Maximum-Likelihood Refinement of Incomplete Models with BUSTER + TNT.

G. Bricogne
MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, England,
and LURE, Bâtiment 209D, Université Paris-Sud, 91405 Orsay, France.
gb10@mrc-lmb.cam.ac.uk
http://Gerard2.mrc-lmb.cam.ac.uk

J. J. Irwin
MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, England.
ji10@mrc-lmb.cam.ac.uk
http://Gerard2.mrc-lmb.cam.ac.uk

Abstract

The Bayesian viewpoint had long suggested that structure refinement should be performed by the Maximum-Likelihood (ML) rather than the Least-Squares (LS) method.. ML refinement has been implemented by the combined use of the BUSTER and TNT programs and has been tested on a severely incomplete and imperfect model of crambin. Comparison with LS refinement in the same conditions shows that the ML results are more accurate, and that the log-likelihood gradient map gives clear indications as to the location of the missing atoms. Maximum-entropy techniques implemented in BUSTER are then able to bring out an even clearer picture of how the partial structure should be completed. Much fundamental statistical work remains to be done to allow the construction of likelihood functions better adapted to the type of errors present in macromolecular models at medium resolution.

0. Introduction.

The Bayesian viewpoint has long suggested that structure refinement should be carried out by maximising the log-likelihood gain LLG rather than by minimising the conventional least-squares residual [1,2,3] : only the maximum-likelihood (ML) method can take into account the uncertainty of the phases associated to model incompleteness and imperfection by suitably down-weighting the corresponding amplitude constraints. It was predicted [3,4] that ML refinement would allow the refinement of an incomplete model by using the structure factor statistics of randomly distributed scatterers to represent the effects of the missing atoms, in such a way that the latter would not be wiped out; and that the final LLG gradient map would then provide indications about the location of these missing atoms. As will be shown below, these predictions have now been confirmed by actual tests.

This contribution ends with a discussion of the two main concerns at the moment in the fields of structure refinement and validation where Bayesian methods have much to offer, namely (1) getting better reliability indicators for the final results of structure refinement, i.e. the design of new likelihood functions better suited to the nature of errors in macromolecular models at medium resolution; and (2) ensuring that these indicators are effectively optimised during refinement.

1. Least-squares: ills and current remedies.

In small-molecule studies, where the data to parameter ratio is huge, the error-covariance matrix gives a wealth of accuracy estimates, which can be cast into more readable form (e.g. TLS analysis of thermal parameters). The Luzzati error model and plot [5] can also be used to estimate final positional accuracy.

With macromolecules, however, the data to parameter ratio is never huge, even with restraints. In these circumstances least-squares (LS) structure refinement can produce overfitting artefacts by moving faster towards agreement with moduli than towards correctness of the phases, because its shift directions assume the current model phases to be error-free constants. R-factors and Luzzati plots then become misleading. Furthermore, when the model is very incomplete, density for the missing part tends to disappear rather than improve during LS refinement.

The current remedies rely on cross-validation (CV) [6] as a powerful device for detecting the onset of overfitting. It is based on the simple notion that overfitting amounts to fitting "noise" rather than “signal” in the data, which causes a loss of predictive power towards data not used in the fit. It must be borne in mind, however, that using CV in this way as a stopping criterion in a LS refinement only guarantees optimality *along the least-squares path*: it does not guarantee that the solution reached is optimal in a global sense. Assessing the accuracy of the results in the absence of an error-covariance matrix is not straightforward; the safest method available at present for estimating r.m.s. coordinate error seems to be a Luzzati plot from cross-validated σ_A values .[7]

2. Seeking a more radical cure.

Improvements on the current state of the art (least-squares refinement with cross-validation by R^{free}) seem desirable in two related directions, in both of which the fundamental techniques of structure factor statistics occupy a central position.

Firstly, since the LS path is deflected towards a premature fit to the moduli by excessive confidence in the current phases, it is natural to think of a feed-back mechanism whereby the current estimate of model error would be converted into a representation of the uncertainty on the phases, so that the latter could be used with more caution. Exercising this caution, however, necessarily involves altering weights (variances), which is not allowed within the least-squares method: the latter must therefore be abandoned in favour of the maximum likelihood (ML) method (see a similar argument about the treatment of non-isomorphism in §2.4.1 of [8]).

Secondly, since in the ML method the model now parametrises its own uncertainty, the question arises of choosing an adequate error model. It will be argued that the Luzzati error model is not suited to the heavily-restrained macromolecular setting, and that a new class of statistical models is required.

3. Maximum Likelihood vs. Least-Squares.

ML refinement offers an attractive generalisation over LS [1,2,3,4] by allowing the refinement of parameters which modulate the variances of the model structure factors: the latter are no longer handled as *values* but as *probability distributions*, in which variances and covariances can represent both model imperfection and model incompleteness. According to the standard protocol outlined in [1] the probability distributions for model structure factors are integrated over the phase to yield predicted distributions of model amplitudes; substituting the observed values of these amplitudes then yields the likelihood Λ of the model. All parameters can then be refined by maximisation of L or of $L = \log \Lambda$. The error covariance matrix is the final Hessian of L, if it can be calculated. It should be recalled that ML estimation is only an approximation to Bayesian estimation, and that the full force of the latter should be invoked whenever the maximum of L is not so pronounced as to dominate over prior probability in the application of Bayes's theorem.

4. ML refinement using BUSTER and TNT.

To ascertain the impact of taking phase uncertainty into account on the path followed during structure refinement, we have used BUSTER [4] and TNT [9] on a test data set for crambin [10] suffering from both model imperfection and model incompleteness, and compared the results of LS and ML refinements from these data.

Model incompleteness resulted from taking only residues 1-27 (60% of the atoms) as the fragment to be refined; the remaining 40% (residues 28-46) was modelled through a non-uniform distribution for the missing atoms, defined by a mask for that region which had been extensively smoothed then blurred by a B-factor of 250. The expectation values and variances for the structure factor contributions from this pool of random atoms were calculated within BUSTER according to the equations in §2.1.0 of [4].

Model imperfection was introduced by heating fragment 1-27 to 1000°K then regularising it, using XPLOR [11], thereby creating positional errors with an r.m.s. value of about 1.0Å. This imperfection was treated statistically through a Luzzati model parametrised by a refinable "imperfection B factor" B^{impf}, similar to the quantity B^{glo} used in the parametrisation of non-isomorphism in heavy-atom derivatives (see §2.4.1 of [7]). This B^{impf} intervenes in the calculation of expectation values $\left\langle F^{impf}(\mathbf{h}) \right\rangle$ and variance parameters $\sigma_2^{impf}(\mathbf{h})$ for the structure factor contributions from the imperfect fragment according to:

$$\left\langle F^{impf}(\mathbf{h}) \right\rangle = D(\mathbf{h}) \times F^{frag}(\mathbf{h}) \qquad (1)$$

$$\sigma_2^{impf}(\mathbf{h}) = \left(1 - D(\mathbf{h})^2\right) \times \left\langle \left|F^{frag}(\mathbf{h})\right|^2 \right\rangle_{d^*_{\mathbf{h}}} \qquad (2)$$

where

$$D(\mathbf{h}) = \exp\left[-\frac{1}{4} B^{impf} (d^*_{\mathbf{h}})^2 \right] \qquad (3).$$

The expectation values for the imperfect fragment and random atoms contributions, and the variances or covariances caused by imperfection and incompleteness, are added and used as arguments of elliptic Rice likelihood functions [12], in combination with any experimental phase information which may be available.

Refinement was carried out against 1.5Å synthetic data calculated from the correct whole crambin structure, without solvent, with 3% r.m.s. noise added. The reference LS refinement was performed using TNT in the conventional way. The ML refinement proceeded as follows. At each cycle BUSTER refined the values of overall scale and B factors and of B^{impf} by maximum-likelihood, and calculated the value, gradient and Hessian of the log-likelihood gain L with respect to the quantities

$F^{\mathrm{frag}}(\mathbf{h})$. This “osculating LS” approximation to L was passed on to TNT where is was used to generate parameter gradients (AGARWAL command) and curvatures, and to carry out one cycle of positional refinement on the fragment structure.

In these conditions ML refinement clearly outperformed LS refinement, giving a mean-square distance to the correct positions of 0.176 (ML) instead of 0.415 (LS). Examination of histograms of positional errors showed that, apart from a small number of outliers corresponding to model atoms near the boundary with the missing region, *the ML fit is much tighter than the LS fit*.

Visualisation of the time course of the refinement showed, as anticipated, that not only the end point but *the entire path of the refinement is altered* by switching from LS to ML. This may be understood by noting that the contribution to structure factor variances from model imperfection, given by eq. (2) above, increases sharply with resolution, so that high-resolution contributions to the gradient maps are filtered out in the early stages then gradually switched on as refinement proceeds. This feature leads to considerable *increases in the radius of convergence* of the refinement.

Furthermore the ML method produced a final LLG gradient map displaying highly significant, correct connected features for the missing part (40%) of the molecule, while the final LS difference map showed no such features (see Figs. 1 and 2). This *enhances the possibilities of bootstrapping* from an otherwise unpromising molecular replacement starting point to a complete structure. Essentially the same behaviour was observed at 2.0Å resolution, and with experimental rather than calculated data.

Other prototypes for ML structure refinement have been built and tested by Read [13] (using XPLOR and an intensity-based LLG) and by Morshudov [14] (using PROLSQ [15] and the Rice LLG). The BUSTER+TNT prototype has the advantage of being able to use external phase information by means of the elliptic Rice function [12], as well as prior information about non-uniformity in the distribution of the missing atoms in incomplete models. It also allows the ML refinement of an incomplete model to be carried out in conjunction with phase permutation for those strong amplitudes which are most poorly phased by that model, i.e. have the largest renormalised $|E|$'s ; or in conjunction with maximum-entropy updating of the distribution of random atoms, initially taken as essentially featureless within the given envelope. Using the method of joint quadratic models of entropy and LLG described in [22] before and after refinement of the incomplete model produced the updated distributions shown in Figs. 3 and 4, demonstrating clearly the advantage of carrying out ML refinement within the integrated statistical framework provided by BUSTER. This “after-burner” establishes a seamless continuity between the middle game of structure determination and the end game of structure refinement.

5. Limits of the Luzzati model.

In the test calculations reported above, examination of partially converged models during or after refinement at lower resolution leads to the obvious conclusion that *questions of accuracy concerning the results of macromolecular refinement at medium resolution are fundamentally different from the same questions posed and studied for small molecules at high or very high resolution*. In the latter case it is reasonable to treat the model errors on the positions of different atoms as statistically independent and thus to use Luzzati's treatment for the errors they induce on the structure factors. Macromolecular refinement, on the other hand, is so heavily restrained that the model positional errors at any stage are highly correlated. This affects such crucial quantities as the effective number of degrees of freedom in the error statistics, and the magnitude of the uncertainty along each of these degrees of freedom. The Luzzati model is then inappropriate as a means of relating positional error statistics to structure factor statistics, and hence as a means of constructing a good likelihood function for ML refinement.

6. New error models for macromolecular structures.

In a macromolecular refinement, model positional errors will be correlated through “regular perturbations” of a restrained macromolecular structure, i.e. perturbations compatible with the restraints which propagate positional errors between atoms or groups of atoms. New error models are required for deriving the structure factor statistics associated to random regular perturbations.

This may be illustrated by a simple physical analogy, for the physical aspects of which the reader is referred to [16]. The assumption of statistically independent random perturbations of atomic positions underlies not only the Luzzati model in structure factor statistics, but also the Einstein model of thermal motion in crystals and the Debye model of thermal effects on scattering. What is now needed in the field of structure factor statistics is the equivalent of the Born & von Kármán lattice-dynamical model of thermal motion, and of the use of these lattice normal modes in the parametrisation of anisotropic B factors and of thermal diffuse scattering.

An attractive possibility – if computer limitations can be overcome – would be to use the softest lattice ‘normal modes’ with wave vector $\mathbf{q}=\mathbf{0}$ from the Hessian matrix of the restraint function and parametrise the joint positional uncertainty model in terms of the variances of normal coordinates along these modes. This correlated positional error model could then be converted into a parametrised joint probability distribution of complex structure factors, then of amplitudes, which would yield the best likelihood function for refining both the structural model parameters and the mean-square normal coordinates describing the errors. At the end of the refinement, this error model would embody the description of the accuracy of the refinement results.

7. Maximum-likelihood refinement for non-macromolecular problems.

The two main sources of bias in macromolecular LS refinement results, namely the low observation-to-parameter ratio and the inadequate treatment of phase uncertainty, are also present in other fields of crystallography, in particular in Rietveld refinements of powder structures [17] and in multipole refinements of accurate electron densities [18-21]. In the powder case the notion of phase can be generalised to that of a hyperphase,[2] the loss of hyperphase information comprising both that which results from the overlap of different Bragg reflexions and from the ordinary loss of phase for these Bragg reflexions. In this instance, hyperphase-mediated bias is even more pernicious than the phase-mediated bias considered above and is the likely cause of numerous recently diagnosed pathologies in test Rietveld LS refinements. The probability distributions and likelihood functions for powder data derived in [2] will enable the incorporation of hyperphase uncertainty into the refinement and yield a maximum-likelihood Rietveld method which can be expected to cure the observed biases of the current LS method.

8. Validation and error models.

The use of cross-validation in the choice of refinable model parameters and in the validation of refinement results [7] has so far been based on the conventional crystallographic R-factor, which is not a particularly optimal criterion from the statistical point of view. In particular, concern has arisen about possible dangers of its use in the presence of non-crystallographic symmetries, since data belonging to the test set may happen to be strongly correlated to data which are being fitted, thus creating misleadingly low values for the free R-factor. The problem is clearly that the R-factor definition makes no reference to any predictable variability in statistical dispersion from one data item to another, nor to expected patterns of correlation in this dispersion.

The Bayesian viewpoint gives an unequivocal answer to this dilemma. Retaining the idea of cross-validation as a measure of the predictive power of a statistical model towards yet unseen data (already present in the scheme proposed in §8.1 of [22]) it leads naturally to suggesting that the free R-factor be replaced by the *free log-likelihood gain* L^{free} calculated over the same test data set. This viewpoint is none other than that formulated in [1] and [4] and does require that the predictions from the fit of the actively used data be couched in terms of a conditional probability distribution for the test data, from which the free LLG (e.g. from the model at the preceding cycle) can be calculated by the standard procedure.

Since the strong correlations between amplitudes created by non-crystallographic symmetries can be taken into account in the calculation of likelihoods [3], the use of L^{free} should be immune to the problems encountered by R^{free} in this case. In less problematic cases L^{free} can still be expected to perform better, in view of the Neyman-Pearson optimality property [1], provided the likelihood functions used are capable of correctly representing the state of knowledge (or uncertainty) prevailing at each stage. In the refinement context this adds to the urgency of the developments outlined in §7.

9. Conclusion.

It has been shown that maximum-likelihood structure refinement, long advocated by the first author, is greatly superior to conventional least-squares refinement by virtue of its ability to deal correctly with the phase uncertainties introduced by model imperfection and incompleteness. The end results are more accurate, the radius of convergence is increased, and the final log-likelihood gradient map gives useful indications as to the location of missing atoms.

The increase in radius of convergence may rapidly overturn the present reliance on simulated annealing [11] as a means of getting out of local least-squares minima: the automatic “blurring” of the LLG gradient maps in the early stages of the refinement will largely suppress such spurious minima. It is thus conceivable that simulated annealing might be dispensed with altogether in the future, any possible bifurcation being handled through phase permutation techniques.

The optimal performance of ML refinement will depend crucially on the design and implementation of better statistical error models in real space as the basis for obtaining better likelihood functions in structure factor space. Much remains to be done in this area, as well as in making better use of off-diagonal interactions during the likelihood-maximisation process itself.

Acknowledgements.

We wish to thank Dale Tronrud and Lynn Ten Eyck for advice on the use of TNT and for making minor changes at our request. We are grateful to Angela Loh and to

Digital Equipment Corp. for the loan of generous amounts of state-of-the-art computer equipment.

This research was supported in part by a Tage Erlander Guest Professorship from the Swedish Natural Sciences Research Council (NFR) in 1992-93, and by an International Research Scholars award from the Howard Hughes Medical Institute since 1993.

References.

[1] G. Bricogne, *Acta Cryst.* A**44**, 517-545 (1988).
[2] G. Bricogne, *Acta Cryst.* A**47**, 803-829 (1991).
[3] G. Bricogne, In *The Molecular Replacement Method. Proceedings of the CCP4 Study Weekend 31 January – 1st February 1992*, edited by W. Wolf, E.J. Dodson and S. Gover, 62-75, Daresbury Laboratory, Warrington (1992).
[4] G. Bricogne, *Acta Cryst.* D**49**, 37-60 (1993).
[5] V. Luzzati, *Acta Cryst.* **5**, 802-810 (1952).
[6] A.T. Brünger, *Acta Cryst.* D**49**, 24-36 (1993).
[7] A.T. Brünger, *Methods in Enzymology* **276** (in the press).
[8] E. de La Fortelle and G. Bricogne, *Methods in Enzymology* **276** (in the press).
[9] D.E. Tronrud, L.F. Ten Eyck, and B.W. Matthews, *Acta Cryst.* A**43**, 489-501 (1987).
[10] W.A. Hendrickson and M.M. Teeter, *Nature* **290**, 107-109.
[11] A.T. Brünger, J. Kuriyan, and M. Karplus, *Science* **235**, 458-460 (1987).
[12] G. Bricogne, *Methods in Enzymology* **276** (in the press).
[13] R.J. Read, (1996), To appear in *Macromolecular Refinement*, edited by M. Moore and E.J. Dodson, Daresbury Laboratory, Warrington (1996).
[14] G. Morshudov, To appear in *Macromolecular Refinement*, edited by M. Moore and E.J. Dodson, Daresbury Laboratory, Warrington (1996).
[15] J.H. Konnert and W.A. Hendrickson, *Acta Cryst.* A**36**, 344-349 (1980).
[16] B.T.M. Willis and A.W. Pryor, *Thermal Vibrations in Crystallography*, Cambridge University Press, Cambridge (1975).
[17] H.M. Rietveld, *Acta Cryst.* **22**, 151-152 (1967) ; *J. Appl. Cryst.* **2**, 65-71 (1969).
[18] F.L. Hirshfeld, *Acta Cryst.* B**27**, 769-781 (1971).
[19] N.K. Hansen and P. Coppens, *Acta Cryst.* A**34**, 909-921 (1978).
[20] M.A. Spackman and R.F. Stewart, In *Methods and Applications in Crystallographic Computing*, edited by S.R. Hall and T. Ashida, 302-320, Clarendon Press, Oxford (1984).
[21] B.M. Craven, H.P. Weber, and X.M. He, *The POP Refinement Procedure.* Technical Report, Dept. of Crystallogr., University of Pittsburgh (1987).
[22] G. Bricogne, *Acta Cryst.* A**40**, 410-445 (1984).

Figures.

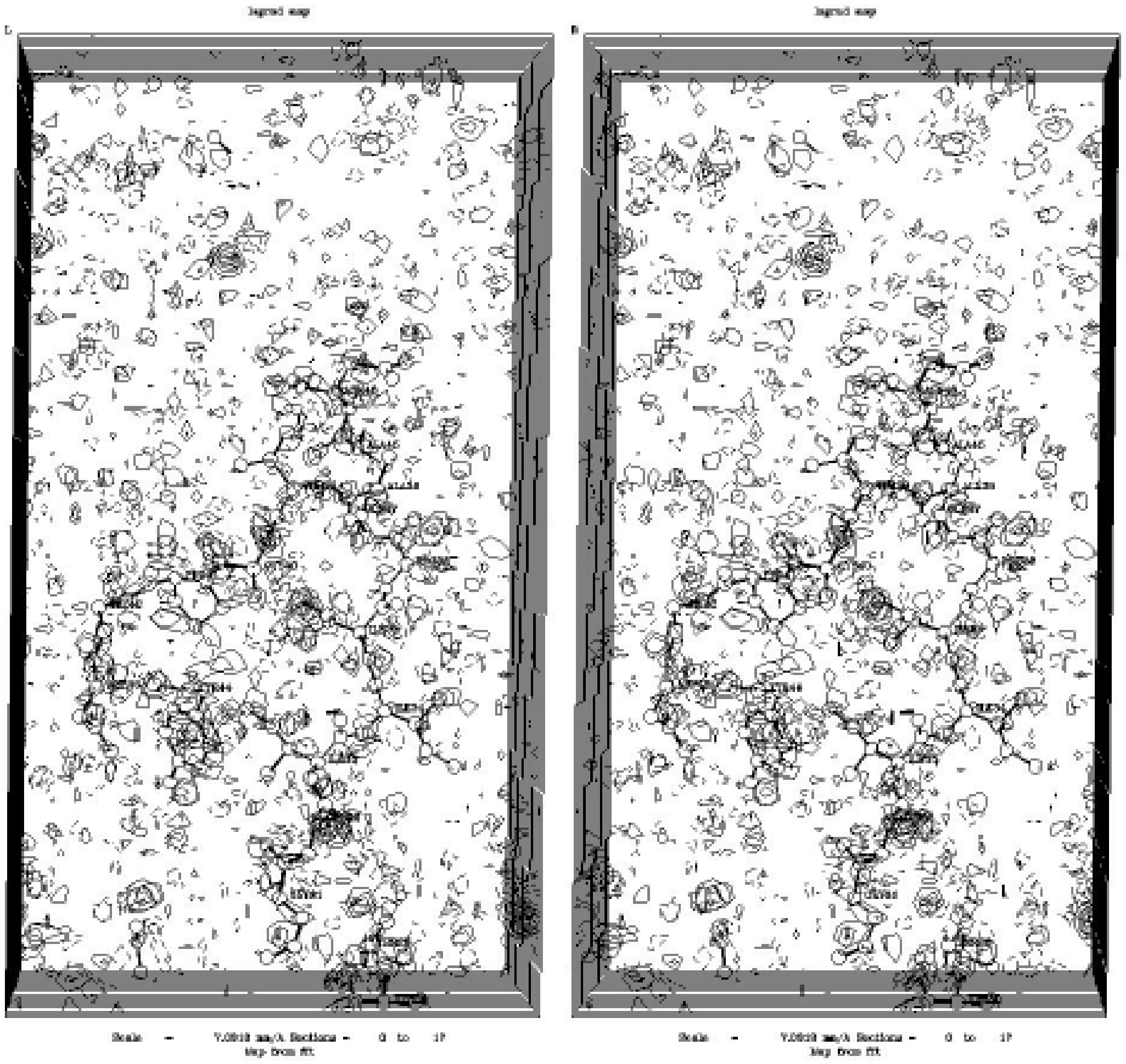
lsgrad map
L
lsgrad map
R
Scale = 7.0910 mm/A Sections = 0 to 17
Map from fft
Scale = 7.0910 mm/A Sections = 0 to 17
Map from fft

Figure 1 The log-likelihood gradient map at the end of LS refinement. The missing structure is drawn for reference. There is very little reliable information to help complete the refined partial structure.

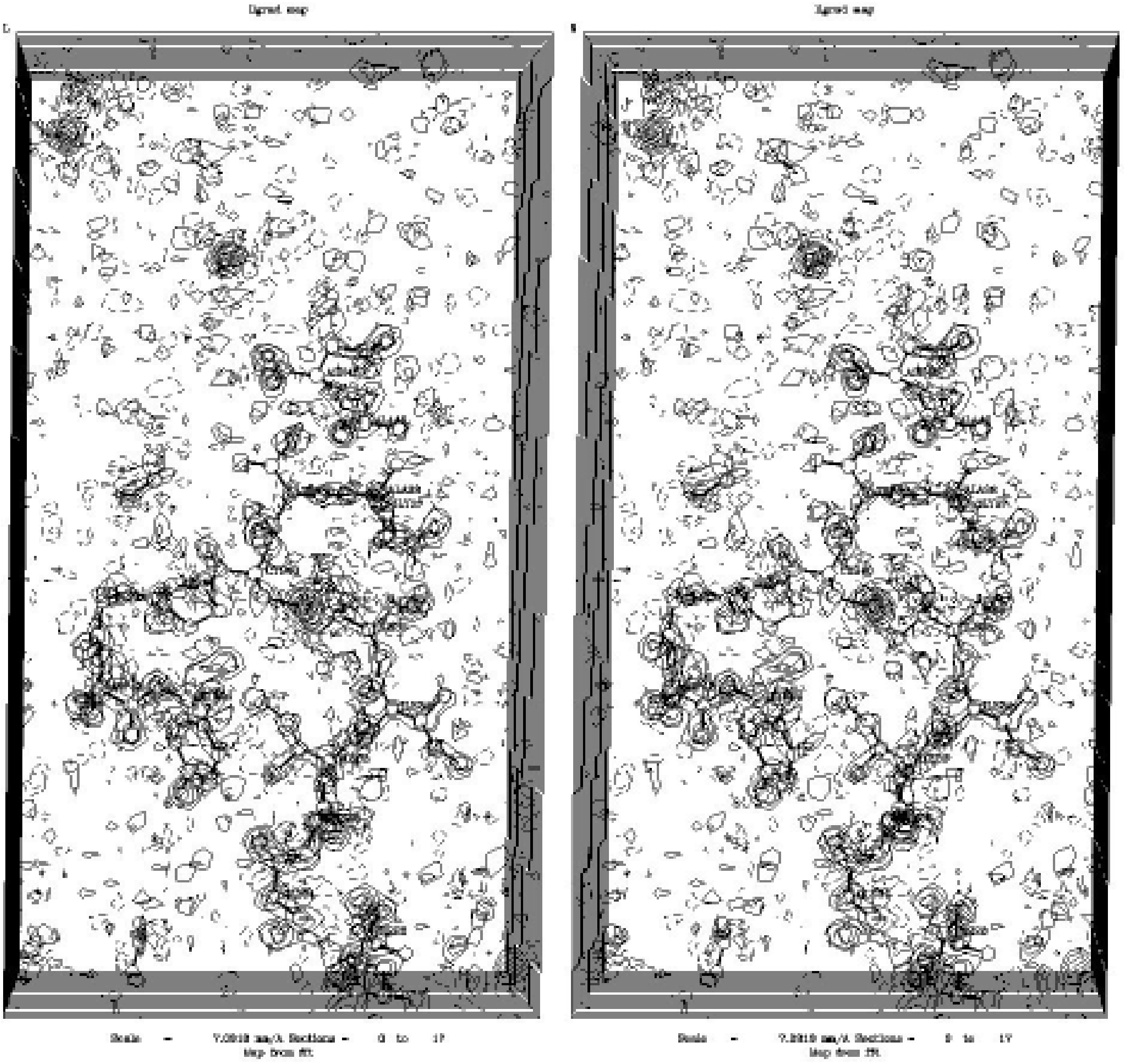
Lgrad map
L
Lgrad map
R
Scale = 7.0910 mm/A Sections = 0 to 17
Map from fft
Scale = 7.0910 mm/A Sections = 0 to 17
Map from fft

Figure 2 The log-likelihood gradient map at the end of ML refinement. There is considerably more reliable information to help complete the partial structure after it has been refined.

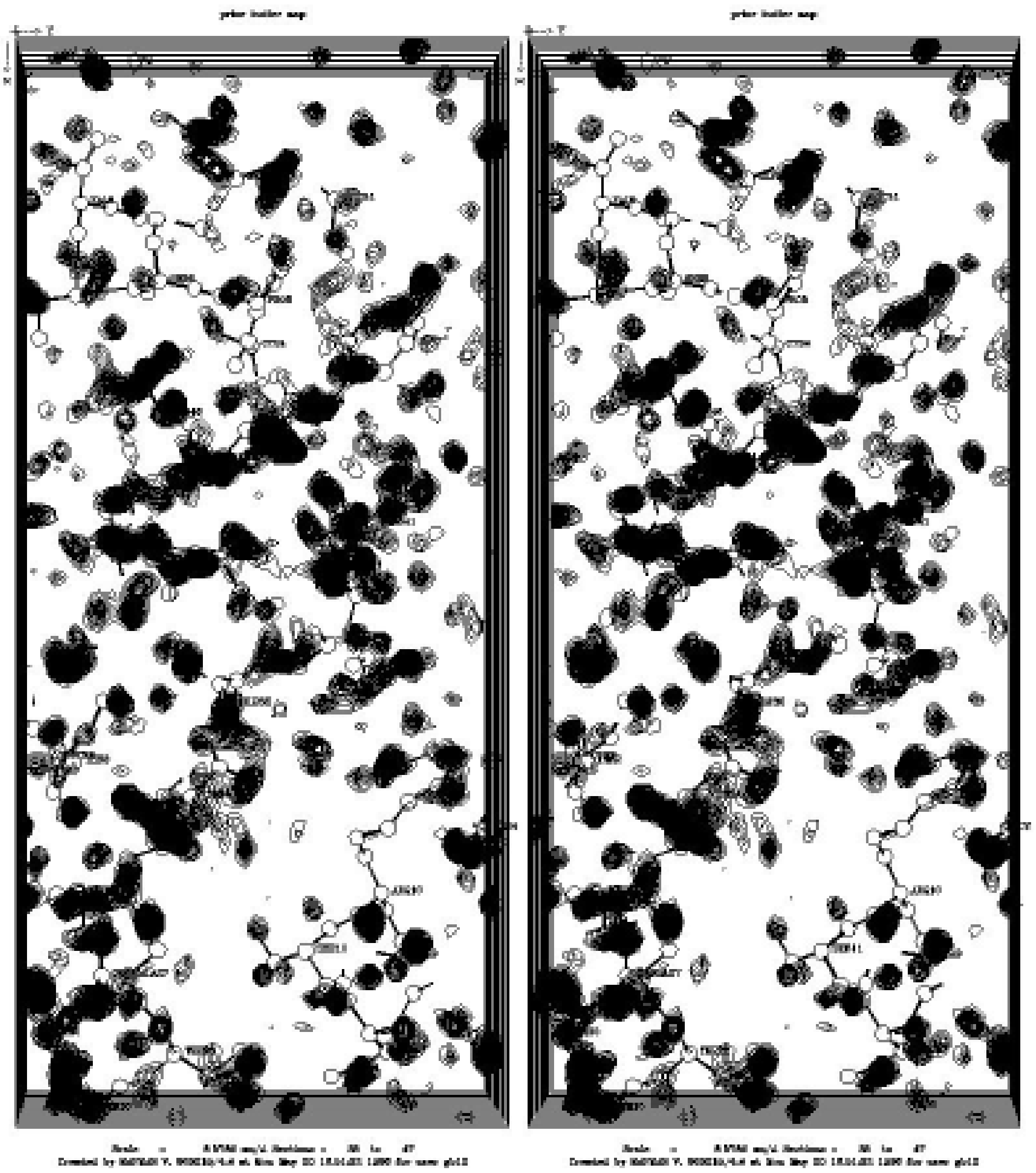

Figure 3 Maximum-entropy update of the distribution of missing atoms before ML refinement of the partial structure.

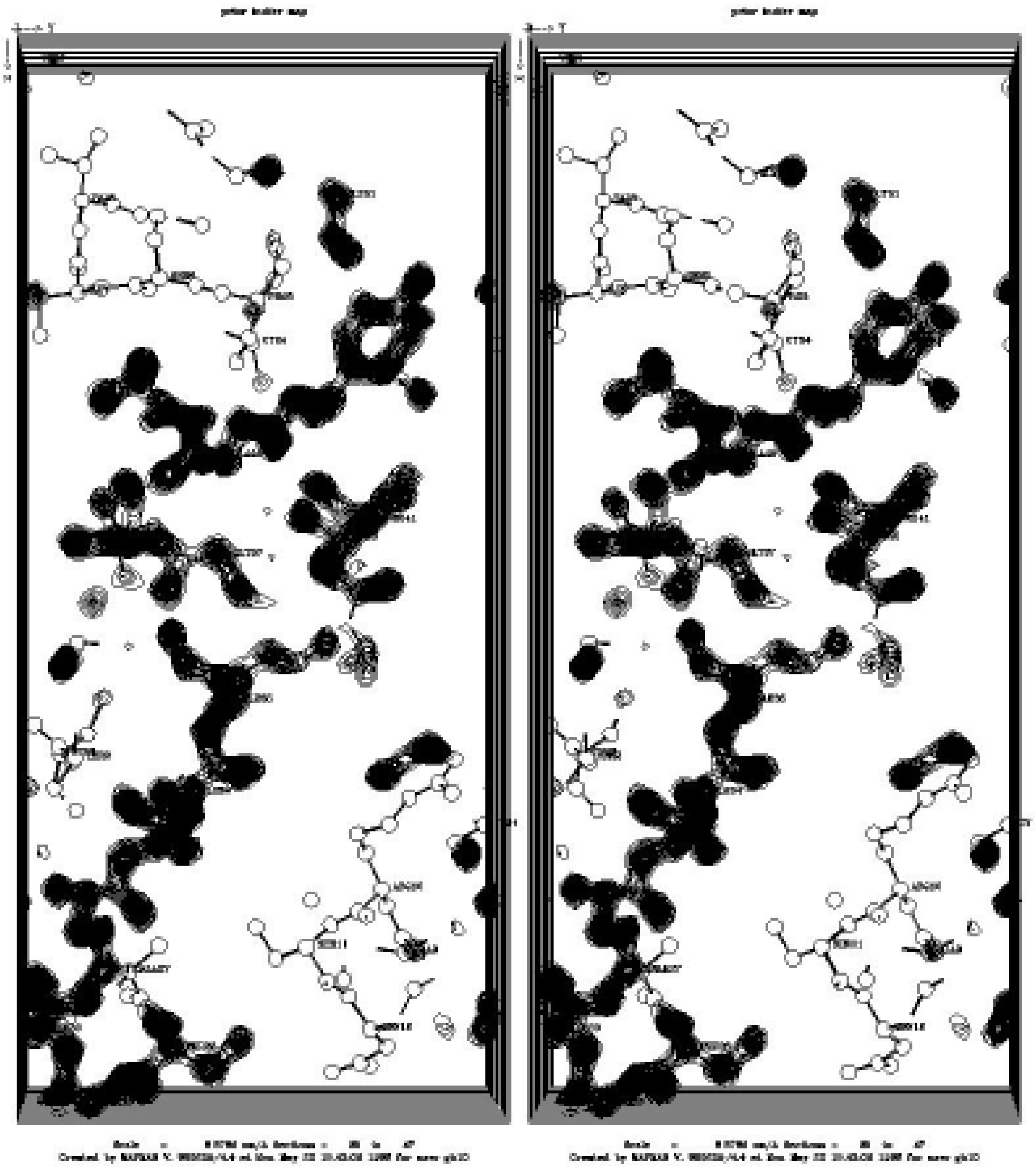

Figure 4 Maximum-entropy update of the distribution of missing atoms after ML refinement of the partial structure.