Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Extending the limits of molecular replacement through combined simulated annealing and maximum-likelihood refinement

Paul D. Adams,[a] Navraj S. Pannu,[b] Randy J. Read[c] and Axel T. Brunger[d]*

[a]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA, [b]Department of Mathematical Sciences, University of Alberta, Edmonton, Alberta T6G 2G1, Canada, [c]Department of Medical Microbiology and Immunology, University of Alberta, Edmonton, Alberta T6G 2H7, Canada, and [d]The Howard Hughes Medical Institute and Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA

Correspondence e-mail: brunger@laplace.csb.yale.edu

Received 16 March 1998
Accepted 6 May 1998

Phases determined by the molecular-replacement method often suffer from model bias. In extreme cases, the refinement of the atomic model can stall at high free *R* values when the resulting electron-density maps provide little indication of how to correct the model, sometimes rendering even a correct solution unusable. Here, it is shown that several recent advances in refinement methodology allow productive refinement, even in cases where the molecular-replacement-phased electron-density maps do not allow manual rebuilding. In test calculations performed with a series of homologous models of penicillopepsin using either backbone atoms, or backbone atoms plus conserved core residues, model bias is reduced and refinement can proceed efficiently, even if the initial model is far from the correct one. These new methods combine cross-validation, torsion-angle dynamics simulated annealing and maximum-likelihood target functions. It is also shown that the free *R* value is an excellent indicator of model quality after refinement, potentially discriminating between correct and incorrect molecular-replacement solutions. The use of phase information, even in the form of bimodal single-isomorphous-replacement phase distributions, greatly improves the radius of convergence of refinement and hence the quality of the electron-density maps, further extending the limits of molecular replacement.

1. Introduction

The number of solved macromolecular structures is ever increasing. As the structural database increases in size, the molecular-replacement method (Hoppe, 1957; Rossmann & Blow, 1962) is used more frequently to derive initial phases for a new structure. This method relies upon the existence of some structural similarity between the known and unknown structures. There have been several advances in the molecular-replacement technique which extend its scope (Turkenburg & Dodson, 1996). The direct-rotation search (Delano & Brunger, 1995) combined with Patterson correlation (PC) refinement (Brunger, 1990), followed by a correlation-coefficient translation search (Fuginaga & Read, 1987) have greatly improved the ability to find the correct molecular-replacement solution, in some cases allowing the correct solution to be found using only a polyalanine model (Delano & Brunger, 1995). The fast correlation-coefficient translation search (Navaza & Vernoslova, 1995) has introduced the possibility of complete six-dimensional molecular-replacement searches and rapid trials with different initial models.

An outstanding problem in the molecular-replacement method is model rebuilding and subsequent refinement of the unknown structure. The initial phases for the unknown structure are derived from the known structure. This can lead

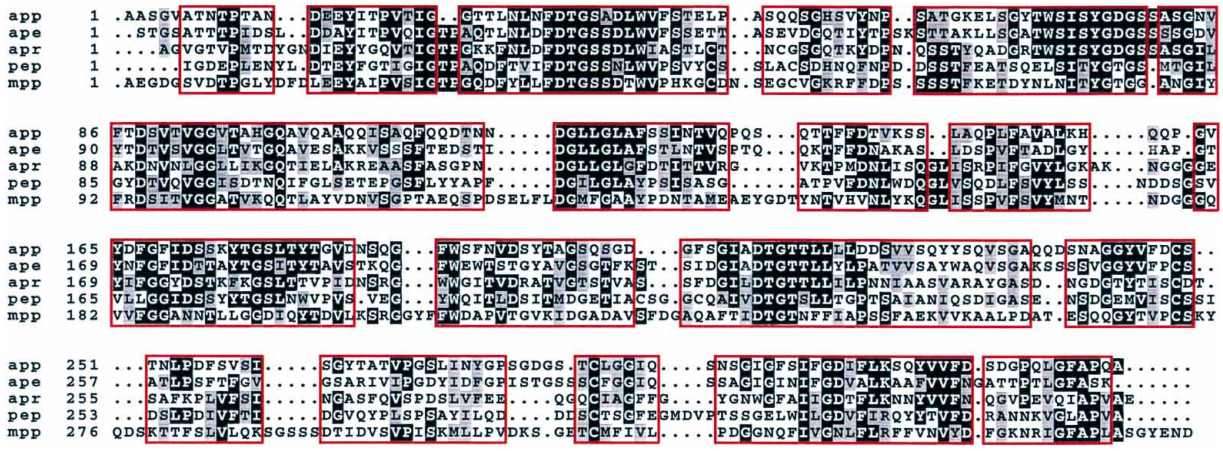

Figure 1
Multiple sequence alignment for selected aspartic proteinases: 3APP (penicillopepsin; Sielecki *et al.*, 1991), 4APE (endothiapepsin; Blundell *et al.*, 1990), 2APR (rhizopuspepsin; Suguna *et al.*, 1987), 3PEP (porcine pepsin; Abad-Zapatero *et al.*, 1990) and 1MPP (*Mucor pusillus* pepsin; Newman *et al.*, 1993). Sequence alignment generated by the program *ClustalW* (Thompson *et al.*, 1994). Regions of amino-acid identity with respect to the penicillopepsin sequence are shaded black. Conservative amino-acid substitutions are shaded grey. The conserved regions which lack either insertions or deletions in the alignment are shown boxed in red. These regions were used in both molecular replacement and subsequent refinements.

to model bias, especially in the case where the search model has significant differences from the true structure. An additional problem is that parts of the true structure may be missing from the search model, *e.g.* when only a polyalanine model is used. Thus, even after the correct placement of the search model has been found, the initial phases may not reveal the missing parts of the structure and may show electron density which confirms the search model instead of the true structure. Correct placement of the model by manual fitting to these model-biased electron-density maps is often impossible. Partial automation of refinement and reduction of model bias is therefore an essential part of a solution of the phase problem by molecular replacement.

The underlying cause for model bias during refinement is a low ratio of data to parameters, owing to the lack of atomic resolution diffraction data for macromolecules. During the refinement of the model, it is often the case that initial errors in the model combined with this low data-to-parameter ratio result in overfitting of the diffraction data, and hence errors in the final model. Overfitting is the introduction of systematic errors into the model while still apparently improving the fit of the model to the experimental data. Many developments have been designed to reduce overfitting: cross-validation (Brunger, 1992), powerful optimization techniques such as simulated annealing (Brunger *et al.*, 1987, 1997), simulated annealing in torsion-angle space (Rice & Brunger, 1994) and maximum-likelihood targets (Pannu & Read, 1996; Murshudov *et al.*, 1997; Bricogne, 1997) combined with cross-validation (Kleywegt & Brunger, 1996; Read, 1997). Probabilistic methods for estimating the effects on calculated structure factors of errors in the current model also make it possible to reduce model bias in the calculation of electron-density maps through σ_A weighting (Read, 1986). Furthermore, prior phase information can be included in the process of structure refinement (Arnold & Rossmann, 1988) even if the phase probability distribution is not unimodal (Murshudov

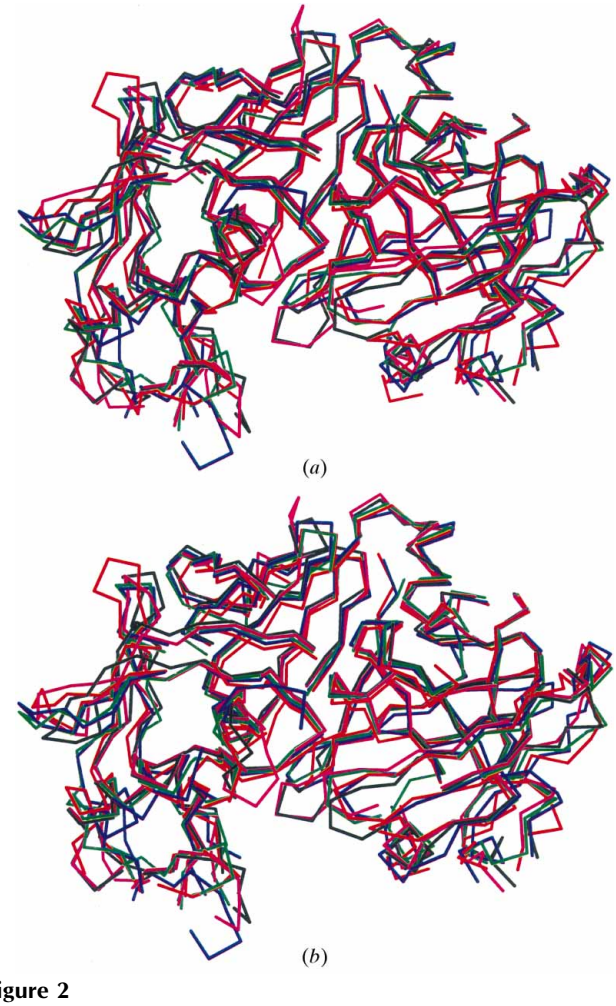

Figure 2
Superposition of the molecular-replacement solutions for both kinds of models: (*a*) polyalanine, (*b*) polyalanine plus conserved residues. Only Cα atoms are shown for conserved residues for each structure: penicillopepsin (black), endothiapepsin (magenta), rhizopuspepsin (green), porcine pepsin (blue) and *M. pusillus* pepsin (red). Breaks in the chains indicate insertions or deletions (see text for details).

Table 1
Molecular-replacement solutions for models consisting of polyalanine only.

The level of sequence identity between each model and the penicillopepsin (3APP) sequence was calculated from multiple sequence alignment (Fig. 1). For the rotation searches, the number in parenthesis is the highest error peak. The ten highest rotation-function peaks were tested by translation searches. The translation search with the highest correlation coefficient always produced the correct solution, even in the case of 4APE where the highest rotation-search peak is incorrect. For the translation searches, the value in parenthesis is the highest error peak among all translation searches for the particular model. The value of the rotation function (RF) is the product between the selected rotated vectors of the Patterson map calculated from the model and the interpolated values of the Patterson map calculated from the observed data (Huber, 1965; Steigemann, 1974). The translation correlation coefficient is calculated between squared structure factors. The backbone coordinate r.m.s. deviation (RMSD) was calculated between the penicillopepsin crystal structure and the search model, placed according to the molecular-replacement solution, after application of the appropriate origin shifts and symmetry operations (see text for details).

		Rotation search				Translation search				
Molecule	Identity (%)	θ_1 (°)	θ_2 (°)	θ_3 (°)	RF	X (Å)	Y (Å)	Z (Å)	Correlation	RMSD (Å)
3APP	100	0.0	0.0	0.0	3.05 (1.76)	−0.08	−0.16	0.13	0.64 (0.43)	0.25
4APE	52	197.3	90.0	314.6	1.44 (2.10)	−9.88	−5.44	27.65	0.46 (0.29)	1.50
2APR	36	287.6	70.0	235.3	2.23 (1.85)	46.22	0.86	−3.52	0.40 (0.34)	2.44
3PEP	31	196.4	30.0	93.6	2.09 (1.76)	34.14	−0.16	23.48	0.38 (0.33)	2.88
1MPP	25	63.0	80.0	1.3	1.79 (1.57)	22.67	−2.36	23.85	0.34 (0.28)	2.94

et al., 1997; Pannu *et al.*, 1998). Thus, even single isomorphous replacement (SIR) phases from one heavy-atom derivative can be useful. Here, it is shown that these developments in refinement allow the successful use of poorer models than previously possible for molecular-replacement phasing. A series of realistic starting models obtained from homologous structures of the aspartic proteinase penicillopepsin were tested in molecular replacement and subsequent refinement against the penicillopepsin diffraction data. It is shown that the new refinement methods reduce the model bias inherent in the molecular-replacement method and greatly extend the radius of convergence, allowing refinement even in cases where manual rebuilding of the initial model appears impossible.

2. Materials and methods

2.1. Penicillopepsin diffraction data and coordinates

Diffraction data for the native crystal and a $K_3UO_2F_5$ derivative of the aspartic proteinase penicillopepsin (Sielecki *et al.*, 1991) were used. Native diffraction data were available from a maximum Bragg spacing of d_{max} = 22.0 Å to a minimum Bragg spacing of d_{min} = 1.8 Å in space group $C2$ with unit-cell parameters a = 97.37, b = 6.64, c = 65.47 Å and β = 115.4°. SIR phase probability distributions were used in the MLHL target to the diffraction limit of the derivative (d_{min} = 2.8 Å). Phases were of good quality with an overall figure of merit of 0.45, an overall phasing power of 2.1 and an overall R_{Cullis} of 0.56. The coordinates of the refined structure at a minimum Bragg spacing of d_{min} = 1.8 Å were taken from the Protein Data Bank (Bernstein *et al.*, 1977), with accession code 3APP.

2.2. Sequence searching and alignment

The amino-acid sequence of the penicillopepsin structure (3APP; Sielecki *et al.*, 1991) was used to search the sequence database of known structures (NRL_3D; Pattabiraman *et al.*,

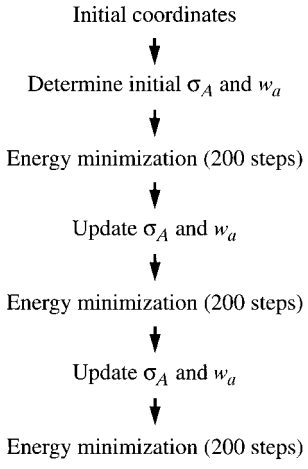

Figure 3
Extensive conjugate-gradient energy-minimization refinement protocol. The cross-validated σ_A estimates for the maximum-likelihood targets, and the weight (w_a) between geometric and crystallographic energy terms were updated at the beginning of each cycle.

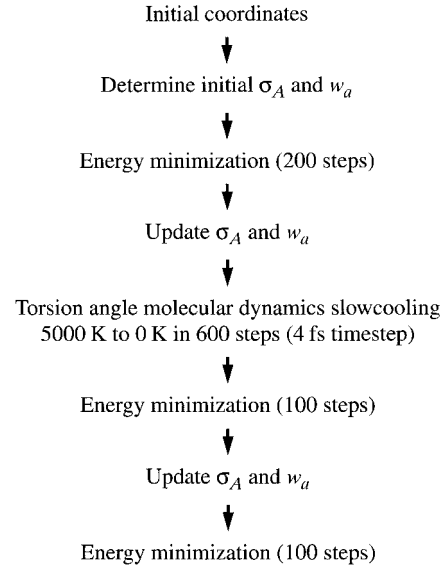

Figure 4
Torsion-angle molecular-dynamics simulated-annealing refinement protocol. The cross-validated σ_A estimates for the maximum-likelihood targets, and the weight (w_a) between geometric and crystallographic energy terms were updated as indicated.

1990). Several structures of homologous proteins were found. Four of these structures were taken, with sequence identity ranging from 100 to 25%. The proteins were: *Mucor pusillus* pepsin (1MPP; Newman *et al.*, 1993), porcine pepsin (3PEP; Abad-Zapatero *et al.*, 1990), rhizopuspepsin (2APR; Suguna *et al.*, 1987) and endothiapepsin (4APE; Blundell *et al.*, 1990). All of these structures had been refined at minimum Bragg spacings between 2.3 and 1.8 Å. The sequences were aligned using the multiple alignment program *ClustalW* (Thompson *et al.*, 1994).

2.3. Search models

The multiple sequence alignment produced conserved segments across all sequences (boxes outlined in Fig. 1), interspersed by insertions and deletions relative to the penicillopepsin sequence. The regions of insertions or deletions were removed from each model including the penicillopepsin model itself (3APP). When comparing sequences, no use was made of the structure of the models, in order to mimic the common case where several sequences of homologous proteins are known yet only one homologous structure is known. Clearly, the use of common structural information derived from all the known structures (Holm *et al.*, 1992; Schmidt *et al.*, 1997) could be used to create better models for molecular replacement when possible (Leahy *et al.*, 1992; Müller *et al.*, 1995). After removal of insertions and deletions, the five resulting models had 294 residues out of a total of 323 residues in penicillopepsin. The same sequence numbering as penicillopepsin was maintained, facilitating convenient comparisons between the partial model and the penicillopepsin crystal structure. In this way, a series of partial models with the same number of residues but differing in coordinate error and amino-acid sequence relative to the penicillopepsin crystal structure was created. Two sets of search models were generated: polyalanine only and polyalanine plus side chains for residues conserved in the multiple sequence alignment of homologous models and penicillopepsin (Fig. 1).

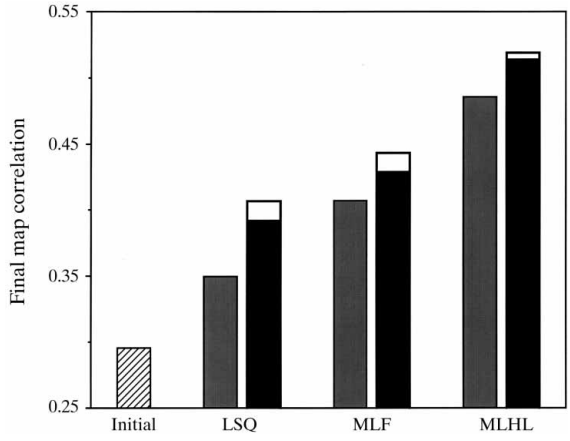

Figure 6
Map correlation coefficients of σ_A-weighted electron-density maps before and after refinement against the native penicillopepsin diffraction data for the polyalanine model derived from *M. pusillus* pepsin (PDB code 1MPP) with an initial r.m.s. deviation of 2.94 Å to the penicillopepsin crystal structure. Correlation coefficients are between cross-validated σ_A-weighted maps calculated from each model and from the published penicillopepsin structure. The initial correlation coefficient is shown as a diagonally hatched bar. Results of extensive conjugate-gradient energy minimization (protocol in Fig. 3) are shown as grey bars. Simulated-annealing refinements (protocol in Fig. 4) were repeated five times with different initial velocities. The numerical averages of the map correlation coefficients for the five refinements are shown as black bars. The best map correlation coefficients from simulated annealing are shown as white bars.

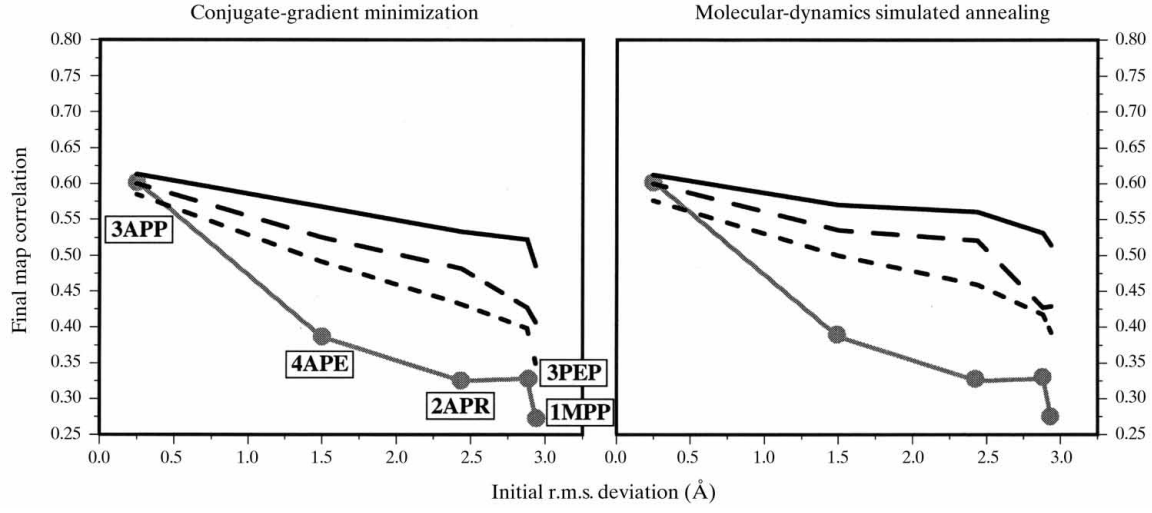

Figure 5
Final overall map correlation coefficients of σ_A-weighted electron-density maps *versus* initial backbone r.m.s. deviation coordinate error to the refined penicillopepsin model, for the five polyalanine molecular-replacement solutions. Results are shown for both extensive energy minimization and torsion-angle molecular-dynamics simulated annealing. The initial map correlation coefficient prior to any refinement is shown in grey and the models are indicated by labels on the left-hand graph and filled circles on both graphs. After refinement, the map correlation coefficients are shown for the least-squares-residual target (LSQ, short-dashed line), maximum likelihood using amplitude information (MLF, long-dashed line) and maximum likelihood using experimental phase information (MLHL, solid line). The simulated-annealing results are the numerical average of five independent refinements. All simulated-annealing runs performed better than minimization runs with the same target (not shown).

2.4. Molecular replacement

The models generated were used as search models in molecular-replacement phasing, against the native penicillopepsin diffraction data. Molecular replacement was carried out in the *Crystallography & NMR System* (*CNS*; Brunger *et al.*, 1998). Native penicillopepsin diffraction data between d = 15 and 4.0 Å were used in a real-space Patterson superposition method (Huber, 1965; Steigemann, 1974). The top ten peaks from this rotation search were each subjected to 20 steps of Patterson correlation (PC) refinement (Brunger, 1990) using a target of squared normalized structure factors (Hauptman, 1982; Fujinaga & Read, 1987). On the basis of the topology of the aspartic proteinases the models were divided into two domains for PC refinement: using the penicillopepsin numbering scheme, residues 6–94 and 209–220 formed one domain, while residues 195–208 and 221–295 formed the other. A fast translation search (Navaza & Vernoslova, 1995; Grosse-Kunstleve & Brunger, 1999) was performed for each of the top ten rotation-function peaks after PC refinement. The highest translation-search peak was further improved by PC refinement (20 steps). In all cases the correct solution was found within the first ten cross-rotation peaks, and was characterized by a significantly higher correlation coefficient in subsequent translation searches with each of these solutions (Tables 1 and 2). Appropriate origin shifts and symmetry operations were applied to the models obtained from molecular replacement, in order to superimpose them on the published penicillopepsin crystal structure (Fig. 2). The polar space group $C2$ results in arbitrary placement of the molecule along the y axis. Therefore, the centres of geometry of each solution and the penicillopepsin crystal structure were superimposed along the y axis.

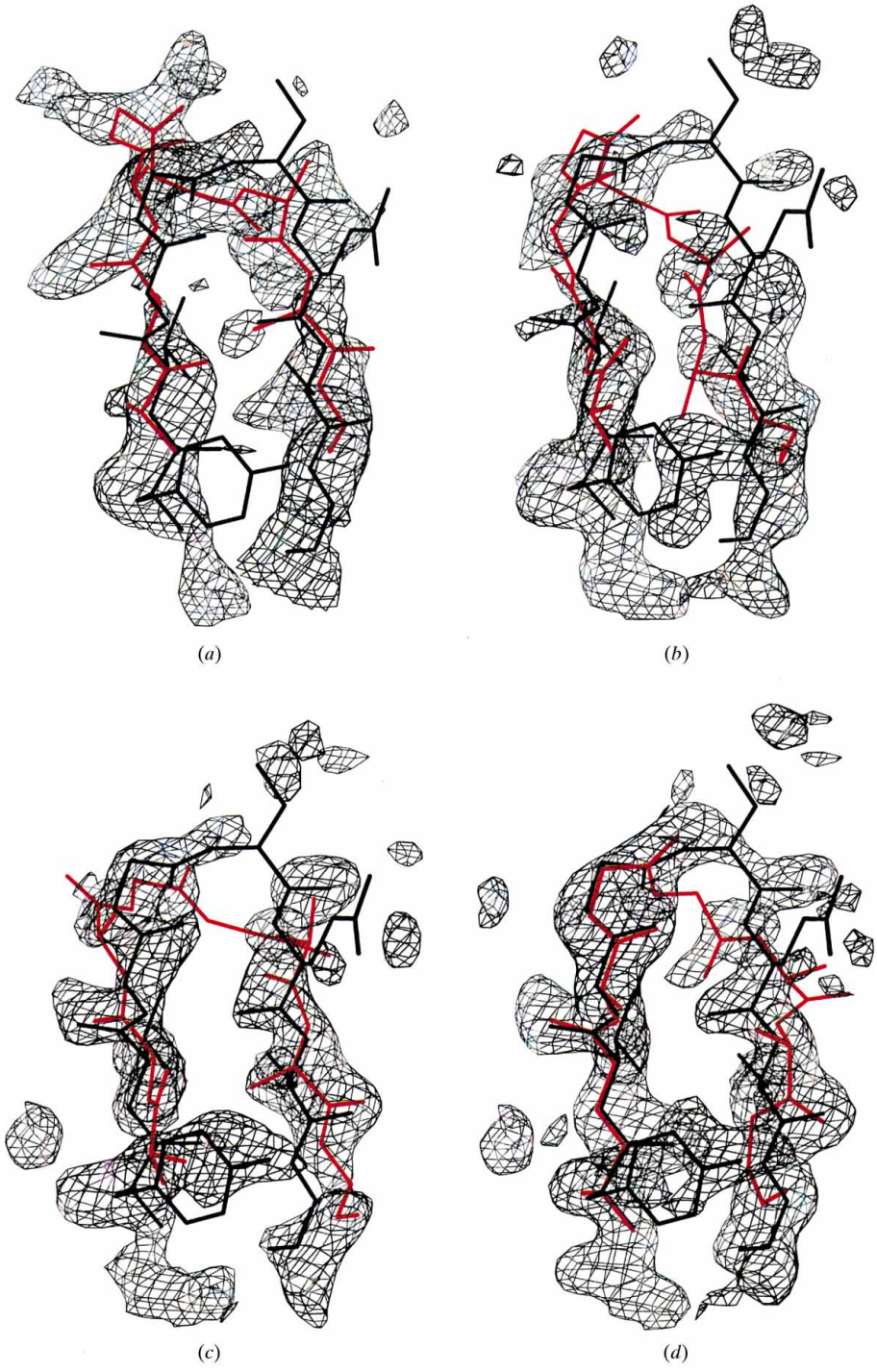

Figure 7
Maximum-likelihood targets significantly decrease model bias in simulated-annealing refinement. Cross-validated σ_A-weighted electron-density maps contoured at 1.25σ for models from simulated-annealing refinement with different targets. The starting model, derived from *M. pusillus* pepsin (PDB code 1MPP), consisted of a polyalanine backbone only and had an initial r.m.s. deviation of 2.94 Å to the penicillopepsin crystal structure. Residues 198–205 are shown with the published penicillopepsin structure in black and the model with the lowest free R value from five independent refinements in red. (*a*) Initial electron-density map prior to refinement, (*b*) after refinement with the LSQ target, (*c*) after refinement with the MLF target, (*d*) after refinement with the MLHL target.

2.5. Refinement protocols

Two different types of optimization methods for refinement were tested: extensive conjugate-gradient minimization and torsion-angle molecular-dynamics simulated annealing as implemented in *CNS*. Refinements were repeated with three different refinement targets: least-squares residual (LSQ), maximum likelihood based on amplitudes (MLF; Pannu & Read, 1996) and maximum likelihood incorporating experimental phase informa-

Table 2
Molecular-replacement solutions.

As Table 1, but for models consisting of polyalanine plus side chains for conserved residues.

	Rotation search				Translation search				
Molecule	θ_1 (°)	θ_2 (°)	θ_3 (°)	RF	X (Å)	Y (Å)	Z (Å)	Correlation	RMSD (Å)
3APP	0.0	0.0	0.0	4.27 (1.96)	0.02	−0.03	0.02	0.80 (0.52)	0.21
4APE	199.1	86.7	323.9	2.41 (1.56)	−9.98	0.83	27.83	0.56 (0.21)	1.38
2APR	287.5	70.0	235.3	2.22 (1.50)	45.99	0.87	−3.52	0.45 (0.17)	2.44
3PEP	196.3	33.3	91.7	1.98 (1.58)	34.09	−0.25	23.18	0.40 (0.20)	2.89
1MPP	63.0	80.0	1.3	1.46 (1.47)	22.54	−2.71	24.25	0.36 (0.16)	2.96

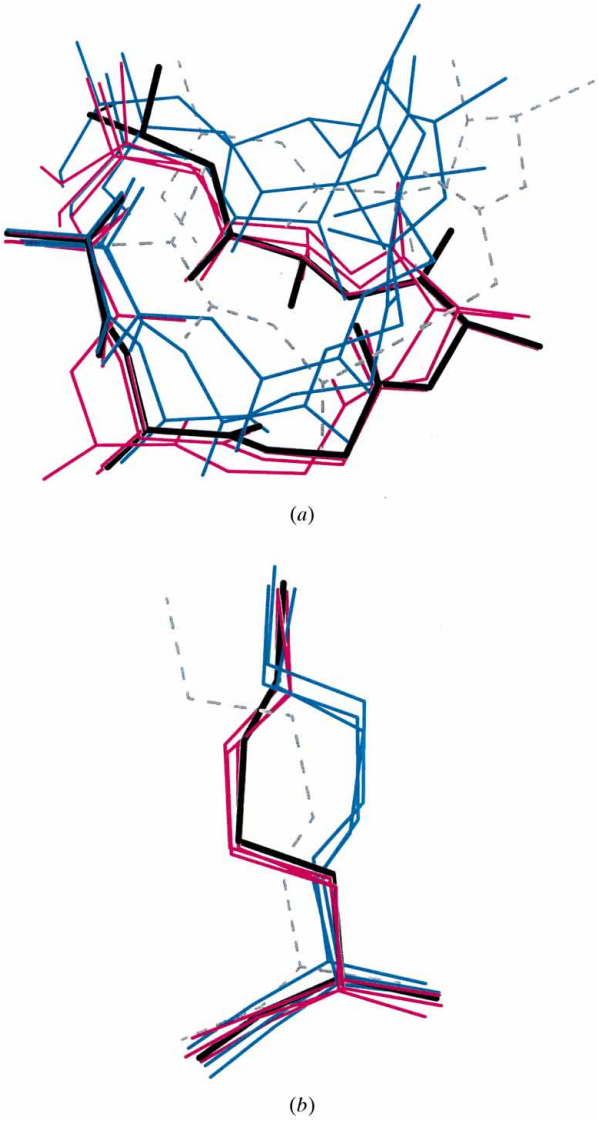

Figure 8
Simulated annealing is able to correct large coordinate errors in the model. The results of all refinements with extensive conjugate-gradient minization with the three different refinement targets are shown in cyan. The best refinements from simulated annealing, as judged by the free R value, for the three refinement targets are shown in magenta. The published penicillopepsin model is shown in black. The initial model prior to refinement is shown in grey (dashed). (*a*) Residues 199–204 in the polyalanine model derived from 2APR. (*b*) Residue 145 in the conserved residue model derived from 4APE.

tion (MLHL; Pannu *et al.*, 1998). In all cases a flat bulk-solvent model (Jiang & Brunger, 1994) and overall anisotropic B-factor correction were applied. Masks for the bulk-solvent correction were created with the program *MAMA* (Kleywegt & Jones, 1993) using the models consisting of polyalanine plus conserved residues. The mask was expanded twice and contracted twice in order to remove fine surface detail and any internal cavities. All native penicillopepsin diffraction data, truncated to the Bragg spacing where data completeness dropped below 90% ($d_{\min}$ = 2.0 Å), were used in the refinements with no amplitude-based cutoff or outlier rejection applied. Refinements using experimental phase information with the MLHL target also included the SIR experimental phase probability distributions which were available to $d_{\min}$ = 2.8 Å.

Extensive conjugate-gradient energy minimization (Fig. 3) was carried out to provide a fair comparison to simulated annealing. Thus, the computational time required for minimization was equivalent to that required for a single trial of simulated annealing. Three cycles of 200 steps of minimization were carried out. The σ_A values (if required by the target) and the weight (w_a) between crystallographic and geometric terms were automatically recalculated at the beginning of each cycle. Simulated annealing using torsion-angle molecular dynamics employed a protocol used previously (Adams *et al.*, 1997). Briefly, initial minimization was followed by slow cooling from 5000 to 0 K in 50 K steps (Fig. 4). For each temperature decrement, six steps of torsion-angle molecular dynamics with a time step of 4 fs were used. Temperature control used the velocity-scaling method. Following molecular dynamics, 200 steps of conjugate-gradient energy minimization were performed. In the case of simulated annealing, five independent trials were performed, each with different initial velocities (Brunger, 1988).

2.6. Comparisons

The R value, free R value, r.m.s. deviation and unweighted phase error between each model and the published penicillopepsin crystal structure were calculated before and after refinement. In addition, the overall map correlation coefficient in the protein region was calculated between the σ_A-weighted $(2m|F_o| - D|F_c|)\exp(i\varphi_c)$ electron-density map (Read, 1986)

from the model and the same map calculated from the penicillopepsin crystal structure.

3. Results

3.1. Refinement of polyalanine models

The polyalanine models were very incomplete, with only 50% of the protein content in the asymmetric unit of the crystal accounted for. In addition, the worst models had significant coordinate errors compared to the penicillopepsin crystal structure; the overall r.m.s. deviation for backbone atoms was approximately 3 Å. Electron-density maps calculated from the molecular-replacement solutions were severely model biased, with electron density covering many incorrectly placed atoms (as shown, for example, for *Mucor pusillus* in Fig. 7*a*). Manual model rebuilding would have been impossible for much of the model at this stage. After refinement the models were improved, all moving closer to the penicillopepsin crystal structure (Fig. 5). However, there were significant differences in the level of improvement produced by the two refinement methods (conjugate-gradient minimization and simulated annealing) and the three refinement targets (LSQ, MLF and MLHL) (Figs. 5, 6, 7*b*, 7*c* and 7*d*).

Simulated annealing has a larger radius of convergence than conjugate-gradient methods (Brunger *et al.*, 1987; Brunger, 1988; Gros *et al.*, 1989; Rice & Brunger, 1994; Adams *et al.*, 1997). The poor polyalanine models used here are yet another stringent test of the method. In several regions of the search models significant deviations from the penicillopepsin crystal structure were present, as is commonly observed in models derived from molecular replacement. In many cases simulated annealing was able to correct large errors in main-chain position, producing shifts of up to 5 Å in loop regions (Fig. 8*a*). The automatic correction of these large errors produced models with better correlation coefficients (Fig. 5). In contrast, even extensive conjugate-gradient minimization remained stalled in local minima close to the starting model (Fig. 8*a*).

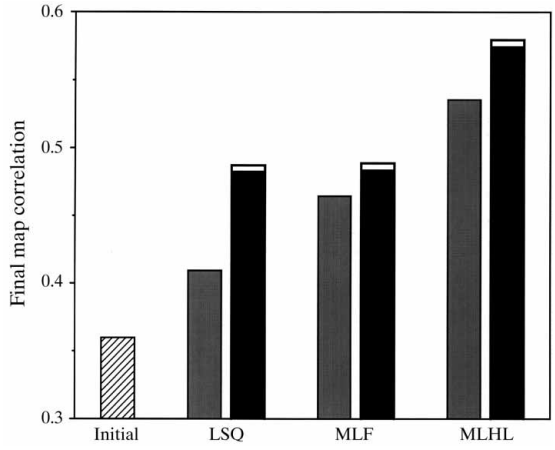

Figure 10
Map correlation coefficients of σ_A-weighted electron-density maps before and after refinement against the native penicillopepsin diffraction data for the polyalanine model plus side chains for conserved residues derived from porcine pepsin (PDB code 3PEP) with an initial r.m.s. deviation of 2.89 Å to the penicillopepsin crystal structure. Correlation coefficients are between cross-validated σ_A-weighted maps calculated from each model and from the published penicillopepsin structure. The initial correlation coefficient is shown as a diagonally hatched bar. Results of extensive conjugate-gradient energy minimization are shown as grey bars. Simulated-annealing refinements were repeated five times with different initial velocities. The numerical averages of the map correlation coefficients for the five refinements are shown as black bars. The best map correlation coefficients from simulated annealing are shown as white bars.

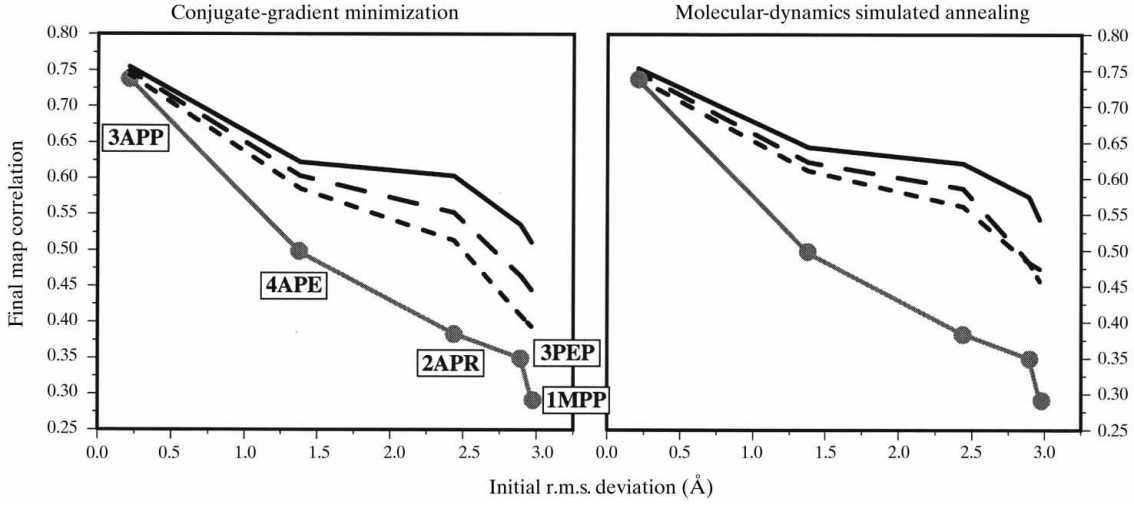

Figure 9
Final overall map correlation coefficients of σ_A-weighted electron-density maps *versus* initial backbone r.m.s. deviation coordinate error to the penicillopepsin crystal structure for the five models consisting of polyalanine residues plus side chains for conserved residues. Results are shown for both extensive energy minimization and torsion-angle molecular-dynamics simulated annealing. The initial map correlation coefficient prior to any refinement is shown in grey and the models are indicated by labels on the left-hand graph and filled circles on both graphs. After refinement, the map correlation coefficients are shown for the least-squares residual target (LSQ, short-dashed line), maximum likelihood using amplitude information (MLF, long-dashed line) and maximum likelihood using experimental phase information (MLHL, solid line). The simulated-annealing results are the numerical average of five independent refinements. All simulated-annealing runs performed better than minimization runs with the same target (not shown).

For a given refinement target, simulated annealing always produced better models than conjugate-gradient minimization. Even the worst model from multiple simulated-annealing trials is better than that obtained from conjugate-gradient minimization. In addition, the stochastic nature of molecular dynamics produces different models when several trials are performed each with different initial velocities (Brunger, 1988). Often, at least one model is produced which is significantly better than the others, with up to a 10% improvement in map correlation coefficient compared to the numerical average of the independent trials (Fig. 6). In addition, the averaging of structure factors from the best models in the ensemble can further improve the quality of electron-density maps (Rice *et al.*, 1998).

By accounting for the statistical effects of errors in the model, the maximum-likelihood method can significantly further improve macromolecular refinement (Pannu & Read, 1996; Murshudov *et al.*, 1997) especially when combined with simulated annealing (Adams *et al.*, 1997). This is particularly true when the model is very incomplete, as in the case of polyalanine. The standard least-squares residual (LSQ) target overfits the diffraction data, resulting in significant model bias even after extensive refinement (Fig. 7*b*), making manual model building exceedingly difficult or impossible. When cross-validated estimates of model errors (Kleywegt & Brunger, 1996; Read, 1997) and errors in the observed diffraction data were taken into consideration with the maximum-likelihood target based on amplitudes (MLF), the refinement produced a model which was closer to the penicillopepsin crystal structure and also resulted in less model-biased maps (Fig. 7*c*). Even when the model was still incorrectly placed, the σ_A-weighted $(2m|F_o| - D|F_c|)\exp(i\varphi_c)$ electron-density maps showed either the correct chain trace or significantly reduced electron density for incorrectly placed atoms.

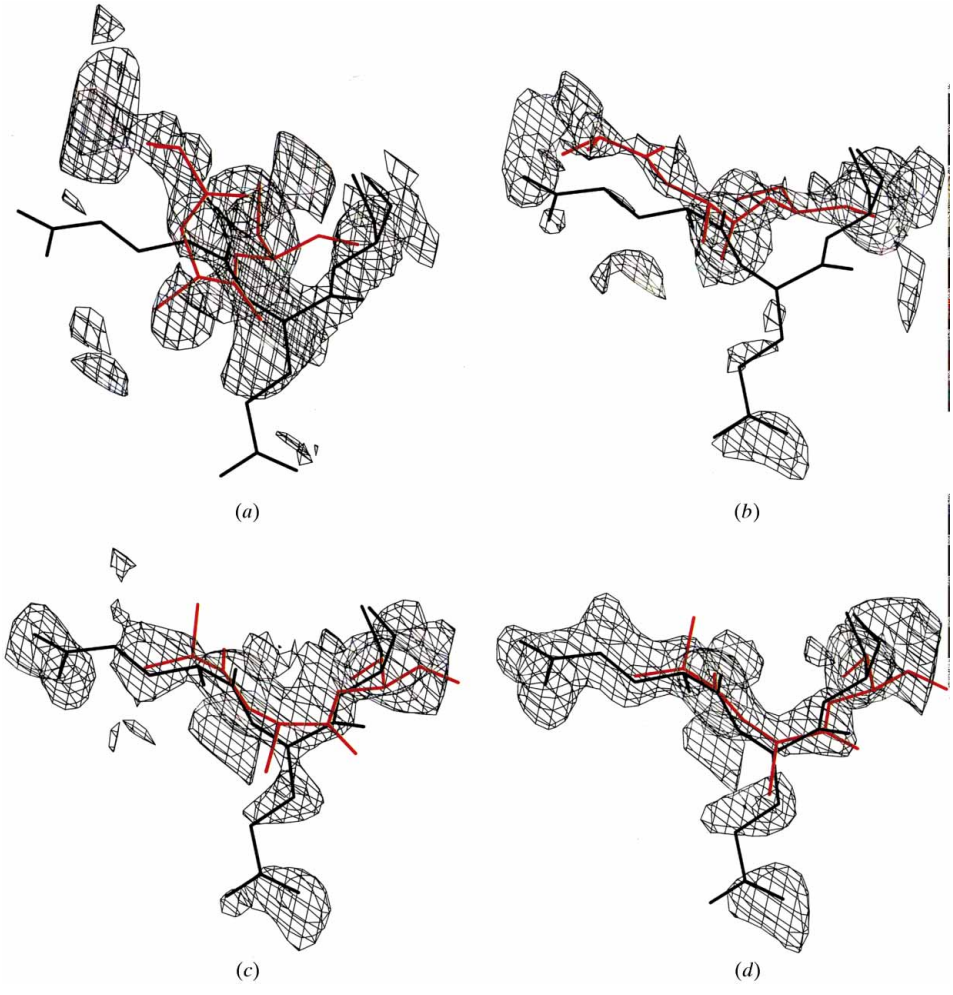

Figure 11
Maximum-likelihood targets combined with simulated-annealing refinement result in significantly better models and electron-density maps. Cross-validated σ_A-weighted electron-density maps contoured at 1.25σ for models from simulated-annealing refinement with different targets. The starting model, derived from porcine pepsin (PDB code 3PEP), consisted of a polyalanine backbone with side chains for conserved residues and had an initial r.m.s. deviation of 2.89 Å to the penicillopepsin crystal structure. Residues 50–52 are shown with the published penicillopepsin crystal structure in black, and the model with the lowest free R value from five independent refinements in red. (*a*) Initial electron-density map prior to refinement, (*b*) after refinement with the LSQ target, (*c*) after refinement with the MLF target, (*d*) after refinement with the MLHL target.

The incorporation of experimental centroid phase information into macromolecular refinement has been suggested previously (Brunger, 1988; Arnold & Rossmann, 1988). More recently it has become practical to also include bimodal or ambiguous phase information in the form of phase probability distributions within the maximum-likelihood formulation (Murshudov *et al.*, 1997; Pannu *et al.*, 1998). The use of bimodal SIR phase information dramatically improved the refinement of the polyalanine models, with map correlation coefficients up to 20% higher than refinements using the MLF target (Figs. 6 and 7*d*). The resulting electron-density maps showed significantly less model bias and a larger percentage of correctly placed atoms (Fig. 7*d*).

3.2. Refinement of polyalanine models including conserved residues

The inclusion of side chains for conserved residues in the homologous molecular-replacement search models increased the percentage of the protein content in the asymmetric unit accounted for, ranging from approximately 60% for the worst model to 80% for the penicillopepsin model. Consequently, models with higher corre-

lation coefficients were obtained as model completeness increased (Figs. 5, 9 and 10). However, those models with large coordinate errors, even though they were more complete, still produced initial maps with significant model bias (as shown, for example, for porcine pepsin in Fig. 11*a*), and the problems of overfitting during refinement with the least-squares-residual target persisted (Fig. 11*b*). The same trends as for the polyalanine case were observed. For a given target, simulated annealing always gave significantly better models than extensive conjugate-gradient energy minimization (Figs. 9 and 10) and is capable of correcting larger errors in models (Fig. 8*b*). For a given refinement optimization method (minimization or simulated annealing), maximum-likelihood targets gave better models than the least-squares residual (Figs. 11*a*, 11*b* and 11*c*).

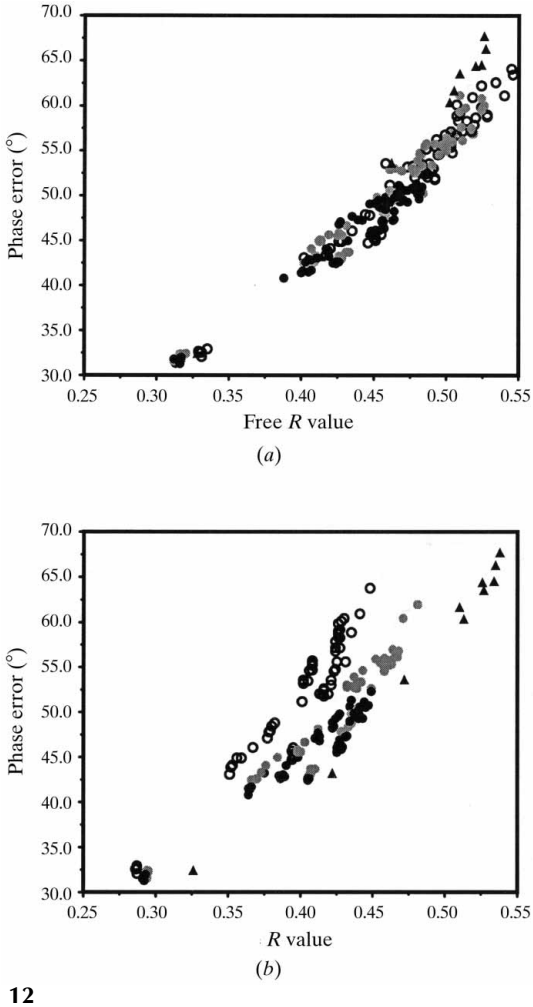

Figure 12
The free *R* value is strongly correlated with the phase error of the model, while the standard *R* value is not. Phase errors are calculated compared to the published penicillopepsin crystal structure. Circles indicate results from both minimization and simulated annealing (five trials) for each refinement target: LSQ (open white), MLF (filled grey) and MLHL (filled black) for all models tested, resulting in a total of 180 data points for each graph. Black triangles indicate initial *R* values and phase errors prior to any refinement. (*a*) Free *R* values, (*b*) *R* values calculated from the working set.

3.3. The free *R* value is a reliable indicator of phase error

Cross-validation of the *R* value (the 'free' *R* value) is an excellent indicator of the success of the refinement (Brunger, 1992; Kleywegt & Brunger, 1996). It can therefore be used to validate a molecular-replacement solution. The homologous models refined here provide an excellent demonstration of the discriminatory power of the free *R* value. There is a strong correlation between the phase error for the model after refinement and the free *R* value (Fig. 12). The standard *R* value is poorly correlated with the phase error, and is also highly dependent on the refinement target used (Fig. 12). As a result of overfitting, standard *R* values for models refined with the least-squares-residual (LSQ) target significantly underestimate the phase error of the model relative to the other targets, whereas the free *R* value is a faithful indicator of phase error for all targets.

4. Conclusions

The types of errors present in the homologous models of penicillopepsin are typical of those present in other molecular-replacement solutions: rigid-body shifts of secondary-structure elements, misthreading of residues and missing main-chain and side-chain atoms. Misthreaded residues and missing atoms are also common problems in models built into maps derived from experimental phases. The series of models generated in this paper therefore represent a realistic test for state-of-the-art crystallographic refinement methods. Of the methods tested, the best for refinement of a model derived from molecular replacement, in the absence of experimental phase information, is simulated annealing using the MLF target. If experimental phase information is available it can dramatically improve the refinement, even if phase probability distributions are bimodal. Thus, in the case of a difficult molecular-replacement solution with a distantly related homologous model, even a single isomorphous derivative is useful. The inclusion of this information in the refinement will significantly decrease the manual intervention required, allowing productive refinement of difficult molecular-replacement solutions even when model building into electron-density maps is stalled. We have shown here that the application of new advances in refinement methodology significantly extend the scope of molecular replacement as a phasing method.

We would like to thank Dr R. W. Grosse-Kunstleve, Dr Y. Shamoo and Mr L. M. Rice for critical readings of this manuscript and Dr M. N. James for making the experimental data for penicillopepsin available. This work was supported by a grant from the National Science Foundation, ASC 93–181159 to ATB, a grant from the Natural Sciences and Engineering Research Council of Canada to NSP and grants from the Alberta Heritage Foundation for Medical Research and the Medical Research Council of Canada, MT11000 to RJR.

References

Abad-Zapatero, C., Rydel, T. J. & Erickson, J. (1990). *Proteins Struct. Funct. Genet.* **8**, 62–81.

Adams, P. D., Pannu, N. S., Read, R. J. & Brunger, A. T. (1997). *Proc. Natl Acad. Sci. USA*, **94**, 5018–5023.

Arnold, E. & Rossmann, M. G. (1988). *Acta Cryst.* A**44**, 270–282.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F. Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.

Blundell, T. L., Jenkins, J. A., Sewell, B. T., Pearl, L. H., Cooper, J. B., Tickle, I. J., Veerapandian, B. & Wood, S. P. (1990). *J. Mol. Biol.* **211**, 919–941.

Bricogne, G. (1997). *Methods Enzymol.* **276**, 361–423.

Brunger, A. T. (1988). *J. Mol. Biol.* **203**, 803–816.

Brunger, A. T. (1990). *Acta Cryst.* A**46**, 46–57.

Brunger, A. T. (1992). *Nature (London)*, **355**, 472–474.

Brunger, A. T., Adams, P. D. & Rice, L. M. (1997). *Structure*, **5**, 325–336.

Brunger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Acta Cryst.* D**54**, 905–921.

Brunger, A. T., Kuriyan, J. & Karplus, M. (1987). *Science*, **235**, 458–460.

Delano, W. L. & Brunger, A. T. (1995). *Acta Cryst.* D**51**, 740–748.

Fuginaga, M. & Read, R. J. (1987). *J. Appl. Cryst.* **20**, 517–521.

Gros, P., Betzel, C., Dauter, Z., Wilson, K. S. & Hol, W. G. J. (1989). *J. Mol. Biol.* **210**, 347–367.

Grosse-Kunstleve, R. W. & Brunger, A. T. (1999). In preparation.

Hauptman, H. (1982). *Acta Cryst.* A**38**, 289–294.

Holm, L., Ouzounis, C., Sander, C., Tuparev, G. & Vriend, G. (1992). *Protein Sci.* **1**, 1691–1698.

Hoppe, W. (1957). *Acta Cryst.* **10**, 750–751.

Huber, R. (1965). *Acta Cryst.* A**19**, 353–356.

Jiang, J.-S. & Brunger, A. T. (1994). *J. Mol. Biol.* **243**, 100–115.

Kleywegt, G. J. & Brunger, A. T. (1996). *Structure*, **4**, 897–904.

Kleywegt, G. J. & Jones, T. A. (1993). *CCP4/ESF-EACBM Newslett. Protein Crystallogr.* **28**, 56–9.

Leahy, D. J., Axel, R. & Hendrickson, W. A. (1992). *Cell*, **68**, 1145–1162.

Müller, T., Oehlenschläger, F. & Buehner, M. (1995). *J. Mol. Biol.* **247**, 360–372.

Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.

Navaza, J. & Vernoslova, E. (1995). *Acta Cryst.* A**51**, 445–449.

Newman, M., Watson, F., Roychowdhury, P., Jones, H., Badasso, M., Cleasby, A., Wood, S. P., Tickle, I. J. & Blundell, T. L. (1993). *J. Mol. Biol.* **230**, 260–283.

Pannu, N. S., Murshudov, G. N., Dodson, E. J. & Read, R. J. (1998). *Acta Cryst.* D**54**, 1285–1294.

Pannu, N. S. & Read, R. J. (1996). *Acta Cryst.* A**52**, 659–668.

Pattabiraman, N., Namboodiri, K., Lowrey, A. & Gaber, B. P. (1990). *Protein Seq. Data Anal.* **3**, 387–405.

Read, R. J. (1986). *Acta Cryst.* A**42**, 140–149.

Read, R. J. (1997). *Methods Enzymol.* **277**, 110–128.

Rice, L. M. & Brunger, A. T. (1994). *Proteins Struct. Funct. Genet.* **19**, 277–290.

Rice, L. M., Shamoo, Y. & Brunger, A. T. (1998). *J. Appl. Cryst.* **31**, 798–805.

Rossmann, M. G. & Blow, D. M. (1962). *Acta Cryst.* A**15**, 24–51.

Schmidt, R., Gerstein, M. & Altman, R. B. (1997). *Protein Sci.* **6**, 246–248.

Sielecki, A. R., Fujinaga, M., Read, R. J. & James, M. N. (1991). *J. Mol. Biol.* **219**, 671–692.

Steigemann, W. (1974). PhD thesis, Technische Universität München.

Suguna, K., Bott, R. R., Padlan, E. A., Subramanian, E., Sheriff, S., Cohen, G. H. & Davies, D. R. (1987). *J. Mol. Biol.* **196**, 877–900.

Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). *Nucleic Acids Res.* **22**, 4673–4680.

Turkenburg, J. P. & Dodson, E. J. (1996). *Curr. Opin. Struct. Biol.* **6**, 604–610.