Ab initio solution of macromolecular crystal structures without direct methods

Airlie J McCoy[1], Robert D Oeffner[1], Antoni G Wrobel[2], Juha RM Ojala[3], Karl Tryggvason[3,4], Bernhard Lohkamp[5], Randy J Read[1*]

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

[2]Department of Clinical Biochemistry, Cambridge Institute for Medical Research, University of Cambridge, Wellcome Trust/MRC Building, Hills Road, Cambridge CB2 0XY, UK

[3]Division of Matrix Biology, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, 171 77 Stockholm, Sweden

[4]Cardiovascular and Metabolic Disorders Program, Duke-NUS Medical School, 16957 Singapore

[5]Division of Molecular Structural Biology, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, 171 77 Stockholm, Sweden

[*]Corresponding author: rjr27@cam.ac.uk, Tel: +44 1223 336500

Classification: Biological Sciences – Biophysics and Computational Biology

Keywords: macromolecular crystallography, likelihood, *ab initio* phasing

Abstract

The majority of macromolecular crystal structures are determined using the method of molecular replacement, in which known related structures are rotated and translated to provide an initial atomic model for the new structure. A new theoretical understanding of the signal-to-noise ratio in likelihood-based molecular replacement searches has been developed to account for the influence of model quality and completeness, as well as the resolution of the diffraction data. Here we show that, contrary to current belief, molecular replacement need not be restricted to the use of models comprising a substantial fraction of the unknown structure. Instead, likelihood-based methods allow a continuum of applications depending predictably on the quality of the model and the resolution of the data. Unexpectedly, our new understanding of the signal-to-noise ratio in molecular replacement leads to the finding that, with data to sufficiently high resolution, fragments as small as single atoms of elements usually found in proteins can yield *ab initio* solutions of macromolecular structures, including some that elude traditional direct methods.

Significance Statement

It is now possible to make an accurate prediction of whether or not a molecular replacement solution of a macromolecular crystal structure will succeed, given the quality of the model, its size, and the resolution of the diffraction data. This new understanding allows the development of powerful new structure-solution strategies, and leads to the unexpected finding that, with data to sufficiently high resolution, fragments as small as single atoms can be placed as the basis for *ab initio* structure solutions.

Introduction

Over the past century, determination of novel crystal structures has evolved from an exercise in logic identifying the locations of single atoms by inspecting diffraction patterns (1) or vector maps (2), through the development of direct methods for small molecules (3) and of isomorphous replacement (4, 5) or anomalous diffraction (6, 7) phasing for molecules as large as proteins.

Currently, about 80% of protein structures are solved by the method of molecular replacement (8), exploiting prior structural knowledge of related proteins. In principle, molecular replacement (MR) involves rotational and translational searches over many possible placements of a molecular model within the unit cell of an unknown structure. The most sensitive method of evaluating the fit to the observed data is a likelihood function (9, 10) that accounts for the effect of measurement errors in the observed diffraction intensities (11). Potential solutions are scored by the log-likelihood-gain on intensities (*LLGI*), the sum of the log-likelihoods for individual reflections minus the log-likelihoods for an uninformative model (see Methods).

Success in MR depends on the signal-to-noise of the search, which varies according to two parameters in the likelihood function: D_{obs} characterises the precision of each measurement, taking values near 1 for moderately well-measured data and only taking values near 0 for extremely weak data; σ_A measures the quality of the model in terms of the fraction of a crystallographic structure factor that it explains. The resolution-dependent value of σ_A for each reflection can be estimated from the fraction (f_P) of the X-ray scattering power accounted for by the model (where the total scattering power is the sum of the squares of the scattering factors for the atoms in the crystal), its estimated accuracy (RMS error Δ), and the

resolution (d) of the reflection (9), with (optionally) a correction for the effect of disordered solvent described by the parameters f_{sol} and B_{sol}:

$$\sigma_A = \sqrt{f_P\left[1 - f_{sol}\exp\left(-\frac{B_{sol}}{4d^2}\right)\right]}\exp\left(-\frac{2\pi^2}{3}\frac{\Delta^2}{d^2}\right) \quad (1a)$$

$$\sigma_A \approx \sqrt{f_P}\exp\left(-\frac{2\pi^2}{3}\frac{\Delta^2}{d^2}\right) \quad (1b)$$

The simpler expression in equation (1b) neglects the effect of disordered solvent at low resolution.

The signal for an MR search can be estimated prior to the calculation as the expected value, or probability-weighted average, of the *LLGI* for a correctly placed model. The expected value of the contribution of one reflection, $\langle LLGI\rangle_{hkl}$, can be approximated simply by $D_{obs}^4\sigma_A^4/2$ (see Methods), an approximation that is particularly good for the low values of $D_{obs}\sigma_A$ characterising the difficult cases of most interest. In the following, we refer to the total expected *LLGI*, summed over all reflections, as the *eLLG*.

The variance of *eLLG* can similarly be approximated as the sum over all reflections of $D_{obs}^4\sigma_A^4$, leading to the conclusion that the expected signal-to-noise ratio in an MR search will be proportional to $\sqrt{eLLG}$ (see Methods). By the same reasoning, the signal-to-noise ratio achieved in a particular search will be proportional to $\sqrt{LLGI}$. The theoretical deduction that confidence in an MR solution can be judged simply by the *LLGI* value has been validated by analysing a database of nearly 22,000 MR calculations, where an *LLGI* of 60 or more in a 6-dimensional rotation/translation search typically indicates a correct solution. (See Figure 1, which also shows that the required signal scales with the number of degrees of freedom in the search.) The database of test calculations also reveals that the translation function Z-score (TFZ: the number of standard deviations by which the translation function peak exceeds its

mean) is roughly on the same scale as $\sqrt{LLGI}$, though the exact relationship depends on the number of primitive symmetry operators; this justifies the success of TFZ as a measure of confidence (10).

An *LLGI* at the level required to distinguish the correct solution from up to millions of alternatives can be achieved by predictable trade-offs among model quality, completeness and resolution of the data used. For example, this theoretical insight explains why it is possible to place individual α-helices with better than random success in the Arcimboldo pipeline (12), but also why it is a great advantage to have data extending beyond 2 Å resolution: helices are preserved very well, so that Δ is small and data to the highest resolution will contribute to the signal. The theory also predicts, correctly, that calculations limited to around 10 Å resolution can give unambiguous MR solutions for ribosome structures, because of the large numbers of diffraction observations available to that resolution with the large ribosomal unit cell. Importantly, it also allows researchers to anticipate when MR is unlikely to succeed, so that they avoid fruitless calculations.

This new insight led us to consider the most extreme example of a small fragment, *i.e.* a single atom. A single atom is a perfect partial model (Δ=0), for which $\sigma_A^2 = f_P$ and hence $\langle LLGI \rangle_{hkl} \propto f_P^2$ for well-measured data regardless of the resolution. With high-resolution data containing a sufficient number of reflections, the *eLLG* can rise to a substantial number. This is particularly true for atoms that are somewhat heavier than average. For instance, the square of the scattering power of a sulphur atom (*i.e.* the fourth power of its scattering factor) is about 50 times greater than that of a carbon atom at a very low resolution such as 10 to 20 Å; because scattering drops off less rapidly for sulphur, that ratio increases to about 300 at 1 Å resolution. This effect is amplified if a sulphur atom is better ordered than the average atom in the structure, because its relative scattering power becomes even greater.

Furthermore, only half as much signal should be required to place a single atom with 3 degrees of freedom compared to a molecule with 6 degrees of freedom (Figure 1). Our new insights predict that, for crystals that contain up to a few thousand unique ordered atoms and diffract beyond about 1 Å resolution, there should be a significant signal in a likelihood search carried out by translating a single sulphur atom over all of its possible positions. Even if the placement of the first atom is ambiguous, the signal will increase quadratically with the number of atoms placed (Figure 2), allowing the ambiguity to be resolved.

Results

Test calculations on a number of systems proved the principle of single-atom MR: it was indeed possible to find sulphur atoms in a variety of protein crystals, as well as phosphorus atoms in one RNA crystal tested (Table S1). The largest structure that yielded to this approach was that of aldose reductase (PDB entry 3bcj) (13). The protein has a mass of 36kDa with 2525 non-hydrogen atoms (2606 including ligands) and no atom heavier than sulphur, and the deposited data extend to 0.78 Å resolution. The *eLLG* for a sulphur atom with a B-factor equal to the average in the crystal is 4.0, or 12.6 for a well-ordered sulphur atom with a B-factor reduced by only 1 $Å^2$. MR implemented in Phaser was able to locate up to 10 atoms with clear signal (Table 1).

A structure comprising a few atoms can then serve as a seed for structure completion by using log-likelihood-gradient maps to select locations for new nitrogen atoms (as a surrogate for other types) that improve the MR likelihood score (14) (see Methods). Starting from as few as the first 2 atoms placed by MR, the structure of aldose reductase was extended successfully by log-likelihood-gradient completion. The result was a model with 3051 atoms (some accounting for solvent molecules and for static disorder) that yields an *LLGI* of 483292 and an R-value of 12.9% (Figure 3). In contrast, all attempts to solve this structure by direct

methods or their dual-space variants (15, 16) have failed. As far as we can determine, it is the largest reported *ab initio* structure containing nothing heavier than the sulphur atoms found in natural protein sequences, although larger *ab initio* structures containing metal ions have been solved (17).

The new formulation predicts that it should also be possible to place sulphur atoms in smaller structures at lower resolution. This was crucial in solving a previously unknown structure, the N-terminal domain (residues 22-95) of Shisa3, which crystallised in space group $P4_32_12$ and diffracted to 1.39 Å resolution. The protein did not have detectable sequence identity with any protein in the PDB, so there was no template structure for traditional MR. The *eLLG* calculations predict that there should be some signal for placing well-ordered sulphur atoms, giving an *eLLG* of 4.0 for a sulphur atom with a B-factor reduced by 1.5 $Å^2$ from the average. Indeed, up to 7 of the 8 sulphur atoms in this protein could be placed with good signal (Table 1).

Log-likelihood-gradient completion is expected to work more poorly at resolutions where atomic peaks are not resolved. Nonetheless, this succeeded in expanding the Shisa3 structure to a total of 56 atoms, with the additional atoms largely corresponding to well-ordered main-chain oxygen and nitrogen atoms. At this point, the phase information was sufficient to enable phase improvement by density modification in Parrot (18), and the resulting map could be interpreted in terms of an atomic model in ARP/wARP (19). A hybrid approach exploiting direct methods algorithms implemented in ACORN (17, 20) or in SHELXE (21) was also able to expand a partial structure obtained by single-atom MR. This succeeded when starting from as little as one pair of sulphur atoms (Figure 4). The structure, which contains no α-helices and represents a novel protein fold, was refined to an R-value of 11.5% and has

been deposited in the PDB with accession code 5m0w. Details of the structure will be discussed elsewhere.

Discussion

This work brings together high resolution *ab initio* phasing and low resolution MR in one unified framework that spans the continuum of data and model quality, with the *eLLG* directing the tailoring of structure solution to the optimal path for the data available. It demonstrates the considerable practical impact, compared to traditional direct methods, of accounting rigorously for the effects of sources of error in a likelihood target. It is also important to note that these results have been obtained by a deterministic algorithm. Direct methods, in contrast, are invariably implemented within a random multi-solution framework, an approach that should also improve the outcome of single-atom MR. Finally, the results were obtained without taking advantage of any other information that would typically be present, *e.g.* from single-wavelength anomalous diffraction (SAD) effects in crystals with intrinsic anomalous scatterers such as sulphur, or even from isomorphous replacement experiments. A proper accounting for the effects of uncertainty, as demonstrated here, should allow us to extend our approach to use even weak information from these other sources.

Methods

Formalism for the *eLLG* and its approximation

The likelihood function used to score MR solutions is based on the Rice distribution (9, 10), modified to account for the effect of measurement errors in the observed intensities (11). For acentric reflections, this is given by

$$p_a(E_e; E_C) = \frac{2E_e}{1-D_{obs}^2\sigma_A^2} \exp\left[-\frac{E_e^2+(D_{obs}\sigma_A E_C)^2}{1-D_{obs}^2\sigma_A^2}\right] \mathrm{I}_0\left(\frac{2D_{obs}\sigma_A E_e E_C}{1-D_{obs}^2\sigma_A^2}\right) \quad (2)$$

where E_e (an effective normalised structure factor amplitude) and D_{obs} (an estimate of its precision) are derived from the observed intensity and its standard error, E_C is the normalised structure factor amplitude calculated from the placed model, σ_A is the fraction of the calculated structure factor that is correlated with the true structure factor and I_0 is a modified Bessel function of order 0.

The *eLLG* is defined as the probability-weighted average of the logarithm of the likelihood ratio, integrated over all pairs of observed and calculated normalised structure factors. The contribution of a single reflection to the *eLLG* is defined in equation (3).

$$\langle LLGI \rangle_{hkl} = \int_0^\infty \int_0^\infty p(E_e, E_C) \ln\left(\frac{p(E_e;E_C)}{p(E_e)}\right) \mathrm{d}E_e \mathrm{d}E_C \tag{3a}$$

where, for the acentric case,

$$p_a(E_e, E_C) = \frac{4E_e E_C}{1-D_{obs}^2\sigma_A^2} \exp\left(-\frac{E_e^2+E_C^2}{1-D_{obs}^2\sigma_A^2}\right) \mathrm{I}_0\left(\frac{2D_{obs}\sigma_A E_e E_C}{1-D_{obs}^2\sigma_A^2}\right) \tag{3b}$$

and

$$p(E_e) = 2E_e \exp(-E_e^2) \tag{3c}$$

The Maclaurin series expansion of the integrand of equation (3a) for the acentric case, to fourth order in $D_{obs}\sigma_A$, is given in equation (4):

$$p_a(E_e, E_C) \ln\left(\frac{p_a(E_e;E_C)}{p_a(E_e)}\right) \approx a + bD_{obs}^2\sigma_A^2 + cD_{obs}^4\sigma_A^4 \tag{4a}$$

where

$$a = 4\mathrm{e}^{-E_e^2-E_C^2} E_e E_C \left[\ln\left(E_e \mathrm{e}^{-E_e^2}\right) - \ln\left(E_C \mathrm{e}^{-E_C^2}\right)\right] \tag{4b}$$

$$b = 4\mathrm{e}^{-E_e^2-E_C^2} E_e E_C (1-E_e^2)(1-E_C^2) \left[1 + \ln\left(E_e \mathrm{e}^{-E_e^2}\right) - \ln\left(E_C \mathrm{e}^{-E_C^2}\right)\right] \tag{4c}$$

$$c = \mathrm{e}^{-E_e^2 - E_C^2} E_e E_C \left\{ 6 + 4E_C^2[E_C^2 - 3] - 4[3 - 6E_C^2 + 2E_C^4]E_e^2 + [4 - 8E_C^2 + 3E_C^4]E_e^4 + [2 - 4E_e^2 + E_e^4][2 - 4E_C^2 + E_C^4] \left[\ln\left(E_e \mathrm{e}^{-E_e^2}\right) - \ln\left(E_C \mathrm{e}^{-E_C^2}\right) \right] \right\} \quad (4d)$$

The double integrals over *a* and *b* both evaluate to zero, whereas the double integral over c yields $1/2$. Figure S1 shows that $D_{obs}^4 \sigma_A^4/2$ is an excellent approximation to $\langle LLGI \rangle_{hkl}$, especially for the smaller values of $D_{obs}\sigma_A$ that would be encountered in difficult structure solutions. Though the forms of the probability distributions for the centric case are different, the same result is achieved by integrating a series expansion, *i.e.* that $\langle LLGI \rangle_{hkl}$ is approximately equal to $D_{obs}^4 \sigma_A^4/2$.

The variance of $\langle LLGI \rangle_{hkl}$ is defined in equation (5).

$$\sigma^2(\langle LLGI \rangle_{hkl}) = \langle LLGI^2 \rangle_{hkl} - \langle LLGI \rangle_{hkl}^2 \quad (5a)$$

where

$$\langle LLGI^2 \rangle_{hkl} = \int_0^\infty \int_0^\infty p(E_e, E_C) \ln\left(\frac{p(E_e;E_C)}{p(E_e)}\right)^2 \mathrm{d}E_e \mathrm{d}E_C \quad (5b)$$

For the small values of $D_{obs}\sigma_A$ that characterise difficult cases, equation (5a) will be dominated by the first term (as the second term will have a value of the order of $D_{obs}^8 \sigma_A^8$). The Maclaurin series expansion of the integrand of equation (5b) for the acentric case, to fourth order in $D_{obs}\sigma_A$, is given in equation (6):

$$p_a(E_e, E_C) \ln\left(\frac{p_a(E_e;E_C)}{p_a(E_e)}\right)^2 \approx 4\mathrm{e}^{-E_e^2 - E_C^2} E_e E_C (1 - E_e^2)^2 (1 - E_C^2)^2 D_{obs}^4 \sigma_A^4 \quad (6)$$

The double integral over this single term yields simply $D_{obs}^4 \sigma_A^4$. The same result is obtained for the contributions of centric reflections to the variance of the *eLLG*.

Because the variance of $\langle LLGI \rangle_{hkl}$ is proportional to $\langle LLGI \rangle_{hkl}$ itself, the variance of the total *eLLG*, summed over all reflections, is also proportional to the total *eLLG*. Therefore, the signal-to-noise ratio for any *eLLG* is proportional to $\sqrt{eLLG}$, regardless of how that *eLLG* is achieved through a combination of model quality, completeness, data quality and data resolution. Similarly, the value of *LLGI* obtained in an MR search will indicate the confidence that can be placed in the corresponding solution, regardless of how the *LLGI* was achieved. Indeed the translation function Z-score, which is used as a measure of confidence in a MR solution (10), is seen to be roughly proportional to the square root of the *LLGI* in the database of MR calculations.

Mathematical derivations

Series approximations and integrals used in the derivation of equations (3) to (6) were computed with *Mathematica* (22), which was also used to prepare Figures 2 and S1.

Single-atom MR protocol

In the single-atom MR protocol, the first step is to carry out translation searches for a specified number of the heavier atoms expected in the structure. For the trials summarised in Table S1, the search looked for 4 atoms unless fewer sufficiently heavy atoms were expected. In the next step, log-likelihood-gradient completion (described in the next section) was used to complete each of the potential few-atom solutions by adding nitrogen atoms as surrogates for all remaining atom types. Refinement, at each step, of the occupancies of the nitrogen atoms compensates for the difference in scattering power compared to other atom types, such as carbon or oxygen. The log-likelihood-gradient completion continues to convergence, when no further peaks are identified.

The test cases in Table S1 were chosen from the PDB based initially on the criteria that data extending to atomic resolution (1.2 Å or better) were deposited in the form of intensities rather than amplitudes, and that there were no atoms heavier than S in the structure. The initial set was supplemented with several cases at lower than 1 Å resolution in which there are atoms heavier than S, as the success rate was otherwise low in this resolution range. Note that the *LLGI* per atom after the initial search for individual heavier atoms provides a reasonable diagnostic indication of success. For the cases where the protocol succeeded, *LLGI*/atom ranged from 21.5 to 272.2 with a mean of 88.3, whereas for cases where the protocol failed, *LLGI*/atom ranged from 19.0 to 43.5 with a mean of 28.4. The difference in *LLGI*/atom distributions for the data from Table S1 is illustrated in Figure S2 by a box plot, generated with BoxPlotR (23).

Log-likelihood-gradient completion

In a log-likelihood-gradient map, peaks show positions where the addition of atoms of a specified type would tend to increase the corresponding likelihood target. The single-atom MR algorithm implemented in Phaser computes a log-likelihood-gradient map corresponding to the MR likelihood function, but does so by using the equivalent functionality required for handling singletons (reflections with only one member of a Friedel pair, hence no anomalous scattering phase information) in the SAD likelihood target (14). Peak-picking is carried out using the same defaults as for log-likelihood-gradient SAD completion, *i.e.* peaks above 6 times the RMS value of the map are selected, unless the deepest hole in the map has a greater magnitude. Log-likelihood-gradient completion is iterative, with the addition of atoms increasing the signal in subsequent log-likelihood-gradient maps.

Acknowledgments

This research was supported by a Principal Research Fellowship from the Wellcome Trust (R.J.R.: 082961/Z/07/Z), and grants from the NIH (R.J.R.: P01GM063210), the Swedish Research Council (B.L.: 2007-5648 and K.T.), the Knut and Alice Wallenberg Foundation (K.T.), the Novo Nordisk Foundation (K.T.) and the Röntgen Ångström Cluster (B.L.: 349-2013-597). The research was facilitated by a Wellcome Trust Strategic Award (100140) to the Cambridge Institute for Medical Research. The diffraction data were collected on beamline ID14-3 at the European Synchrotron Radiation Facility (ESRF), Grenoble, France. We are grateful to the Local Contact at the ESRF for providing assistance in using beamline ID14-3 as well as Doreen Dobritzsch for help with the data collection.

References

1. Bragg WL (1913) The structure of some crystals as indicated by their diffraction of X-rays. *Proc. Roy. Soc. A* **89**:248-277.

2. Patterson AL (1934) A Fourier series method for the determination of the components of interatomic distances in crystals. *Phys. Rev.* **46**:372-376.

3. Hauptman H, Karle J (1953) Solution of the Phase Problem. I. The Centrosymmetric Crystal. American Crystallographic Association Monograph No. 3.

4. Cork JM (1927) LX. The crystal structure of some of the alums. *Lond., Edinb. Dubl. Philosoph. Mag. J. Sci.* **4**:688-698.

5. Perutz MF (1956) Isomorphous replacement and phase determination in non-centrosymmetric space groups. *Acta Cryst.* **9**:867-873.

6. Bijvoet JM (1954) Structure of optically active compounds in the solid state. *Nature* **173**:888-891.

7. Hendrickson WA (1985) Analysis of protein structure from diffraction measurement at multiple wavelengths. *Trans. Am. Cryst. Assoc.* **21**:11-21.

8. Rossmann MG, Blow DM (1962) A method of positioning a known molecule in an unknown crystal structure. *Acta Cryst.* **15**:24-31.

9. Read RJ (2001) Pushing the boundaries of molecular replacement with maximum likelihood. *Acta Cryst*. **D57**:1373-1382.

10. McCoy AJ *et al*. (2007) Phaser crystallographic software. *J. Appl. Cryst.* **40**:658-674.

11. Read RJ, McCoy AJ (2016) A log-likelihood-gain intensity target for crystallographic phasing that accounts for experimental error. *Acta Cryst.* **D72**:375-387.

12. Rodríguez Martínez DD *et al*. (2009) ARCIMBOLDO: crystallographic ab initio protein structure solution below atomic resolution. *Nat. Methods* **6**:651-653.

13. Zhao HT *et al*. (2008) Unusual binding mode of the 2S4R stereoisomer of the potent aldose reductase cyclic imide inhibitor fidarestat (2S4S) in the 15 K crystal structure of the ternary complex refined at 0.78 A resolution: implications for the inhibition mechanism. *J. Med. Chem.* **51**:1478-1481.

14. McCoy AJ, Read RJ (2010) Experimental phasing: best practice and pitfalls. *Acta Cryst*. **D66**:458-469.

15. Weeks CM, DeTitta GT, Miller R, Hauptman HA (1993) Applications of the minimal principle to peptide structures. *Acta Cryst.* **D49**:179-181.

16. Sheldrick GM, Hauptman HA, Weeks CM, Miller M, Usón I (2001) Direct methods. **In** *International Tables for Macromolecular Crystallography*, Vol. *F*, eds Arnold E, Rossmann M (Kluwer Academic Publishers, Dordrecht), pp 333–345.

17. Dodson EJ, Woolfson MW (2009) ACORN2: new developments of the ACORN concept. *Acta Cryst.* **D65**:881-891.

18. Cowtan K (2010) Recent developments in classical density modification. *Acta Cryst.* **D66**:470-478.

19. Langer G, Cohen SX, Lamzin VS, Perrakis A (2008) Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7. *Nature Protocols* **3**:1171-1179.

20. Foadi J *et al*. (2000) A flexible and efficient procedure for the solution and phase refinement of protein structures. *Acta Cryst*. **D56**:1137-1147.

21. Thorn A, Sheldrick GM (2013) Extending molecular-replacement solutions with SHELXE. *Acta Cryst.* **D69**:2251-2256..

22. Wolfram Research (2015) *Mathematica* v.10. Wolfram Research, Champaign, Illinois, USA.

23. Spitzer M, Wildenhain J, Rappsilber J, Tyers M (2014) BoxPlotR: a web tool for generation of box plots. *Nature Methods* **11**:121-122.

24. Adams PD *et al.* (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Cryst.* **D66**:213-221.

25. Winn MD *et al*. (2011) Overview of the CCP4 suite and current developments. *Acta Cryst*. **D67**:235-242.

Author contributions. The theory and computer code were developed by A.J.M., R.D.O. and R.J.R., database MR calculations and analyses were carried out by R.D.O., and A.G.W. carried out single-atom MR calculations on known structures. J.R.M.O. and K.T. cloned, expressed and crystallised the Shisa3 N-terminal domain, and B.L. solved its structure. The manuscript was written by A.J.M. and R.J.R., with all authors contributing to revisions.

The authors declare no conflict of interest.

Availability. Coordinates and data for the Shisa3 N-terminal domain structure have been deposited with the PDB under accession number 5m0w. Computer code for the program Phaser, including the eLLG calculations, is available as open source within the Phenix (24) and CCP4 (25) packages.

Table 1. Progress of single-atom MR.

	Aldose reductase (3bcj)				Shisa3 (5m0w)			
Atom number	*LLGI*	*TFZ*	Atom type	ΔB (Å^2)	*LLGI*	*TFZ*	Atom type	ΔB (Å^2)
1	22	4.2	S	-0.9	19	6.1	S	-1.5
2	67	8.8	S	-0.4	57	8.3	S	-1.0
3	154	12.7	P	-0.7	80	6.1	S	-1.8
4	243	12.7	P	-0.2	122	8.3	S	-1.6
5	346	13.3	S	0.3	161	8.3	S	-0.1
6	463	14.4	S	0.1	221	10.1	S	-0.4
7	613	16.5	S	-0.2	297	11.3	S	-1.4
8	691	12.0	P	1.2	–	–		–
9	829	15.7	S	0.1	–	–		–
10	908	11.8	S	1.3	–	–		–

LLGI = log-likelihood-gain on intensities, *TFZ* = translation function Z-score, ΔB = refined difference from overall average B-factor. Note that the searches become more unambiguous as more well-ordered S or P atoms are placed because, for equal atoms, the total *LLGI* should be proportional to the square of the number of atoms placed.

Figure Legends

Figure 1. Confidence in MR solution as function of final *LLGI* score. The final refined *LLGI* score provides a clear diagnostic for success in MR. The 3 curves show how the success rate for placing the first copy by MR varies with *LLGI* in 3 different space group symmetry classes: P1 (only 3 rotational degrees of freedom; red; total of 263 MR trials), polar (3 rotational and 2 translational degrees of freedom, with an arbitrary origin along one axis; blue; 4738 MR trials) and non-polar (3 rotational and 3 translational degrees of freedom; black; 16,740 MR trials).

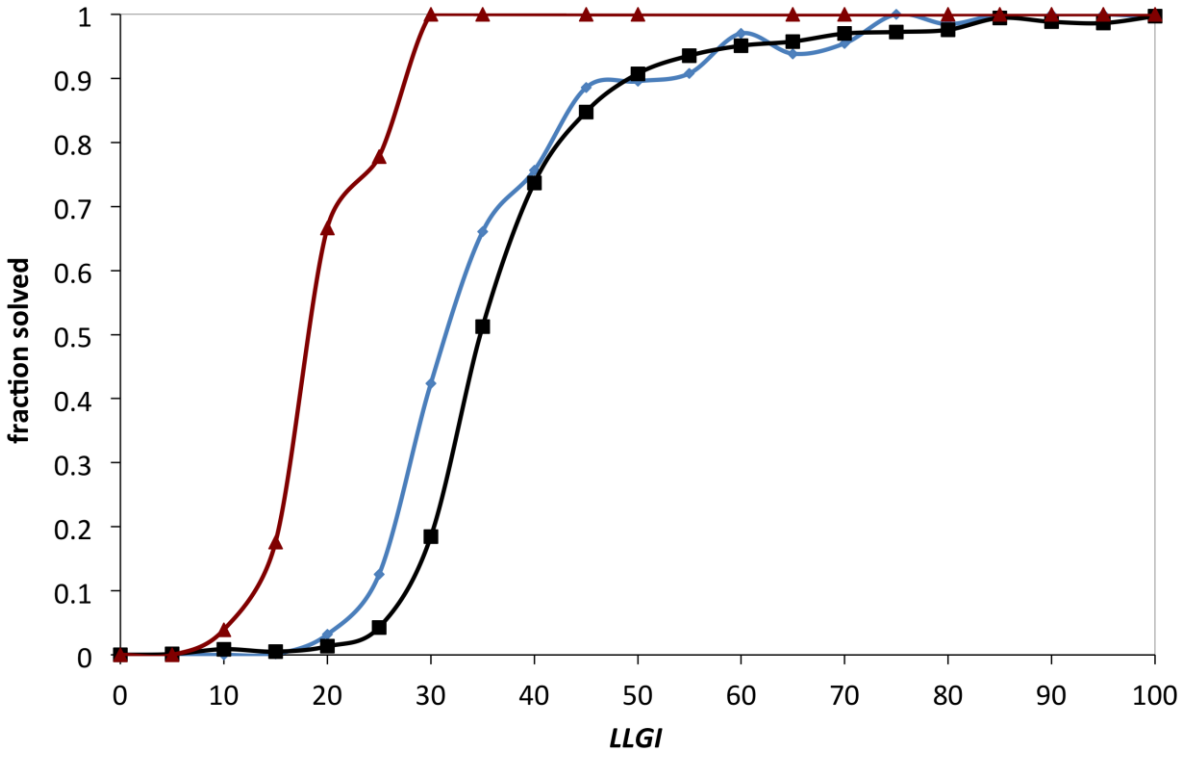

Figure 2. Increase in *eLLG* with resolution and number of atoms. The 3 curves show how the *eLLG* increases with the number of atoms placed (1 atom: blue curve, 2 atoms: orange curve, 3 atoms: green curve) and with increasing numbers of reflections to higher resolution. The calculations are based on the aldose reductase test case (3bcj), for which the data extend to 0.78 Å resolution and the heaviest atoms are sulphurs. It is assumed that B-factors for the best-ordered S atoms will be lower than the mean for the whole structure; by choosing a B-factor reduced by just 1.3 $Å^2$ from the mean, the actual *LLGI* values obtained from placing single S atoms (Table 1) can be reproduced fairly well. The *eLLG* values rise rapidly with resolution, as the number of observed reflections increases and the relative scattering power of the S atoms increases.

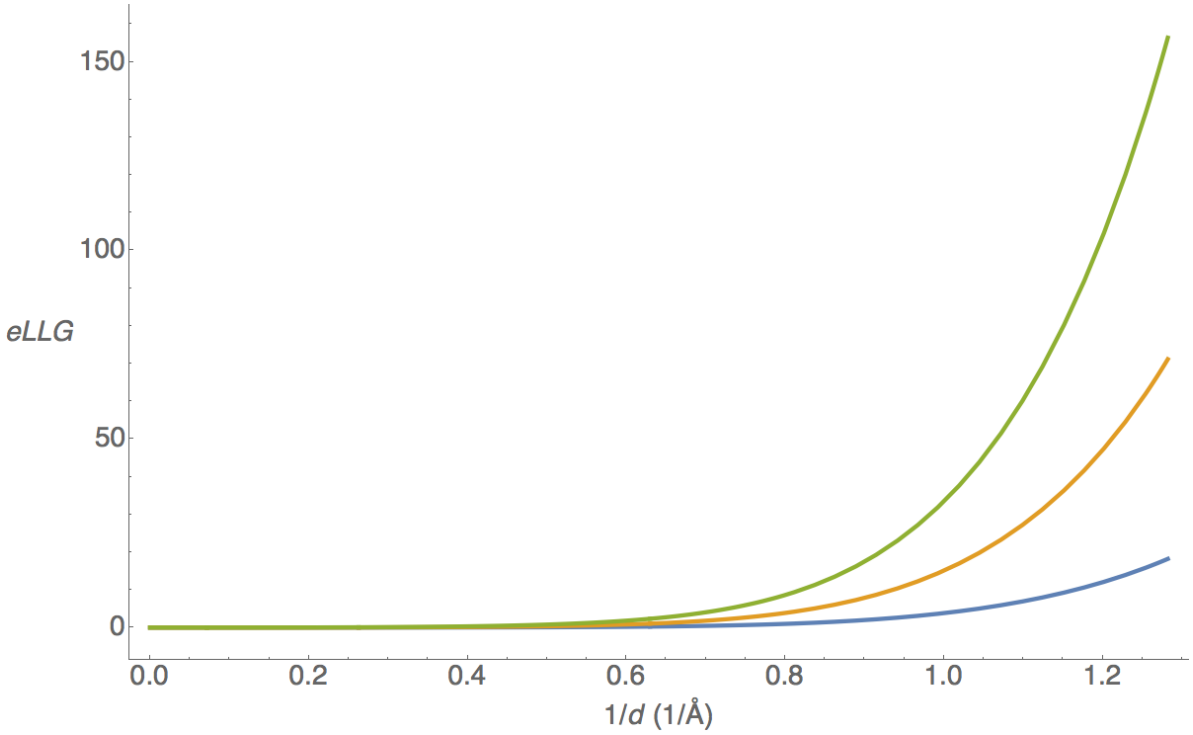

Figure 3. Single-atom model and electron density for aldose reductase. Two single sulphur atoms were placed by MR, then nitrogen atoms were positioned using the log-likelihood-gradient completion algorithm. Atoms forming the sequence Tyr-Pro-Phe and its environment are shown as grey spheres, and the electron density map phased with the atomic model is shown in magenta, contoured at 2.3 times the rms electron density. Refined occupancies allow the nitrogen atoms to serve as surrogates for all atom types.

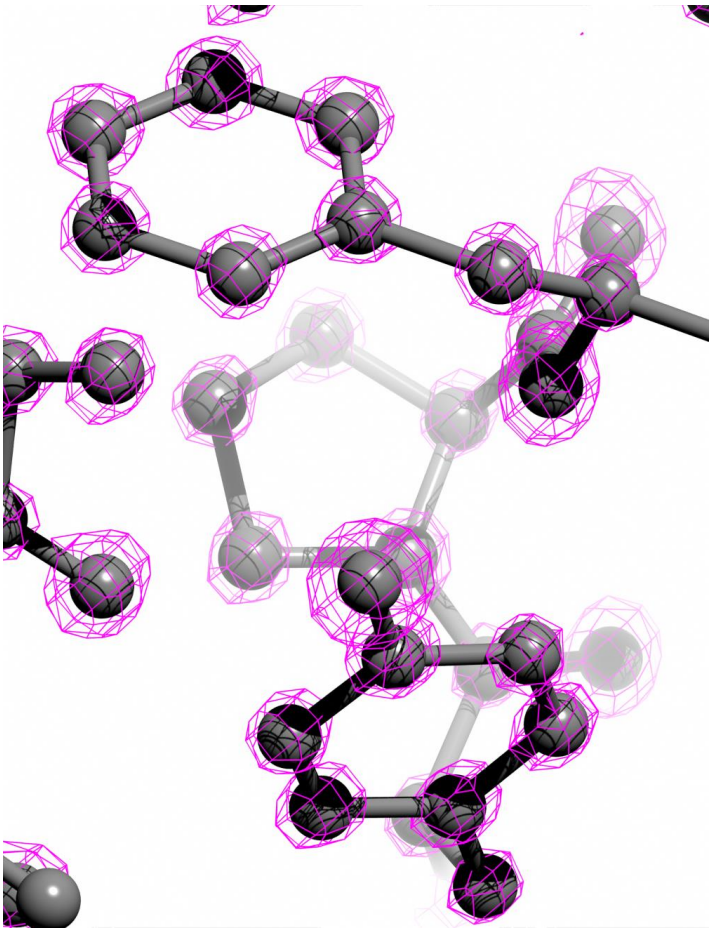

Figure 4. Two-sulphur model and phase-extended density for Shisa3. The two sulphur atoms shown as spheres were placed individually by MR, then the program ACORN was used to refine the phase information, giving the map shown in magenta lines, contoured at 0.6 times the rms electron density.

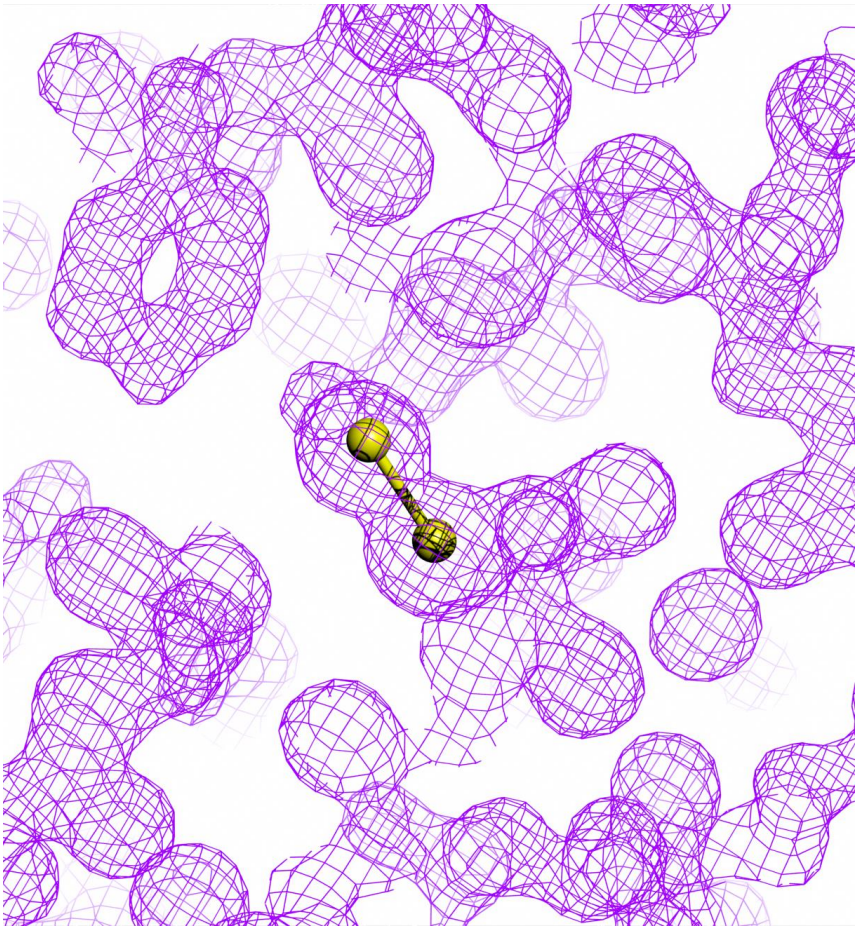

Figure S1. Series approximation of *eLLG*. The blue curve shows the contribution of a single reflection to the total *LLGI*, $\langle LLGI \rangle_{hkl}$, evaluated by numerical integration, compared to the fourth order series approximation ($D_{obs}^4 \sigma_A^4/2$) in orange, as a function of the combined measure of data and model quality, $D_{obs}\sigma_A$.

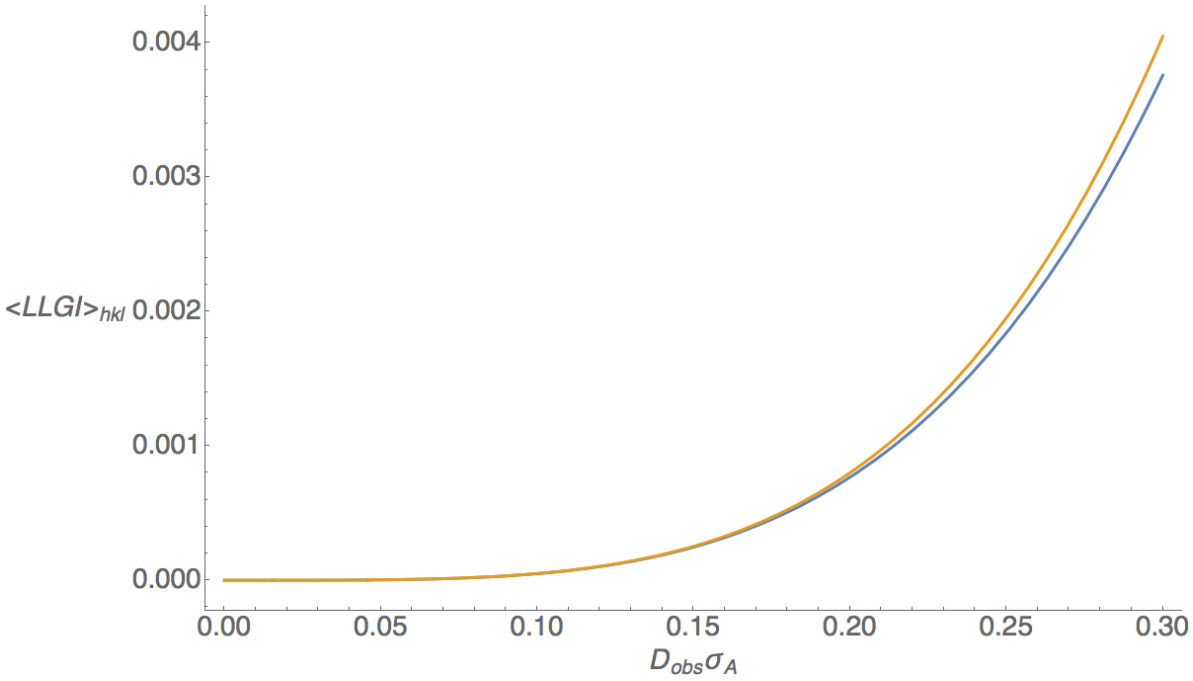

Figure S2. Box plot comparing *LLGI*/atom with success and failure in single-atom MR.

The box plot presents the distributions of *LLGI* per atom for the successful and unsuccessful single-atom MR trials tabulated in Table S1.

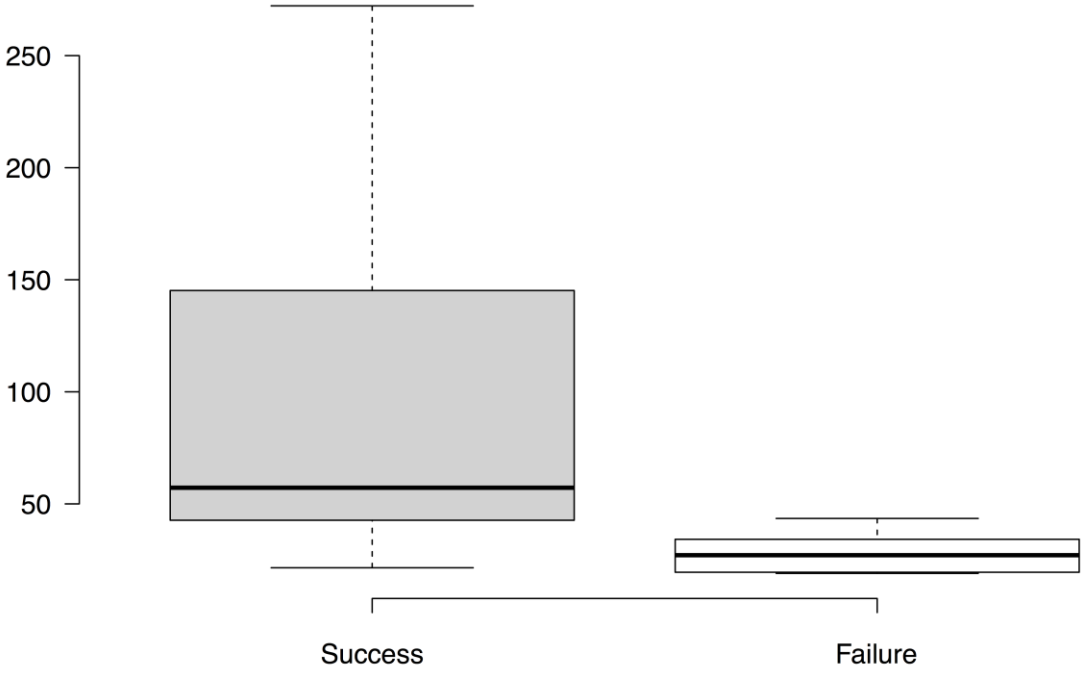

Table S1. Tests of single-atom MR protocol. Test cases are sorted in order of resolution. Cases in which the single-atom MR protocol failed to correctly place the heaviest atoms are highlighted in italics.

PDB code	Name	Res.* (Å)	Space group	# of data	# non-H atoms	Heavy atoms	Single-atom *LLGI*	*LLGI* per atom†	TFZ (last)	# N added	Final *LLGI*	R-factor	Map CC
1ejg	crambin	0.54	$P2_1$	112233	327	6S	1089	272.2	35.2	487	163053	0.096	0.944
2vb1	lysozyme	0.65	P1	187165	1545	10S	582	145.5	22.8	1305	290124	0.11	0.947
2wfi	cyclophilin G	0.75	$P2_12_12_1$	206114	1920	10S	388	97.0	16.7	1767	238296	0.131	0.969
1pq7	*trypsin*	*0.80*	*P1*	*165919*	*2158*	*7S*	*76*	*19.0*	*6.5*	*0*	*65*	*0.831*	*0.062*
2h5c	α-lytic protease	0.82	$P3_221$	189474	2286	8S	86	21.5	5.9	1665	213475	0.141	0.971
5hbs	RBP	0.89	$P2_12_12_1$	104105	1580	8S	176	44.0	10.7	1655	105359	0.147	0.963
1iee	lysozyme	0.94	$P4_32_12$	72347	1490	10S	212	53.0	12.5	1300	63911	0.173	0.924
3o5p	*FKBP51*	*0.95*	*$P2_12_12_1$*	*65672*	*1449*	*4S*	*137*	*34.2*	*5.6*	*0*	*124*	*0.813*	*0.010*
5d99	RNA hairpin	0.97	$P4_3$	38820	859	26P	102	25.5	9.5	749	37322	0.156	0.947
1k6u	BPTI	1.00	$P4_32_12$	32278	673	6S	136	34.0	10.9	552	25144	0.196	0.936
1exr	calmodulin	1.00	P1	77150	1650	4Ca, 8S	255	63.8	11	1164	48722	0.212	0.871
1a6m	myoglobin	1.00	$P2_1$	65676	1445	1Fe, 2S	128	42.7	7.2	1517	57045	0.17	0.921
5d14	*IL-8*	*1.00*	*$P3_121$*	*45396*	*761*	*4S*	*78*	*19.5*	*4.5*	*2*	*110*	*0.789*	*0.005*
4dp7	plastocyanin	1.08	$P2_12_12_1$	33774	1045	1Cu, 3S	229	57.2	8.4	2	291	0.711‡	0.326
4qmc	*PLA2*	*1.09*	*$P4_3$*	*52671*	*1250*	*16S*	*101*	*25.2*	*6.4*	*4*	*171*	*0.752*	*0.029*
1ctj	cytochrome C6	1.10	R3	32653	918	1Fe, 3S	581	145.2	17.3	675	17427	0.227	0.844
5f6e	*Ubc9*	*1.12*	*$P2_1$*	*67878*	*1865*	*7S*	*116*	*29.0*	*6.2*	*5*	*215*	*0.763*	*0.051*
5e0g	*3CL protease*	*1.20*	*C2*	*44825*	*1490*	*11S*	*174*	*43.5*	*7*	*1*	*198*	*0.775*	*0.025*
3po0	SAMP1	1.55	$P2_12_12_1$	11674	736	1Cd, 1S	146	146.0	10	21	687	0.555¶	0.508

* Resolution

† Four atoms were placed in all test cases except 1a6m (3) and 3po0 (1)

‡ R-factor = 0.275 after phase improvement with ACORN and model-building with ARP/wARP

¶ R-factor = 0.298 after phase improvement with ACORN and model-building with ARP/wARP