told. Protein Sci. 25(3):720–33.

[114] Rice SO. (1945). Mathematical Analysis of Random Noise. Bell Syst Tech J. 24(1):46–156.