

Potential use of real-space refinement in protein structure determination

MICHAEL S. CHAPMAN* AND ERIC BLANC at *Department of Chemistry and Institute of Molecular Biophysics, Florida State University, Tallahassee, FL 32306-3015, USA. E-mail: chapman@sb.fsu.edu*

(Received 10 June 1996; accepted 21 October 1996)

Abstract

Through testing refinement protocols using free R -factor estimates of model quality, it is shown that real-space refinement can be a useful addition to conventional reciprocal-space refinement, even for protein structures with poor electron-density maps derived from multiple isomorphous replacement. By alternating real- and reciprocal-space refinements, starting with an experimental map, then calculating $2F_o - F_c$ maps, it is demonstrated with the structure of HMG-CoA reductase, that quick automatic refinement can yield a model with a free R factor 1.5% better than exhaustive reciprocal-space refinement, and within 1% of a model that was interactively rebuilt and refined repeatedly.

1. Introduction

Real-space refinement (Diamond, 1971) ceased to be the refinement method of choice with the development of efficient stereochemically restrained reciprocal-space methods (Hendrickson, 1985). There has been recent renewed interest in real-space methods for the refinement of macromolecular structures with high non-crystallographic symmetry, for example virus capsids (Chapman & Rossmann, 1996). The electron-density maps of such structures can be of exceptional quality following phase refinement through non-crystallographic symmetry (Rossmann *et al.*, 1992). Refinement against the electron density (implicitly including phase information) can yield models that are slightly better (Chapman, 1995) than those refined only against amplitudes, with programs such as *PROLSQ* (Hendrickson, 1985) or *X-PLOR* (Brünger, Kuriyan & Karplus, 1987). However, in protein structure determination, the phase errors and map qualities are generally much poorer. As expected, real-space refinement, alone does not rival the precision of reciprocal-space refinement (see below). However, real-space refinement is already used for the modest goal of facilitating model building using molecular graphics (Jones, Zou, Cowan & Kjeldgaard, 1991). The work reported here explores whether real-space methods, supplementing reciprocal-space methods, are harmful, or improve the refinement of crude initial models. As the effectiveness of real-space refinement has already been demonstrated in cases with high phase quality (Chapman & Rossmann, 1996), this report focuses on the most difficult cases, those with the high phase error of medium-resolution multiple isomorphous replacement.

In restrained real-space refinement (Chapman, 1995), the following residual is minimized,

$$R_p = \sum_{\mathbf{x}} [s\rho_{\text{map}}(\mathbf{x}) + k - \rho_{\text{model}}(\mathbf{x}, P)]^2 + \sum_{\text{bonds, angles...}} w_L [L_{\text{model}}(P) - L_{\text{standard}}(P)]^2,$$

where $\rho(\mathbf{x})$ is the electron density at a grid point, \mathbf{x} (near an

atom), either in the map, or calculated from a model, P is the set of refining atomic parameters (coordinates); L is the set of stereochemical parameters; w are weights and s and k are scaling constants. It is implicitly dependent on phases that are needed to calculate $\rho_{\text{map}}(\mathbf{x})$, differing from reciprocal-space methods in which the first term is replaced by $\sum_{\mathbf{h}} w_{\mathbf{h}} [|F_{\text{obs}}(\mathbf{h})| - s|F_{\text{model}}(\mathbf{h}, P)]^2$, where $|F|$ is the structure amplitude for reflection \mathbf{h} . In reciprocal space, all atomic parameters are interdependent, because each structure amplitude is dependent on all atoms. By contrast, real-space refinement is local: interdependency is limited to stereochemical interactions and overlap of electron density.

Atomic refinement is a non-linear optimization that is only marginally overdetermined, has a finite convergence radius and is subject to overfitting (Brünger, 1992). The outcome depends upon the data-to-parameter ratio and accuracy of the starting model. Starting-model accuracy depends on: (1) map quality, and (2) how well the model has been fitted to the map with interactive computer programs (Jones *et al.*, 1991). One potential benefit of real-space refinement is to reduce the dependence of refinement on starting model, through prior optimization against the electron density. Secondly, implicit inclusion of phases could result in improved conditioning through increased data-to-parameter ratio. At about 3 Å resolution, the number of diffraction data do not greatly exceed the number of refined parameters. Refinement is possible only with added data in the form of stereochemical restraints (Hendrickson, 1985) or reduction of parameters through stereochemical constraints (Sussmann, Holbrook, Church & Kim, 1977). In most applications of reciprocal-space refinement, only amplitudes are used, ignoring the typically inferior experimental phase information. Addition of either explicit or implicit phase restraints has proved beneficial in several test and real refinements (Arnold & Rossmann, 1988; Chapman, 1995; Rees & Lewis, 1983) when the phases are of exceptional quality. Here, even poor phases will be used (implicitly in real-space refinement), maximizing the data-to-parameter ratio when the model is poor.

Previous test uses of phase restraints (Arnold & Rossmann, 1988; Chapman, 1995; Rees & Lewis, 1983) have focused either on simulated data calculated from small molecules or on actual virus structures. In both cases the phases are more precise than likely in a typical protein structure determination. For protein structures, there is concern that the large phase error, derived from isomorphous replacement methods, and consequent poor map quality may inhibit model improvement during real-space refinement, or even lead to deleterious effects. With the development of free R factors, unbiased indicators of model quality (Brünger, 1992), it is now possible to test refinement protocols with the large experimental errors encountered during a real protein structure determination.

The potential of real-space methods was tested through repeated (re-)refinements of the structure of 3-hydroxy-3-

methyl glutaryl (HMG)-CoA reductase (Lawrence, Rodwell & Stauffacher, 1995), a typically difficult structure to determine. It was a particularly suitable test, because unrefined coordinates were available that had been interactively fit into the electron density with the care typical of an actual structure determination. Also, the experimental map was typically poor, as illustrated by its low correlation with electron density calculated from the refined coordinates ($r = 0.76$) compared with $r = 0.88$ for the final $2F_o - F_c$ map (Lawrence *et al.*, 1995). The structure had been determined by multiple isomorphous replacement yielding a combined figure of merit of 0.65. The map was improved by solvent flattening after which it was possible to locate a non-crystallographic diad. Following initial twofold averaging, a polyalanine model was built that was used to generate an improved molecular mask for the final electron-density averaging and solvent flattening. It was into this averaged map that the complete model was originally built. However, many difficult structure determinations do not have the advantage of non-crystallographic symmetry, so the unaveraged map was used for the following tests of real-space refinement.

2. Methods

The success of refinements was evaluated through free R factors [R_T^{free} (Brünger, 1992)] and coordinate deviation from a 'target' structure. The model, as originally reported (Lawrence *et al.*, 1995), had been refined extensively allowing individual temperature factors to vary. A target structure, more appropriate for appraising refinement protocols at the initial stages, was derived from the reported model by resetting the temperature factors to 20.0 \AA^2 then refining the atomic positions in reciprocal space with an additional 30 cycles with fixed temperature factors. Thus, the target structure (X in Table 1) incorporated the improvements of extensive manual rebuilding (Lawrence *et al.*, 1995), was optimized with fixed temperature factors, and represented the best model likely to be obtained after an initial batch of refinement. The models resulting from each refinement test are compared with this target, including an exhaustive reciprocal-space refinement, which is referred to as the 'yardstick' (W). The original structure determination had preceded publication of R_T^{free} , but our additional refinement enabled calculation of a pseudo R_T^{free} . The diffraction data were randomly split into a working set (95%) used for reciprocal-space refinements and for $2F_o - F_c$ map calculations, and an evaluation set (5% = 1028 reflections) used exclusively for R_T^{free} calculation. The working set gave a low data-to-positional parameter ratio of 1.2 (excluding stereochemical restraints) that is common in medium-resolution protein structure refinements.

Reciprocal-space refinements were performed using *TNT* (Tronrud, Ten Eyck & Matthews, 1987). Real-space refinements were performed using *RSRef* (Chapman, 1995). *RSRef* differs from previous dedicated real-space programs (Diamond, 1971; Jones & Liljas, 1984) in that it explicitly accounts for the (finite) resolution of the electron-density map and uses full stereochemical restraints. Pseudo real-space extensions of *TNT*, *X-FLOR* and *PROLSQ*, that use phase restraints in reciprocal space (Arnold & Rossmann, 1988; Rees & Lewis, 1983) also account for resolution. The only practical difference is *RSRef* can be applied very fast locally, without Fourier transformation of an entire asymmetric unit.

RSRef is implemented using the stereochemical and shift-optimizing routines of *TNT*. This facilitated direct comparisons

with reciprocal-space refinements performed exclusively with *TNT*. Weights for bond lengths, angles, torsion angles and contacts were 1.0, 2.0, 2.0 and 3.0 for reciprocal-space refinements and 1.2, 1.5, 3.0 and 3.0 for real-space refinement, respectively. The weights for structure amplitudes and electron-density values were 0.00013 and 90.0, respectively. To the standard stereochemical targets (Tronrud *et al.*, 1987) were added peptide backbone torsion angle restraints (Chapman & Rossmann, 1996). Procedures for real- and reciprocal-space refinements were analogous, except that shifts were calculated from second derivatives in reciprocal space (Tronrud, 1992), but first derivatives for real space (Chapman, 1995). As in the original refinement (Lawrence *et al.*, 1995), non-crystallographic twofold symmetry was neither constrained nor restrained.

Real-space refinement was first attempted on the unrefined model and the unaveraged MIR-based experimental map (Lawrence *et al.*, 1995). This map had been calculated using both the working and evaluation reflections. Thus, subsequent R factors are not strictly 'free', but are a good approximation after extensive refinement from a starting model with high conventional R factor (R^{conv}). To circumvent the limitations of poor experimental phases, the real-space refined model was used to calculate a $2F_o - F_c$ map for further refinement. Map calculation and real-space refinement were alternated three times. Several of the intermediate real-space refined models were used as starting points for conventional reciprocal-space refinement which converged with 21 to 30 cycles.

Convergence of the new protocols towards the target was compared to an exhaustive 120-cycle reciprocal-space refinement of the initial model (during which there was no interactive remodeling). Differences between models were calculated with program *difftomc* (MSC) allowing for ambiguities in side chains: for example, it is difficult to distinguish crystallographically between two rotamers of histidine or asparagine, related by $180^\circ \chi_2$ rotations. Deviations from non-crystallographic symmetry were calculated with the *TNT* program *NCS*, without constraints on the superimposition matrix. R_T^{free} and R^{conv} were calculated with *TNT*'s *RFactor*, using all data between 20 and 3.0 \AA . $2F_o - F_c$ maps were calculated between 20 and 3 \AA , unweighted, with a grid step of 0.9 \AA . Real-space R factors were calculated by analogy to R^{conv} (Chapman, 1995), thereby differing from those proposed earlier (Jones *et al.*, 1991). They were calculated using map grid points within 3.4 \AA of any atom in the asymmetric unit ($\sim 24 \text{ atom}^{-1}$), and considering contributions of atoms within 3.4 \AA (the same cut-offs used in real-space refinement).

3. Results and discussion

The results are summarized in Table 1.

3.1. Generation of the target structure

Upon resetting the B factors, R^{conv} jumped from 19.7 to 25.7%. It fell during refinement to 21.7%. R_T^{free} rose to converge at 30.2%, as the model became less biased to the diffraction data of the evaluation set. R^{conv} of the fully refined model (Z , 19.7%) (Lawrence *et al.*, 1995) and R_T^{free} of the target (X , 30.2%), are typical of well refined structures at $\sim 3 \text{ \AA}$ resolution. The large disparity between R^{conv} and R_T^{free} indicates overfitting of the model to the data of the working set (Brünger, 1992). In

Table 1. Refinement statistics for several refinement protocols

R factors were calculated as described in the text. Root-mean-square deviations (r.m.s.d.) are given in Å for both stereochemical parameters and the deviations between different coordinate sets. The twofold r.m.s.d. refers to deviations from exact non-crystallographic symmetry.

Start model	End model	Refinement/Map ← phasing model	Cycles	R_T^{free}	R^{conv}	R^{FD}	R.m.s.d. versus target	R.m.s.d. bond lengths (Å)	R.m.s.d. angles (°)	R.m.s.d. torsions (°)	R.m.s.d. twofold
A		Not refined	0	37.2*	37.2	111	0.889	0.041	3.17	15.6	0.0
A	Z	Reciprocal-space refined, rebuilt repeatedly by Lawrence <i>et al.</i> (1995)		n.a.	19.7	49.5	0.353	0.016	3.30	10.8	0.76
Z	Y	B's reset to 20.0 Å ² , no refinement	0	n.a.	25.7		0.353	0.016	3.30	10.8	0.76
Y	X → target	Reciprocal space, fixed B	30	30.2†	21.7		0 (0.56 versus W)	0.016	2.42	16.48	0.78
A	W	Reciprocal space	120	32.7	21.9		0.531	0.017	2.51	16.93	0.73
A	B	Real space/Experimental map	21	34.0*	34.0	97.9	0.788	0.038	4.18	16.8	0.40
B	C	Real space/ $2F_o - F_c \leftarrow B$	30	32.8	27.0	65.6 → 49.8	0.639	0.027	4.13	15.72	0.50
C	D	Real space/ $2F_o - F_c \leftarrow C$	30	33.1	25.2	49.6 → 46.6	0.611	0.025	3.94	15.67	0.51
D	E	Real space/ $2F_o - F_c \leftarrow D$	30	33.3	24.6	46.7 → 45.9	0.601	0.025	3.96	15.82	0.54
B	F	Reciprocal space	30	31.8	22.4		0.526	0.017	2.53	17.01	0.61
C	G	Reciprocal space	30	31.8	22.4		0.526	0.017	2.53	17.01	0.61
E	H	Reciprocal space	30	31.7	22.2		0.518	0.016	2.43	16.97	0.60
F	I	Real space/ $2F_o - F_c \leftarrow F$	30	32.6	23.5	48.1 → 44.5	0.542	0.023	3.71	15.99	0.62
I	J	Reciprocal space	30	31.2	21.8		0.511	0.016	2.40	16.87	0.64
J	K	Real space/ $2F_o - F_c \leftarrow I$	30	32.6	23.5	47.6 → 44.0	0.536	0.023	3.68	15.88	0.66
K	L	Reciprocal space	30	31.6	21.5		0.509	0.016	2.35	16.79	0.67

* The experimental map calculated using the evaluation data. The initial model (A) and those from the first batch of real-space refinement were fit to a map containing this data. Thus, strictly, these can only be considered an approximation to R_T^{free} . (The evaluation data were excluded from subsequent $2F_o - F_c$ maps.) † Model Z was previously exhaustively refined against all data (Lawrence *et al.*, 1995). Evaluation data was only excluded from the 30 cycles reported here. Therefore, R_T^{free} may be underestimated, but in 30 cycles R_T^{free} approaches convergence.

fact, the large deviation (0.78 Å) between non-crystallographically related monomers suggests that the residual root-mean-square (r.m.s.) error is at least 0.39 Å, limiting the usefulness of the target in judging refinement protocols.

3.2. The yardstick – exhaustive reciprocal-space refinement of the starting model

R_T^{free} dropped 4.5% to within 2.5% of the target (models A, X and W). R_T^{free} of the target is lower than the yardstick presumably, because of interactive remodeling. This refinement further illustrates overfitting: between cycle 50 and 120 cycles, R_T^{free} remained unchanged (32.7%), but R^{conv} decreased from 22.6 to 21.9%. As R^{conv} depends on the number of refinement cycles and strength of restraints (including implicit phase restraints), R_T^{free} was taken as a more reliable indicator of model reliability.

3.3. Real-space refinement

With a poor map, refinement, exclusively in real space, gives a model (B, $R_T^{\text{free}} = 34.0\%$) that is inferior to reciprocal-space refinement (W, $R_T^{\text{free}} = 32.7\%$). Although worse, the real-space pre-refined model (B) is a good starting point for subsequent reciprocal-space refinement. The resulting model (F) has both coordinates and R_T^{free} (31.8%) closer to the target than possible with reciprocal-space refinement alone (W, $R_T^{\text{free}} = 32.7\%$).

3.4. Alternated real-space refinement and map calculation

Use of a $2F_o - F_c$ map using phases calculated from the real-space-refined model allowed real-space refinement to progress further. R_T^{free} drops from 34.0 to 32.8% (models B and C). Models get slightly worse (models D and E; R_T^{free} increasing) if the map calculation and refinement are iterated. Decreased R^{conv} and deviation from target suggest that the detrimental effects on iteration are minor and temporary – subsequent reciprocal-space refinements generated models (F, G and H) of similar quality.

3.5. Alternated real- and reciprocal-space refinements

Phases might improve faster with reciprocal-space refinement. Each iteration was expanded to 30 cycles of real-space refinement, 30 cycles of reciprocal refinement and $2F_o - F_c$ map calculation. At the end of the second reciprocal-space refinement (model J), R_T^{free} was within 1% of the target. Further refinements were detrimental (model L), presumably because of phase bias (see above).

3.6. Analysis of coordinate changes

Analysis of the coordinate changes is complicated, because the 'true' structure is unknown, and the likely residual errors of models refined by any of the protocols are about as large as the differences between them. The residual error in the target (and

other models) can be estimated by comparing protomers related by the non-crystallographic symmetry that was not imposed. If independent, the remaining experimental errors of models J and X could more than account for the 0.51 Å difference (Table 1). It is, therefore, very difficult to analyze small improvements in refinement protocols, and the cross-validation free R factors (above) are a better indicator. Nevertheless, an attempt was made to compare coordinates resulting from alternated real and reciprocal refinements (J) or from exclusive reciprocal-space refinement (W) both against the target (X), with the expectation that due to random error, some of the atoms of J would be closer than W to X and some further away. This accounts for the scatter in Fig. 1, which also shows that on average the alternating procedure gives coordinates slightly closer to the target.

4. Conclusions

This analysis confirms that real-space refinement is not equal to reciprocal-space refinement for protein structure determination in which experimental maps are typically poor. However, when used as a supplement to reciprocal-space methods, protein models can be improved more than with exclusive use of reciprocal-space methods. The most efficient protocols iterate real-space refinement, reciprocal-space refinement and $2F_o - F_c$ map calculation. Divergence (probably because of phase bias) can be stopped through monitoring R_T^{rec} .

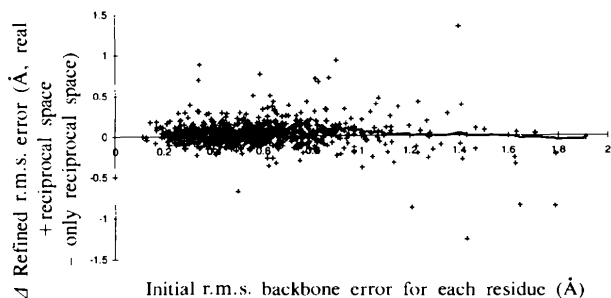


Fig. 1. Comparison of refinement protocols that use or are independent of real-space methods. The x axis shows the difference between the initial model (A of Table 1) and the target (X). The y axis shows the difference in deviation from the target between a model (W) refined exclusively in reciprocal space and a combination real/reciprocal-space refined model (J). Positive values indicate that the combination real/reciprocal-space refinement lead to greater improvement. Thus, points above the horizontal axis represent amino acids for which the combination refinement is better. The trend line was calculated by averaging windows of 50 data points. The graph shows wide scatter, but some average improvement in agreement (for residues within 1.2 Å of the target position) with the use of both real- and reciprocal-space refinement. The wide scatter could be because of widely different outcomes for different residues, but is also caused by the large residual error in the target coordinates suggested by a 0.8 Å deviation between non-crystallographically related monomers. The graph is likely to underestimate the improvement resulting from real-space methods, because the residual error of the target structure is likely to be correlated more to the errors of models refined similarly in reciprocal space than to models refined in real space.

The improvement is nearly as great as can be realized through several labor-intensive rounds of interactive model building and re-refinement. At the start of refinement, real-space methods are likely to improve the conditioning of the optimization by increasing the data-to-parameter ratio, through the use of phase information, and also by restricting the interdependencies of atoms to local interactions. Improved conditioning will frequently lead to improved convergence. During the course of refinement, real-space refinement is likely to force atoms to fit the density in cases where the overfitting of atoms to structure amplitudes (Brünger, 1992) can cause atoms to move away from their correct positions. Even when real-space refinement with $2F_o - F_c$ maps was slightly detrimental, leading to increased R_T^{rec} , subsequent reciprocal-space refinement proceeded past the prior point of convergence. This beneficial effect might result from real-space refinement moving coordinates slightly from the positions of a local minimum in the reciprocal-space optimization. Many have seen similar, but lesser, effects (higher initial R factors, but subsequent improved refinement), following interactive remodeling, random shifts to coordinates, or changes in weights. In an intuitive sense, real-space refinement is automatically accomplishing some of the goals of interactive remodeling. It is possible that real-space methods will be more powerful when combined with omit maps and molecular dynamics to reduce phase bias effects (Hodel, Kim & Brünger, 1992).

We are grateful to Drs Martin Lawrence and Cynthia Stauffacher for encouragement to use the HMG-CoA structure prior to publication, and to Mike Slodebeck for computational advice. This work was generously supported by the National Science Foundation (MSC; BIR9418741) and the Lucille P. Markey Charitable Trust.

References

- Arnold, E. & Rossmann, M. G. (1988). *Acta Cryst.* **A44**, 270–282.
- Brünger, A. T. (1992). *Nature (London)*, **355**, 472–475.
- Brünger, A. T., Kuriyan, J. & Karplus, M. (1987). *Science*, **235**, 458–460.
- Chapman, M. S. (1995). *Acta Cryst.* **A51**, 69–80.
- Chapman, M. S. & Rossmann, M. G. (1996). *Acta Cryst.* **D52**, 129–142.
- Diamond, R. (1971). *Acta Cryst.* **A27**, 436–452.
- Hendrickson, W. W. (1985). *Methods Enzymol.* **115**, 252–270.
- Hodel, A., Kim, S.-H. & Brünger, A. T. (1992). *Acta Cryst.* **A48**, 851–858.
- Jones, T. A. & Liljas, L. (1984). *Acta Cryst.* **A40**, 50–57.
- Jones, T. A., Zou, J.-Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Acta Cryst.* **A47**, 110–119.
- Lawrence, C. M., Rodwell, V. M. & Stauffacher, C. V. (1995). *Science*, **268**, 1758–1762.
- Rees, D. C. & Lewis, M. (1983). *Acta Cryst.* **A39**, 94–97.
- Rossmann, M. G., McKenna, R., Tong, L., Xia, D., Dai, J.-B., Wu, H., Choi, H.-K. & Lynch, R. E. (1992). *J. Appl. Cryst.* **25**, 166–180.
- Sussmann, J. L., Holbrook, S. R., Church, G. M. & Kim, S.-H. (1977). *Acta Cryst.* **A33**, 800–804.
- Tronrud, D. E. (1992). *Acta Cryst.* **A48**, 912–916.
- Tronrud, D. E., Ten Eyck, L. F. & Matthews, B. W. (1987). *Acta Cryst.* **A43**, 489–501.