Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Real-space protein-model completion: an inverse-kinematics approach

Henry van den Bedem,[a]* Itay Lotan,[b] Jean-Claude Latombe[b] and Ashley M. Deacon[a]*

[a]Joint Center for Structural Genomics, Stanford Synchrotron Radiation Laboratory, SLAC, 2575 Sand Hill Road, Menlo Park, CA 94025, USA, and [b]Department of Computer Science, Stanford University, Stanford, CA 94305, USA

Correspondence e-mail: vdbedem@slac.stanford.edu, adeacon@slac.stanford.edu

Received 26 July 2004
Accepted 11 October 2004

Rapid protein-structure determination relies greatly on software that can automatically build a protein model into an experimental electron-density map. In favorable circumstances, various software systems are capable of building over 90% of the final model. However, completeness falls off rapidly with the resolution of the diffraction data. Manual completion of these partial models is usually feasible, but is time-consuming and prone to subjective interpretation. Except for the N- and C-termini of the chain, the end points of each missing fragment are known from the initial model. Hence, fitting fragments reduces to an inverse-kinematics problem. A method has been developed that combines fast inverse-kinematics algorithms with a real-space torsion-angle refinement procedure in a two-stage approach to fit missing main-chain fragments into the electron density between two anchor points. The first stage samples a large number of closing conformations, guided by the electron density. These candidates are ranked according to density fit. In a subsequent refinement stage, optimization steps are projected onto a carefully chosen subspace of conformation space to preserve rigid geometry and closure. Experimental results show that fitted fragments are in excellent agreement with the final refined structure for lengths of up to 12–15 residues in areas of weak or ambiguous electron density, even at medium to low resolution.

1. Introduction

The Protein Structure Initiative (PSI), a National Institute of General Medical Sciences program in the US, aims to reduce the time and associated costs of determining a three-dimensional protein structure. Stimulated in part by funding initiatives such as the PSI, the experimental and computational methods used for X-ray structure determination have been greatly improved. Many of the sample-preparation steps including protein expression, purification and crystallization have been automated and turned into large-scale production facilities (Lesley *et al.*, 2002). Various third-generation synchrotrons now feature highly automated protein crystallography beamlines and allow collection of a complete X-ray diffraction data set in a matter of minutes (Walsh *et al.*, 1999; Cohen *et al.*, 2002; van den Bedem *et al.*, 2003). Such developments require an ever-increasing rate at which macromolecular structures need to be solved. Further automation of all computational aspects of structure determination is therefore highly desirable to avoid it becoming a rate-limiting step (Burley *et al.*, 1999; Adams *et al.*, 2003).

There have also been tremendous advances in automated model-building methods. Various software systems are now capable of building a protein model into an electron-density map without human intervention (Ioerger & Sacchettini, 2003; Levitt, 2001; Perrakis *et al.*, 1999; Terwilliger, 2003). The *PHENIX* project (Adams *et al.*, 2002) aims to automate structure solution from reduced intensity data to a refined model, even at medium to low resolution. Indeed, in favorable cases it is now possible to proceed from diffraction data to an initial model of a new protein structure in a few hours.

However, the degree of completeness of these initial models, *i.e.* the fraction of atoms or residues correctly placed, varies widely depending on the quality of the experimental data and rarely reaches 100%. Determining the atomic coordinates of mobile fragments in the molecule, for instance, remains a challenge. Such fragments may lead to disorder in the crystal, rendering interpretation of the resulting electron density difficult. Manually completing a partial protein model, *i.e.* building the missing residues, is a time-consuming and labor-intensive process which can take a few weeks of work depending on the resolution and size of the structure. Thus, this step still presents a substantial bottleneck to any high-throughput structure-determination effort.

In practice, a large portion of the molecule has often been resolved and the N- and C-termini of a fragment in the initial model are known. The missing main-chain fragment can be modeled as a kinematic chain, with rigid groups of atoms as links and rotatable bonds as joints. Fitting a fragment between two anchor points can thus be interpreted as an inverse-kinematics (IK) problem (Craig, 1989; Manocha & Zhu, 1994; Manocha *et al.*, 1995): given the position and orientation of the end point of a kinematic chain, can the corresponding values of the joint angles be determined? Exploiting this observation, we have combined inverse-kinematics algorithms with a real-space torsion-angle refinement procedure in a two-stage approach to fit a missing main-chain fragment into a protein model.

In a test set of 103 structurally diverse fragments within one protein, the algorithm closed gaps of 12 residues in length to within, on average, 0.52 Å all-atom root-mean-square deviation (aaRMSD[1]) from the final refined structure at a resolution of 2.8 Å. The algorithm has also been tested and used to aid protein-model completion in areas of weak or ambiguous experimental electron density, where an initial model was built using *ARP*/*wARP* (Perrakis *et al.*, 1999) or *RESOLVE* (Terwilliger, 2003). At a resolution of 2.4 Å, it closed a ten-residue gap to within 0.43 Å aaRMSD of the final refined structure. In another case, a 14-residue gap in a 51% complete model built at 2.6 Å was closed to within 0.9 Å aaRMSD. Furthermore, our method was used to correctly identify and build multiple alternative main-chain conformations at a resolution of 1.8 Å.

[1] The algorithm fits a main chain consisting of {N, C^{α}, C^{β}, C, O} atoms. aaRMSD is the square root of the averaged squared distances between all corresponding atoms. It is calculated after the loops are optimally aligned in three dimensions.

2. Background

A variety of techniques have found successful application and widespread use in automated interpretation of electron-density maps. The program *ARP*/*wARP*, for instance, iterates interpretation of the electron-density map, model building and refinement using a hybrid model consisting of a conventional protein model and a set of free atoms (Morris *et al.*, 2002). The program *TEXTAL* (Ioerger & Sacchettini, 2003) employs local pattern-recognition techniques to select regions from a database of previously determined structures, similar to those in the unknown structure. Some automated systems, targeting lower resolution levels, notably *RESOLVE* and *MAID* (Levitt, 2001), start by identifying larger secondary-structure elements using sophisticated template-matching techniques and then connect these 'fits' through loop regions.

Relying on unambiguous experimental data and elementary stereochemical constraints, areas of weak or ambiguous electron density remain a challenge for these approaches. For instance, exposed mobile-loop regions typically have poorly resolved side-chain density or show discontinuous main-chain density even at low contour levels. Patterns in the density may go unnoticed in template-matching techniques for a variety of reasons. The electron density may exhibit multimodal disorder, in which the protein main chain adopts two or more distinct conformations for a number of contiguous residues (Wilson & Brunger, 2000). Nevertheless, at high resolution these programs may provide over 90% of the protein main chain of the final model (Badger, 2003). At resolution levels beyond 2.3 Å, the initial model resulting from these programs is typically a gapped polypeptide chain and only about two-thirds completeness is attained in the range $2.3 \leq d < 2.9$ Å. In the majority of cases, the amino-acid sequence is correctly assigned, so gap lengths and the identity of their residues are known.

In practice, to complete a model the crystallographer manually builds the missing residues onto the partially completed structure using an interactive graphics program. These programs, such as the *X-BUILD* package in *QUANTA*, *Insight*II (both from Accelrys Inc.) and *O* (Jones & Kjeldgaard, 1997) provide a variety of semi-automated tools and techniques to assist the model-building and refinement steps. In *O*, database fragments straddling a gap can be refined against the density using torsion-angle refinement based on grid summation (Jones *et al.*, 1991). Oldfield (2001) developed a method combining a random search of conformation space with grid- and gradient-based refinement techniques to close loops. *Insight*II employs the random-tweak algorithm (Fine *et al.*, 1986; Shenkin *et al.*, 1987) to build fragments.

In robotics, it is well known that for manipulators in a three-dimensional workspace there are a finite number of solutions to the IK problem when the number of degrees of freedom (DoFs) does not exceed six. In the case of a serial manipulator with six revolute joints, which is the most relevant to protein fragments, an analytic solution exists and the number of unique solutions is at most 16 (Raghavan & Roth, 1989).

Gō and Scheraga were the first to study analytical loop closure, limited to six DoFs, in the context of macromolecules

(Gō & Scheraga, 1970). Practical applications of their method and subsequent improvements (Wedemeyer & Scheraga, 1999) are limited; when restricting the DoFs to φ, ψ angles, the loop length can not exceed three residues. Recently, this limitation was overcome by extending the domain to any three, not necessarily consecutive, residues with arbitrary geometry (Coutsias *et al.*, 2004).

In the general case of $N > 6$ dihedral angles, the chain has redundant DoFs; the inverse-kinematics system of equations is underdetermined. Rather than solving directly for the dihedral angles, numerical methods are employed to sample conformational space.

Search methods sample from a set of conformational parameters and include sampling biased by the database distribution of the φ/ψ-angle pairs (Moult & James, 1986), uniform conformational search (Bruccoleri & Karplus, 1987), sampling from a discrete set of φ/ψ pairs (Deane & Blundell, 2000; DePristo *et al.*, 2003) or sampling from a small library of short representative fragments (Kolodny *et al.*, 2005). Extracting candidate fragments from the PDB satisfying conditions on length and geometry started with Jones & Thirup (1986) and was further developed by Fidelis *et al.* (1994), van Vlijmen & Karplus (1997) and Du *et al.* (2003). Various methods exist for optimization of candidate loops, such as molecular dynamics (Bruccoleri & Karplus, 1987; Fiser *et al.*, 2000; Zheng *et al.*, 1992) and Monte Carlo (Abagyan & Totrov, 1994; Collura *et al.*, 1993) simulations.

Another class of methods iteratively solves the inverse-kinematics system of equations. The aforementioned random-tweak method closes a loop by iteratively changing all its DoFs at once until the desired distances between the two termini are reached. It employs the Jacobian of these distances with respect to torsional DoFs to calculate the DoF changes. The cyclic coordinate descent (*CCD*) algorithm (Canutescu & Dunbrack, 2003; Wang & Chen, 1991) adjusts one DoF at a time along the chain to move the final segment of the loop toward the target residue. It is free from singularities and allows constraints on any of the DoFs.

3. Methods

The objective is to automatically fit a missing protein fragment in between two anchor residues, satisfying electron-density constraints. The algorithm assumes rigid peptide geometry; residue-dependent values for bond lengths and bond angles are derived from small-molecule data (Engh & Huber, 1991). To limit the number of DoFs (and thus computational complexity) in the current implementation, side chains are truncated at the C^β atom.

The algorithm proceeds in two stages: candidate generation and refinement. In the first stage, 1000 candidate gap-closing fragments are built using the *CCD* algorithm, while putting additional constraints on the DoFs to take electron density and collision avoidance into account.[2] Next, a cross-correlation density score $r = \sum \overline{\rho^o}\,\overline{\rho^c} \left(\sum \overline{\rho^o}^2 \sum \overline{\rho^c}^2\right)^{-1/2}$ is calculated for these candidates, where $\overline{\rho^o}$ and $\overline{\rho^c}$ denote the normalized observed and calculated density, respectively. The 99th percentile (with a maximum of six fragments) is passed on to stage two, which refines atomic coordinates by minimizing a standard real-space target function (Diamond, 1971; Chapman, 1995; Korostelev *et al.*, 2002). An optimization protocol based on simulated annealing (SA; Kirkpatrick *et al.*, 1983) and Monte Carlo minimization (MCM; Li & Scheraga, 1987) uses the redundant DoFs of the fragment to search for the global minimum of the target function while maintaining ideal peptide geometry and loop closure. Each fragment is subjected to four SA refinement cycles, the two top-scoring fragments of which are retained.

The input to the algorithm is given by the electron density, in most cases a $2mF_o - DF_c$ map (Read, 1986), the partial model and the amino-acid sequence. The program outputs all 12 fragments it retains. It also writes a log file containing the full cross-correlation electron-density score for each fragment. Final conformations will need to be refined using standard maximum-likelihood refinement programs such as *CNS* (Brünger *et al.*, 1998) or *REFMAC* (Murshudov *et al.*, 1997). The implementation of the algorithm uses the following software packages: Clipper (Cowtan, 2004), the CCP4 Coordinate Library (Krissinel, 2004) and the exact IK solver of Coutsias *et al.* (2004).

3.1. Stage 1: generation

Residues flanking the gap in the partial model are denoted N- and C-stationary anchors. The algorithm starts by constructing a protein fragment $\mathcal{C}$ of length L in a random initial conformation (§3.1.1), where residue 0 is a copy of the N-stationary anchor and residue $L - 1$ is a copy of the C-stationary anchor. This chain is attached to either the N- or C-anchor, thus determining the closing direction. The remaining terminal residue in $\mathcal{C}$ is called the mobile anchor.

Upon starting the procedure, the position of the mobile anchor will not coincide with the position of the stationary anchor. The algorithm iteratively adjusts each backbone dihedral angle in turn to satisfy a closure constraint, minimizing the distance between the three backbone atoms of the mobile anchor and the corresponding atoms of the stationary anchor as follows. Working its way down the chain, at residue i the *CCD* algorithm proposes a dihedral angle φ_i that minimizes the distance between the mobile anchor and stationary anchor. Based on φ_i, it also proposes a minimizing angle ψ_i. (In our implementation, we change each DoF in turn, although this is not strictly necessary.) Thus, a proposed angle pair $(\varphi, \psi)_i^p$ is obtained. To guide the fragment, a heuristic electron-density constraint has been added to the *CCD* algorithm. For $(\varphi, \psi)_i^p$, denote by A_i the set of atoms $\{C_i^\beta, C_i, O_i, N_{i+1}, C_{i+1}^\alpha\}$ subject to change by this angle pair, but not affected by changes in angle pair $i + 1$. Electron-density scores are calculated for trial conformations in a square neighborhood $U_{(\varphi,\psi)_i^p}$ about $(\varphi, \psi)_i^p$. A simple and fast local scoring function is used: the sum of electron-density values at atom-center

[2] In future releases, the number of candidate fragments will depend on the length of the fragment and the quality of the electron density.

positions of A_i. The angle pair $(\varphi, \psi)_i$ is then set to the trial position with maximum score. At this point, overlaps of van der Waals surfaces of atoms in A_i and the rest of the protein structure are determined. If no overlaps occur, the new $(\varphi, \psi)_i$ pair is accepted, otherwise the pair is accepted with a probability inversely related to the amount of overlap. The size of $U_{(\phi,\psi)^p}$ is reduced linearly in the number of CCD iterations to allow closure of the chain.

It was found that longer fragments fit the electron density better when built from both the N- and C-stationary anchors to meet in the middle. Fragments of nine or more residues are therefore split in the middle and each half-chain is attached to its corresponding anchor. The terminal residue of each half-chain alternates between acting as stationary anchor and mobile anchor in subsequent iterations.

Each initial conformation is allowed 2000 iterations for closure, up to a preset tolerance distance $d_{\rm closed}$. Chains that do not close are discarded.

3.1.1. Random initial conformations. For each initial conformation, ω_i is considered to be a fixed $N(180, 5.8)$ random variable for all i. Half of the initial conformations are obtained by adjusting each $(\varphi, \psi)_i$ in turn to optimize agreement with the electon density while stereochemical constraints are observed. The remaining 500 initial conformations are purely random and obtained from sampling $(\varphi, \psi)_i$, $i = 0 \ldots L - 1$ angle pairs from PDB-derived distributions. Finite mixtures of bivariate normal distributions were therefore fitted to frequencies calculated from the Top500 database (Lovell *et al.*, 2003) of non-redundant protein structures using the program *EMMIX* (McLachlan *et al.*, 1999). We obtained distributions for each of the 20 amino acids and an additional distribution for residues immediately preceding proline in the amino-acid sequence. The angles φ_0 and ψ_{L-1} remain fixed at their initial values.

3.2. Stage 2: refinement

A candidate fragment is refined by minimizing the least-squares residuals between the observed density $\rho^{\rm o}$ and the density calculated from the model $\rho^{\rm c}$ in some volume V around the fragment,

$$T(q) = \sum_{g_i \in V} [S\rho^{\rm o}(g_i) + k - \rho^{\rm c}(g_i)]^2. \quad (1)$$

The calculated density at each grid point is a sum of contributions of all atoms whose center lies within a cutoff distance from this point. The calculated density contribution of an atom is a sum of isotropic three-dimensional Gaussians (Waasmaier & Kirfel, 1995). The factors S and k scale $\rho^{\rm o}$ to $\rho^{\rm c}$ and are computed once at initialization using the partial model.

3.2.1. Optimization with closure constraints. The redundant DoFs define a subspace of conformation space termed the self-motion manifold. Motions on this manifold do not influence the position and orientation of the end point and thus can be used to move the fragment towards a minimum of the target function (Burdick, 1989; Khatib, 1987). Since this manifold may be very complex, these motions are in general difficult to calculate. We therefore use a local linear approximation of the self-motion manifold; the null-space of the Jacobian matrix of the fragment (Craig, 1989). For an n-DoF fragment in $\mathbb{R}^3$ at conformation q, the Jacobian $J(q)$ is a $6 \times n$ matrix satisfying the equation

$$\dot{x} = J(q)\dot{q}. \quad (2)$$

Thus, $J(q) = {\rm d}f(q)/{\rm d}(q)$, where $f(q)$ is the fragment's forward-kinematics function mapping DoF parameters to end-point position and orientation. The rank of the Jacobian in $\mathbb{R}^3$ is at most 6 and thus the dimensionality of its null space is at least $n - 6$. An instantaneous change in the conformation corresponding to a desired small change in end-point position is calculated by inverting (2). We obtain

$$\dot{q} = J^{\dagger}(q)\dot{x} + N(q)N^T(q)y, \quad (3)$$

where $J^{\dagger}$ is the pseudo-inverse of the Jacobian and $N(q)$ is an orthonormal basis for the null space. The null space can now be used to optimize the target function without affecting the position of the end point. The instantaneous change in position and orientation of the end point, $\dot{x}$, is set to zero and y is taken to be the gradient vector of the target function. Projecting y onto the null space of the Jacobian produces a motion that minimizes the target function without disturbing closure.

3.2.2. Implementation details. The null space of the Jacobian is obtained from a singular-value decomposition of the Jacobian matrix. The null-space basis $N(q)$ is the set of right singular vectors corresponding to vanishing singular values. We derived an analytical expression for the gradient of the target function with respect to the torsional DoFs of the loop. It is calculated using a recursive method (Abe *et al.*, 1984) linear in the number of DoFs of the fragment.

A gradient-descent search for the minimum of the target function is prone to become stuck in local minima. The MCM approach is well known for its ability to overcome this problem. At each step, a large random move in conformation space is proposed, the new conformation is then minimized by gradient descent and the resulting local minimum is accepted or rejected using the Metropolis criterion (Metropolis *et al.*, 1953). Minimization increases the acceptance probability of the trial move, enabling the search to make more progress.

```
for start_temp = high_start_TEMP downto low_start_TEMP {
    temp = start_temp;
    SmoothDensity(start_temp);
    for SA_steps = 1 to 8 {
        for MCM_steps = 1 to NUM_ITERS {
            M = ProposeRandomMove(temp);
            MinimizeMove(M);
            AcceptMove(M);
        }
        temp *= TEMP_dec_factor;
    }
}
```

Figure 1
Pseudo-code for refinement-search protocol.

This comes at the cost of increasing the time of each simulation step.

Two methods are used for generating random moves for MCM. The first is to take a step in a random direction in the null space (Yakey *et al.*, 2001). Before performing minimization, we make sure the closure tolerance has not been exceeded. A second method for generating random steps is an exact IK solver (Coutsias *et al.*, 2004). One of the solutions is chosen at random as the proposed move. The use of an exact solver allows jumping between unconnected parts of the self-motion manifold. The closure constraint is relaxed during the refinement stage and a maximum RMSD of 0.5 Å is allowed at both ends of the loop. By relaxing closure, larger steps can be taken in the null space of the Jacobian.

The refinement protocol is composed of three nested loops, see Fig. 1. The inner loop performs an MCM search by using the two methods described above for generating random trial moves. The middle loop performs SA by gradually reducing the pseudo-temperature of the MCM search. The outer loop enhances the SA protocol by simulating restarts each time at a lower starting pseudo-temperature. The magnitude of attempted random null-space moves is reduced together with the current pseudo-temperature of the simulation to increase the chance that the random moves will be accepted. Decreasing levels of smoothing are applied to the density after each restart. The density map is smoothed by convolving it with an isotropic three-dimensional Gaussian kernel.

4. Results and discussion

The performance of the algorithm was first evaluated on a test set of structurally diverse fragments at various truncated resolution levels. Next, we tested its ability to fully complete initial protein models at comparable resolution levels by closing all the gaps in three initial models, this time using 'real' data. Furthermore, we evaluated the algorithm's ability to identify alternative conformations in a disordered region.

4.1. Performance at various resolutions, fragment lengths and their secondary structure

4.1.1. TM1621. A set of 103 structurally diverse fragments was obtained by creating gaps of length four, eight, 12 and 15 at each even-numbered residue of a test structure, the protein TM1621 (PDB code 1o1z; SCOP classification α/β). TM1621 consists of one chain, with 34% of the residues in ten α-helices

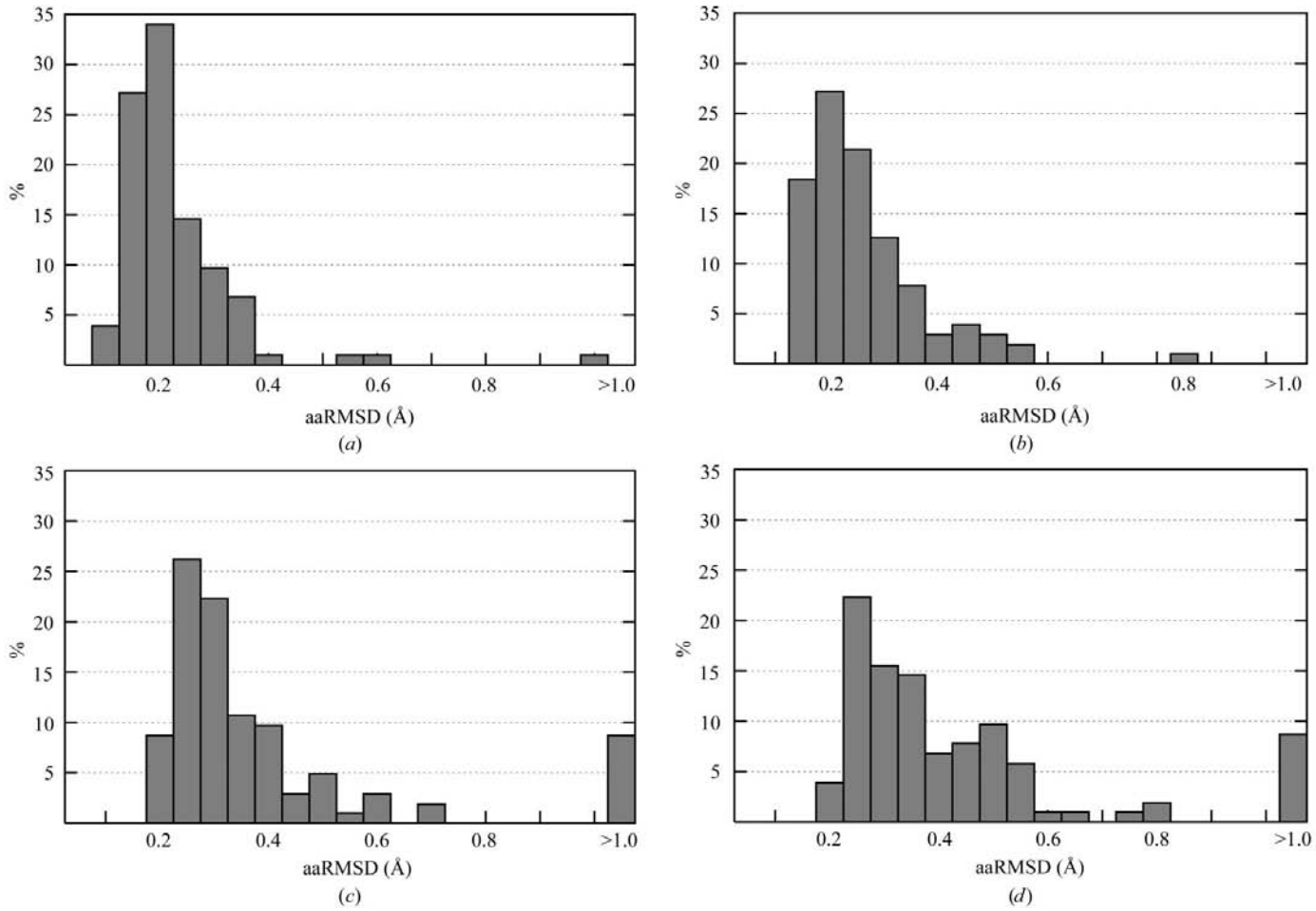

Figure 2
The aaRMSD distribution of 103 fragments with lengths of (*a*) four, (*b*) eight, (*c*) 12 and (*d*) 15 residues of TM1621 at a resolution of 2.0 Å. A total of 9% of 12-residue and 9% of 15-residue fragments have an aaRMSD > 1.0 Å.

and 19% in nine β-sheets. Diffraction data for this 234-residue protein structure had been collected at a resolution of 1.6 Å. To evaluate the performance at various resolution levels, three $2mF_o - DF_c$ electron-density maps were calculated at 2.0, 2.5 and 2.8 Å, using structure factors obtained from the PDB. For each gap, the fragment with the highest cross-correlation electron-density score was selected from the 12 fragments output by the program. Since low-resolution electron-density maps were obtained by truncation, the RMSDs in this section are not typical for their resolution levels.

At a resolution of 2.0 Å, the algorithm successfully closed all 103 gaps of length four to within 1.0 Å and all length eight gaps to within 0.85 Å, as shown in Fig. 2. Wider gaps are more difficult to close; a total of nine 12-residue and nine 15-residue fragments were found to have an aaRMSD greater than 1.0 Å.

To evaluate the effect of secondary structure on aaRMSD, all 12- and 15-residue fragments were classified as helix, strand or 'other'. A fragment is considered a helix or strand only if at least two-thirds of its residues are classified as such. A total of 14 12-residue fragments and eight 15-residue fragments met our criteria for helices. Three 12-residue fragments and no 15-residue fragments were classified as strands. The maximum aaRMSD for the 12-residue strands over all resolutions was 0.3 Å. 4% of non-helical 12-residue fragments were found to have an aaRMSD > 1.0 Å, compared with 36% of helical fragments. For 15-residue fragments, these numbers are 4 and 63%, respectively.

Table 1
Median ($\tilde{x}$) and mean ($\bar{x}$) aaRMSD of fitted fragments to corresponding regions in TM1621 at resolutions of 2.0, 2.5 and 2.8 Å and the percentage of fragments deviating by more than 1.0 Å (p).

	2.0 Å			2.5 Å			2.8 Å		
Length	$\tilde{x}$	$\bar{x}$	p	$\tilde{x}$	$\bar{x}$	p	$\tilde{x}$	$\bar{x}$	p
4	0.13	0.14	0	0.18	0.19	0	0.31	0.32	0
8	0.16	0.18	0	0.23	0.23	0	0.33	0.36	0
12	0.28	0.51	9	0.34	0.41	4	0.41	0.52	4
15	0.33	0.53	9	0.43	0.63	12	0.49	0.76	17

At a resolution of 2.5 Å, all gaps of length four and eight were closed to within 1.0 Å aaRMSD and 0.85 Å aaRMSD, respectively, whereas four 12-residue fragments and 12 15-residue fragments deviated by more than 1.0 Å aaRMSD. The results are depicted in Fig. 3. 1% of non-helical 12-residue fragments were found to have an aaRMSD > 1.0 Å, compared with 21% of helical fragments. For 15-residue fragments, these numbers are 7 and 63%, respectively.

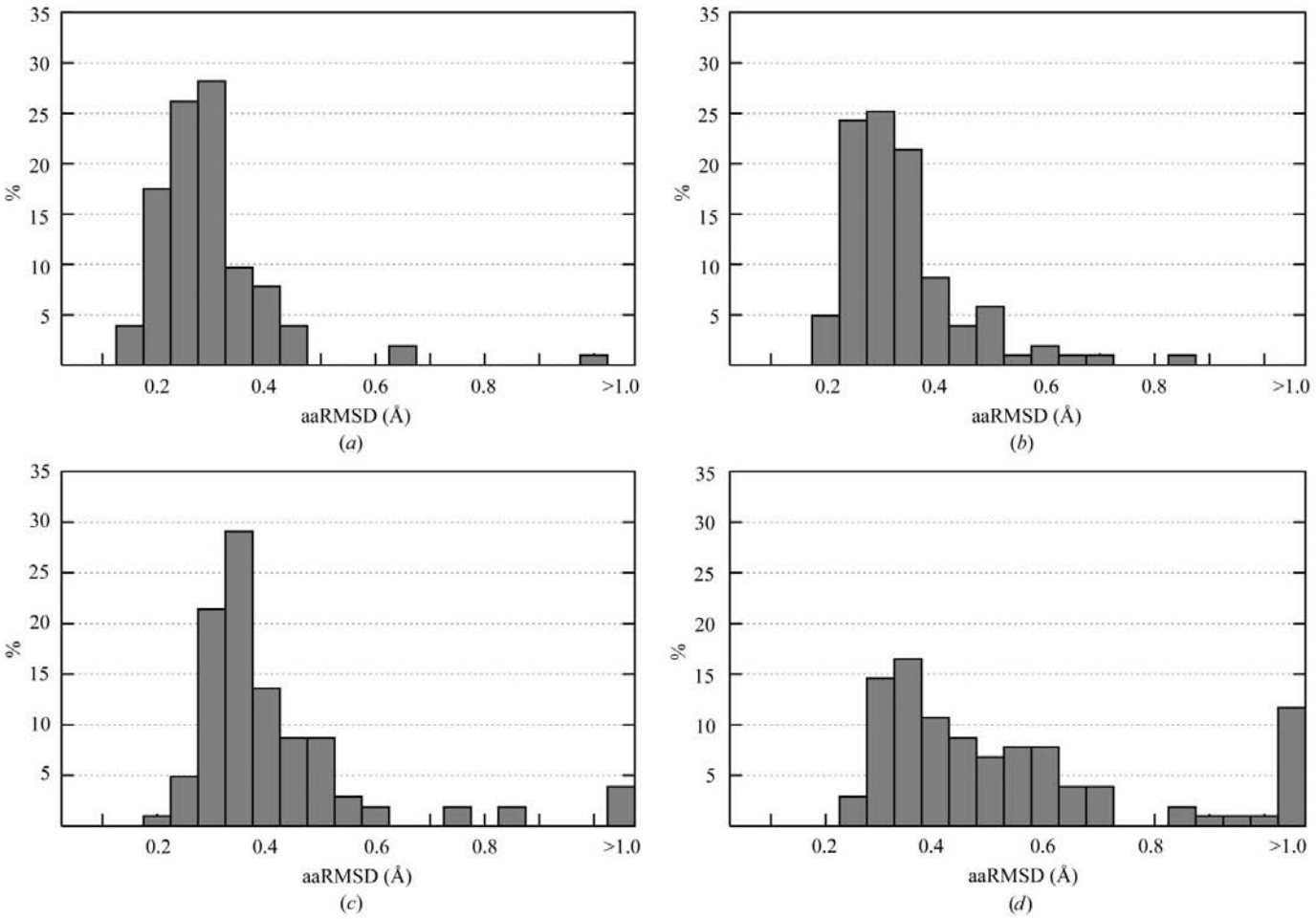

Figure 3
The aaRMSD distribution of 103 fragments with lengths of (*a*) four, (*b*) eight, (*c*) 12 and (*d*) 15 residues of TM1621 at a resolution of 2.5 Å. A total of 4% of fragments of length 12 and 12% of fragments of length 15 have an aaRMSD > 1.0 Å.

Table 2
Average run times (min) on a 2.66 GHz Intel P4 Xeon at various fragment lengths and resolution levels.

The average is calculated over 103 fragments.

	Resolution (Å)		
Length	2.0	2.5	2.8
4	40	29	28
8	92	63	58
12	134	82	73
15	178	105	95

Table 3
Median ($\tilde{x}$) and mean ($\bar{x}$) aaRMSD of 174 fitted fragments to corresponding regions in TM0423 at resolutions of 2.0, 2.5 and 2.8 Å and the percentage of fragments deviating by more than 1.0 Å (p).

	2.0 Å			2.5 Å			2.8 Å		
Length	$\tilde{x}$	$\bar{x}$	p	$\tilde{x}$	$\bar{x}$	p	$\tilde{x}$	$\bar{x}$	p
4	0.18	0.19	0	0.24	0.25	0	0.32	0.32	0
8	0.20	0.22	0	0.28	0.29	0	0.35	0.38	0
12	0.29	0.55	26	0.33	0.50	19	0.40	0.56	19
15	0.34	0.96	38	0.43	0.92	29	0.52	1.19	29

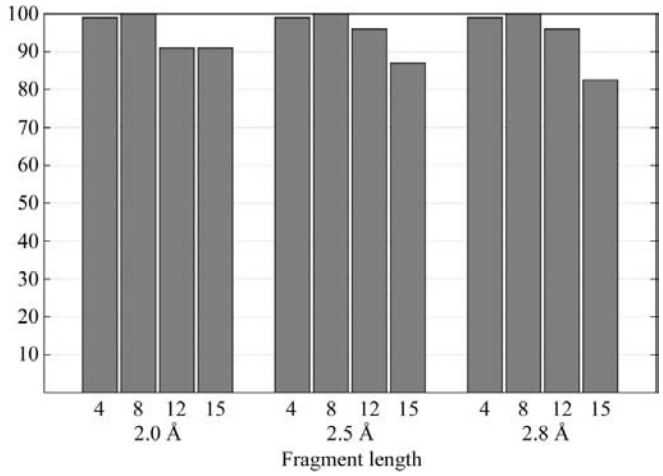

Figure 5
Percentage of fragments closed to within 1.0 Å aaRMSD of the native structure by resolution level.

At a resolution of 2.8 Å, all gaps of length four and eight closed to within 1.05 and 0.75 Å aaRMSD, respectively. Four 12-residue fragments and 18 15-residue fragments deviated by more than 1.0 Å aaRMSD. The results are depicted in Fig. 4.

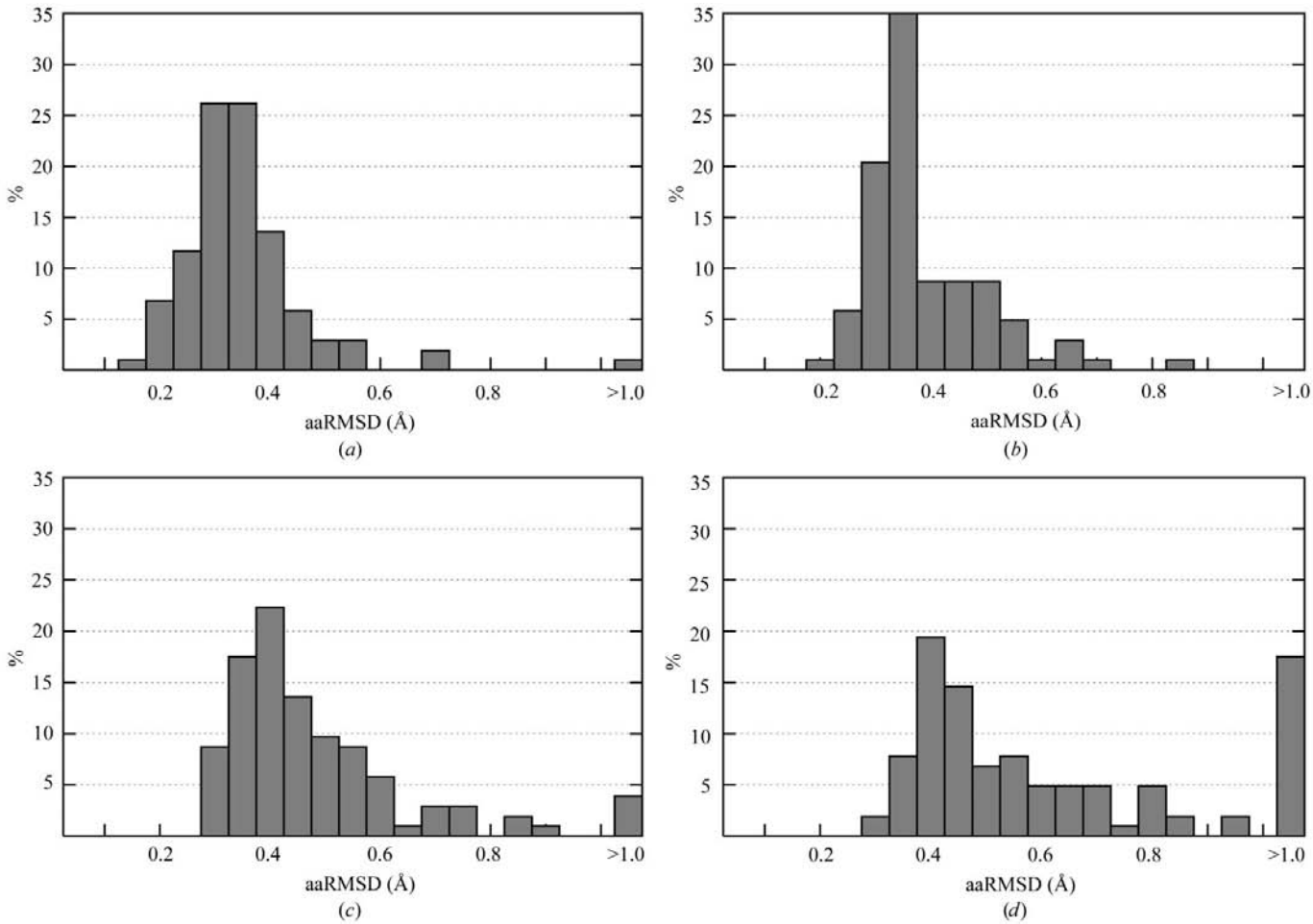

Figure 4
The aaRMSD distribution of 103 fragments with lengths of (*a*) four, (*b*) eight, (*c*) 12 and (*d*) 15 residues of TM1621 at a resolution of 2.8 Å. A total of 4% of fragments of length 12 and 17% of fragments of length 15 have an aaRMSD > 1.0 Å.

2% of non-helical 12-residue fragments were found to have an aaRMSD > 1.0 Å, compared with 14% of helical fragments. For 15-residue fragments, these numbers are 12 and 88%, respectively.

Table 1 summarizes the performance at three resolution levels.

Fig. 5 visually summarizes the performance of the algorithm at the three resolution levels. The histogram depicts the distribution of fragments closed to within 1.0 Å aaRMSD of the native structure at various fragment lengths by resolution level.

4.1.2. Run times. The run time of the algorithm depends on the length of the fragment to be fitted, as well as on the resolution of the diffraction data. Run times vary from about 30 min for short fragments to just under 3 h for the longest fragments at high resolution. Table 2 summarizes average run times calculated while generating the 103 fragments used in this section. All tests were performed on a 2.66 GHz Intel P4 Xeon running RedHat 9. The source code was compiled using gcc 3.2.

It is to be expected that targeting areas of weak or ambiguous electron density, where standard model-building algorithms fail, is computationally expensive. However, average execution times may be reduced by alternating generation and refinement stages together with the introduction of a stopping criteria based on a variety of local scores such as density fit and Ramachandran scores.

4.1.3. TM0423. An equivalent analysis on TM0423 (376 residues; PDB code 1kq3; SCOP classification multi-domain α/β, multi-helical), a protein with a helical domain, gives similar results (see Table 3). TM0423 consists of one chain, with 46% of the residues in 16 helices and 11% in eight β-sheets. The longest helix has length 17 and if a single glycine classified as a hydrogen-bonded turn is included its length is 26.

Clearly, the algorithm performs more modestly when fitting longer fragments. In addition to an increasing median aaRMSD, a larger proportion of fragments deviate by more than 1.0 Å as fragment length increases, particularly when a large number of residues are in α-helical conformation. The latter effect arises from the nature of the *CCD* algorithm; choosing distance-minimizing dihedral angles at every iteration naturally leads to an extended conformation. It has been observed in previous studies that accurately modeling secondary-structure elements may require specialized sampling algorithms (Jacobson *et al.*, 2004). Our current implementation lacks such targeted approaches, yet gives acceptable performance for fragments up to length 12 across all resolutions. For instance, at a resolution of 2.5 Å it correctly builds two out of every three fragments containing eight or more residues in a helical conformation.

Interestingly, lowering the resolution of the data only mildly affects performance (see Fig. 5). We believe that this is the true strength of the algorithm: information from electron density which is at the limit of being interpretable is

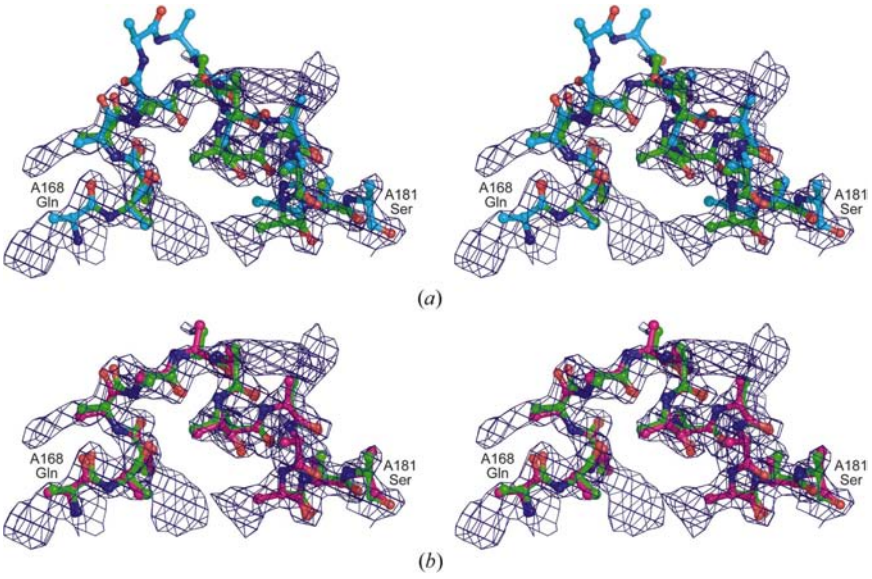

Figure 6
An example illustrating the convergence radius of the refinement stage. Residues *A*168–*A*181 of TM1621 are shown in green. (*a*) The output of the first stage shown in cyan. The aaRMSD to the native structure is 2.72 Å. (*b*) The refined fragment corresponding to (*a*) shown in magenta. The aaRMSD improved by 2.4 Å to 0.31 Å. A closed conformation is maintained throughout refinement. The 2.5 Å $2mF_o - DF_c$ electron-density map is contoured at 1.5σ.

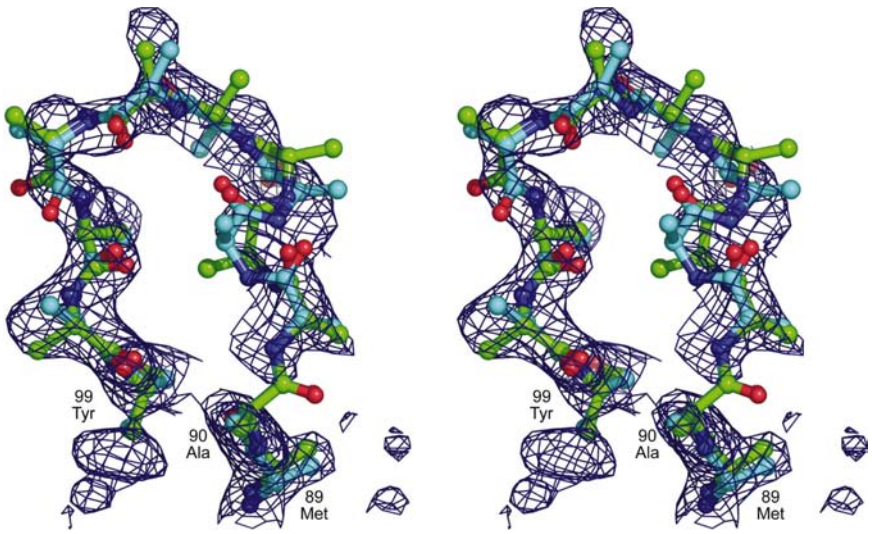

Figure 7
Residues 89–99 of TM1586. The fragment inserted into the model is shown in cyan and the corresponding final refined fragment in green. The aaRMSD between the two fragments is 1.01 Å. The $2mF_o - DF_c$ electron-density map is shown contoured at 0.8σ and is discontinuous around Ala90.

augmented by a closing constraint. The performance of the second stage in the algorithm also reflects this insight. For instance, in §4.1.1 it was found that 12-residue fragments have a mean (median) aaRMSD of 0.41 Å (0.34 Å) to the native structure at a resolution of 2.5 Å. Their corresponding stage-one fragments have mean and median distances of 1.38 and 1.23 Å, respectively. In fact, for one in five fragments, the second stage lowered the aaRMSD by more than 1.5 Å, with a few cases even exceeding 2.5 Å (see Fig. 6).

4.2. Missing fragments

In this section, we present three examples of protein-model completion by inserting main-chain fragments into a gapped initial model at high and medium-to-low resolution. Rather than closing a few selected gaps, we aimed to fully complete each model. Thus, we calculated all missing fragments of length 15 residues or less in each model.

In one instance, the hypothetical protein TM1586, the algorithm was actively used to complete the model and detailed results will appear in a separate publication. The remaining two structures had been completed and refined prior to testing the algorithm. All initial models were obtained from common crystallographic model-building programs.

It was found that residues flanking a gap in partial models do not always fit the density correctly. In these cases, the gap was widened by trimming back one or more residues at the N and/or C end of the gap until the new anchors fit the density satisfactorily. Furthermore, missing fragments of length less than four are extended to length four in this section, again by trimming back residues at both ends of the gap.

The electron-density score of generated fragments and RMSD to the final refined structure cannot expected to be perfectly correlated in areas of poor density. In an extreme case, it may happen that conformations attain a higher score by jumping over to a neighboring empty stretch of density (a β-sheet, for instance) for a few residues. In this section, in addition to the aaRMSD of the best scoring fragment, we therefore report the lowest achieved aaRMSD among the 12 fragments output by the program.

4.2.1. TM1586 at 2.0 Å. An initial model for the 206-residue hypothetical protein TM1586 was produced using *Xsolve*, a fully automated crystallographic data-processing and structure-solution software suite under development at the JCSG. *Xsolve* utilizes standard crystallographic software packages to obtain a protein model.

An experimental electron-density map was obtained from MAD data collected at 2.0 Å with the program *SOLVE* v2.03 (Terwilliger & Berendzen, 1999). The initial model, obtained with *RESOLVE* v2.06, showed gaps between residues 86–98, 107–117 and 142–150. Furthermore, 66 residues were missing at the N-terminus of the molecule. Overall completeness was reported to be 51%. After widening the gap between residues 142–150 by one residue at each end, it was easily closed to within 0.5 Å aaRMSD. The gaps between residues 86–98 and 107–117 proved to be more difficult; the density was too weak for the crystallographer to decide which fragment among the 12 candidates fitted the electron density best.

Table 4
RMSD of fitted fragments in TM1586 and corresponding regions in the final refined structure.

Gap	Length	Secondary structure	aaRMSD (Å) (top score)	aaRMSD (Å) (lowest)
13–27	13	HHHHHHHHH·B···B	2.43	2.39
47–53	5	·SS····	1.08	0.86
89–99	9	HHHHHTTEEEE	1.39	1.01
105–114	8	·BS·····	1.03	0.75
141–151	9	HT·GGGGG·	0.46	0.43

Table 5
RMSD of fitted fragments in TM1742 and corresponding regions in the final refined structure obtained from the PDB.

Gap	Length	Secondary structure	aaRMSD (Å) (top score)	aaRMSD (Å) (lowest)
17–25	7	ETTEE·T	0.72	0.66
56–62	5	HHHHT	0.78	0.78
126–132	5	HHHHH	0.36	0.36
146–148	1	·	0.44	0.40
191–202	10	HHHHHT··GG	0.43	0.43
228–233	4	SSS·	0.22	0.22

The extended *RESOLVE* model was then combined with an *ARP*/*wARP* model and after various rounds of phase improvements the N-terminus was largely recovered. The map was further slightly improved using a combination of *SHELXD* (Schneider & Sheldrick, 2002) and *autoSHARP* (de La Fortelle & Bricogne, 1997; Vonrhein *et al.*, 2005). The model still showed gaps between residues 13–23, 49–52, 89–99 and 105–113. These missing fragments were all located on one face of the molecule and the density remained weak in this area. Three residues at the C-terminus of the first gap did not adequately fit the density and the gap was widened to span residues 13–27. Gap 49–52 was widened to 47–53 and gap 105–113 was extended by one residue at the C-terminus. After generating these fragments and further manual refinement, the resulting structure was subsequently refined with *REFMAC*5. Table 4 shows the aaRMSD of fragments to this final refined model.

The density score and the aaRMSD are poorly correlated, reflecting the weak density in the area of the missing fragments. Even though the first fragment has a fairly high aaRMSD, it still provided a good starting point for manual refinement. Fig. 7 shows residues 89–99 of the final refined structure together with the best fragment that was generated. Note that the main-chain density is discontinuous at the displayed contour level of 0.8σ and that side-chain density is poorly defined.

4.2.2. TM1742 at 2.4 Å. MAD data for the 271-residue putative Nagd protein TM1742 (PDB code 1vjr) was collected at a resolution of 2.4 Å. An initial electron-density map of good quality was obtained using the program *SOLVE* v.2.03 (Terwilliger & Berendzen, 1999) at a resolution of 2.5 Å. Iterative model building using Terwilliger's `resolve_build`

Table 6
RMSD of fitted fragments in molecule *A* of TM0542 and corresponding regions in the manually built structure.

Gap	Length	Secondary structure	aaRMSD (Å) (top score)	aaRMSD (Å) (lowest)
134–142	7	HHHHHHH	0.93	0.78
212–227	14	BS··SSGGGGG·HH	0.91	0.90
256–266	9	ES·SS·SHH	0.87	0.87
272–285	12	·SSEEEEEE·SS	1.15	1.15
318–324	5	HHHHH	0.72	0.72

script resulted in an 88% complete model, with gaps between residues 17–25, 56–62, 129–132, 146–148 and 229–231. Furthermore, the region between residues 191 and 202 had been built incorrectly. The *RESOLVE* model was independently completed and refined. Table 5 summarizes the aaRMSD of top-scoring fragments built with our algorithm to the final refined structure.

4.2.3. TM0542 at 2.6 Å. MAD data for the 376-residue protein TM0542 (malate oxidoreductase) was collected at a resolution of 3.0 Å and a native data set was obtained at 2.6 Å. An electron-density map was calculated with phase extension using the program *SOLVE*. Iterative model building using *SOLVE* revealed that the unit cell contains four NCS-related molecules. Molecule *A* was the most complete of this set of four with 56% of residues placed and gaps between residues 12–89, 134–142, 212–227, 256–266, 272–285 and 318–324. This *RESOLVE* starting model was independently manually completed and refined. The refined model was used to calculate RMSDs for our automatically generated fragments.

The algorithm successfully closed all gaps up to 15 residues in length in the protein. Table 6 summarizes the results.

Fitting a main-chain fragment into the density is rather sensitive to residues being flipped along the chain. This problem is exacerbated by the fact that exposed loop regions typically have poorly resolved side chains in the electron density. Fig. 8 shows an example of a fragment where two consecutive residues are flipped. While the aaRMSD is relatively high at 0.9 Å for this fragment, the C^{α} trace is in excellent agreement with the manually built fragment. The flipped residues are easy to identify and correct for a trained crystallographer.

4.3. Identifying alternative main-chain conformations

Binding of ligands to proteins and protein–protein interactions are typically facilitated by mobile regions in the macromolecule. Such flexible fragments sometimes crystallize in multimodal disordered substates, in which the main chain adopts two or more distinct conformations for a number of contiguous residues. It is generally difficult to recognize features in the resulting areas of overlapping density, even for a trained crystallographer. Here, we show that the techniques introduced in this paper can be extended to support identification and modeling of multiple distinct conformations, even at a resolution of 1.8 Å.

A model for the 398-residue protein TM0755 (PDB code 1vme) was built from a 1.8 Å MAD data set using *ARP*/*wARP*. The structure was completed manually, apart from a short fragment around residue *A*320 and the same fragment around *B*320. The electron density from residues *A*317–*A*323 indicated that this fragment was bimodally disordered. Furthermore, a structurally similar dioxygen-reduction enzyme, rubredoxin oxygen:oxidoreductase (PDB code 1e5d), binds a flavin mononucleotide at the corresponding residues. The absence of this cofactor in TM0755 allows the main chain to adopt other energetically favored conformations. A detailed analysis of TM0755 will appear in a separate forthcoming publication.

Table 7
Main-chain (MC) and side-chain (SC) real-space correlation coefficients for residues *A*316–*A*323 of the final refined model.

Side chains were added manually.

	*A*316	*A*317	*A*318	*A*319	*A*320	*A*321	*A*322	*A*323
MC	0.97	0.92	0.95	0.91	0.87	0.94	0.94	0.97
SC	0.96	0.92	0.92	0.91	0.63	0.76	0.96	0.96

While one conformation was clearly visible in the electron density, the main-chain trace of the alternative conformation was much less obvious. From residue *A*320 to *A*323, the density was particularly ambiguous; the alternative conformation was difficult to identify and not modeled. The algorithm was slightly modified to model the fragment from residue *A*317 to *A*323; half-occupancy was hard-coded and density smoothing was disabled to narrow the radius of convergence of the refinement stage. Runs at four different lengths were attempted. The N-anchor was kept fixed at Ser*A*316 and the C-anchor ranged from Ala*A*320 to His*A*323. In the final run, four out of the final 12 fragments adopted conformation 1, another three adopted conformation 2 and the remaining five fragments did not fit the density meaningfully. The aaRMSDs between these fragments and the final refined model are 0.42 Å for conformation 1 and 0.29 Å for conformation 2. Fig. 9 depicts the final refined model of the two alternative conformations for residues *A*317–*A*321 in an omit electron-density map. In conformation 1, the terminal OH group of Tyr*A*318 is engaged in hydrogen bonds with Glu*B*29 and Lys*B*336. Glu*A*319 is hydrogen bonded to a water molecule and to Lys*B*336. Further stability is provided by a salt bridge between Glu*A*319 and Asp*B*335. Glu*A*321 is hydrogen bonded to a water molecule. In conformation 2, the main-chain rotates to occupy the empty flavin mononucleotide-binding cavity. Fig. 10 depicts the flavin mononucleotide from the enzyme rubredoxin oxygen:oxidoreductase superimposed onto the corresponding residues of its binding site in TM0755. The side chain of Tyr*A*318 rests in the hydrophobic pocket. Ramachandran analysis (Lovell *et al.*, 2003) of the final refined model showed that the dihedral angles of both conformations are all in favored (>98%) or allowed (>99.8%) regions. Real-space correlation coefficients for residues *A*316–*A*323 of the final refined model are listed in Table 7. The $R_{\rm free}$ value for the final model with both conformations present is 0.183, compared with 0.189 when conformation 1 is omitted and 0.187 when 2 is omitted.

5. Conclusions

Existing model-building software sometimes fails to resolve parts of a protein, resulting in an initial structure with gaps. In this study, we present a two-stage approach to modeling missing main-chain fragments, given the anchor points and an electron-density map. IK techniques allowed us to enforce a closure constraint to guide the loop to its final positioning in space, thus augmenting reduced information available in areas of poor electron density. Experimental results demonstrate that our approach yields fragments in good agreement with the final refined structure, even at medium to low resolution, for lengths up to 12–15 residues. Thus, our algorithm extends automation of model building to areas of weak or ambiguous electron density at resolution levels beyond 2.5 Å.

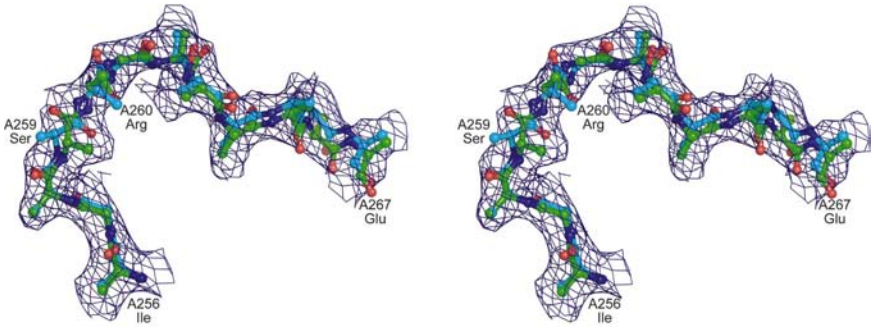

Figure 8
Residues *A*256–*A*267 of TM0542. The top-scoring fragment is shown in cyan and the corresponding manually completed and refined fragment in green. The aaRMSD between the two fragments is 0.87 Å. The fragment is largely correct, apart from residues *A*259 (serine) and *A*260 (arginine) being flipped. The $2mF_o - DF_c$ electron-density map is shown contoured at 1.0σ.

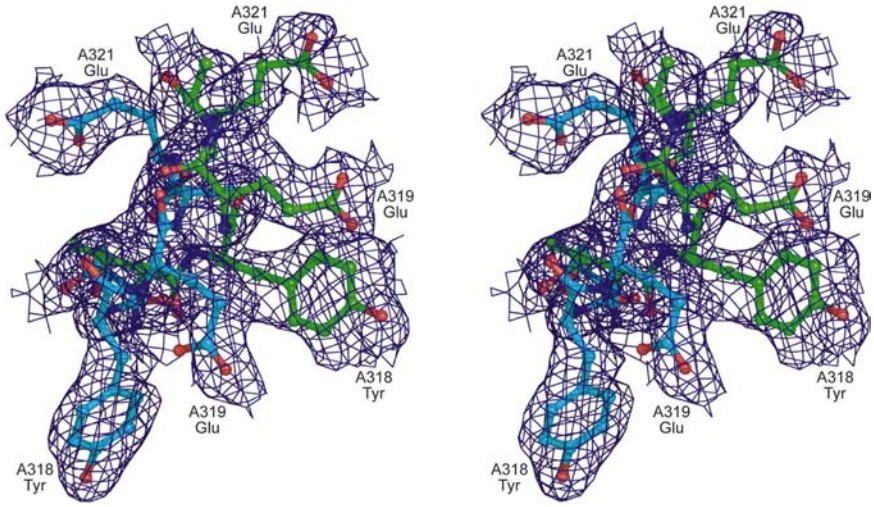

Figure 9
The final refined coordinates of residues *A*317–*A*321 of TM0755 are shown in an omit electron-density map. A total of 33% of the final fragments output by the algorithm converged to conformation 1 (green), while another 25% adopted conformation 2 (cyan). The density is contoured at 0.35σ. The remaining residues of both fragments are omitted for clarity.

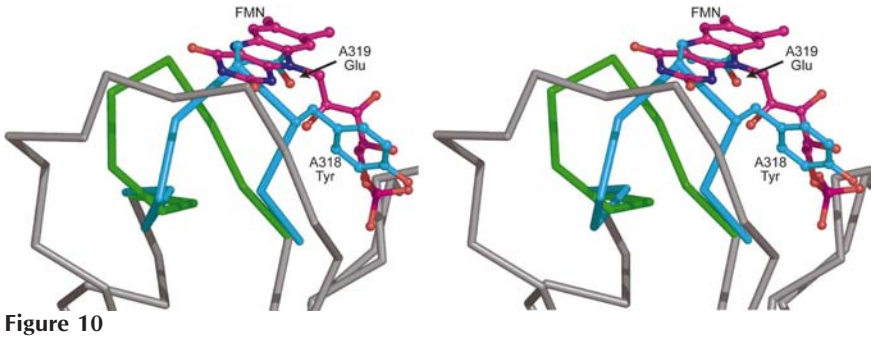

Figure 10
The flavin mononucleotide from rubredoxin oxygen:oxidoreductase superimposed on the corresponding residues of its binding site in TM0755. The C^{α} trace of the final refined coordinates of TM0755 is shown in grey, conformation 1 is shown in green and conformation 2 in cyan.

Fitting a main-chain fragment into areas of poor density is sensitive to residues being flipped along the chain. An important extension to the current algorithm is therefore the ability to identify flipped residues. Although easy to detect and correct manually once the fragment is built, it requires an additional step of human intervention before the model can be submitted to refinement. It is anticipated that elementary heuristic techniques will greatly reduce the occurrence of flipped residues. Similarly, incorporation of specialized algorithms to identify and model secondary-structure elements will enhance the performance in building long α-helices. In cases where the sequence has not been assigned, fragments of various length could be fitted. Using an appropriate score, fragments of correct length and conformation could then be identified.

Advances in all aspects of X-ray crystallography, from protein expression to data processing and instrumentation, are leading to data sets of sufficiently high quality to distinguish alternative main-chain conformations in mobile regions. In §4.3 we have demonstrated that our method can be extended to model alternative conformations, even at a resolution of 1.8 Å.

Inducing a probability measure on conformation space from targeted sampling of self-motion manifolds is another interesting and exciting direction for future research.

6. Software

This algorithm is actively being used in structure determination at the JCSG and work is under way to fully integrate it into *Xsolve*, JCSG's automated data-processing and structure-solution software suite. A software package based on the algorithm

Xpleo is currently under development. It will be available for download at http://smb.slac.stanford.edu/~vdbedem.

Test structures used in this work were solved and deposited as part of the JCSG pipeline (http://www.jcsg.org). The authors would like to thank all members of the JCSG Structure Determination Core at SSRL for their assistance in providing data. In particular, we gratefully acknowledge H. Axelrod, C. L. Rife and M. D. Miller for their help with proteins TM1586 and TM0755. The JCSG is funded by the Protein Structure Initiative of the National Institutes of Health, National Institute of General Medical Sciences (grant P50 GM62411). SSRL operations are funded by DOE BES and the SSRL Structural Molecular Biology program by DOE BER, NIH NCRR BTP and NIH NIGMS. IL was supported in part by a Siebel Fellowship. IL and JCL were also funded by NSF ITR grant CCR-0086013 and a Stanford BioX Research Initiative grant.

References

Abagyan, R. & Totrov, M. (1994). *J. Mol. Biol.* **235**, 983–1002.
Abe, A., Braun, W., Noguti, T. & Gō, N. (1984). *Comput. Chem.* **8**, 239–247.
Adams, P., Grosse-Kunstleve, R., Hung, L.-W., Ioerger, T., McCoy, A., Moriarty, N., Read, R., Sacchettini, J., Sauter, N. & Terwilliger, T. (2002). *Acta Cryst.* D**58**, 1948–1956.
Adams, P. D., Grosse-Kunstleve, R. W. & Brunger, A. T. (2003). *Structural Bioinformatics*, edited by P. E. Bourne & H. Weissig, pp. 75–87. Hoboken, NJ, USA: Wiley–Liss.
Badger, J. (2003). *Acta Cryst.* D**59**, 823–827.
Bedem, H. van den, Miller, M. & Wolf, G. (2003). *Synchrotron Radiat. News*, **16**, 15–19.
Bruccoleri, R. & Karplus, M. (1987). *Biopolymers*, **26**, 137–168.
Brünger, 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.
Burdick, J. (1989). *IEEE Int. Conf. Robot. Autom. (ICRA)*, **1**, 264–270.
Burley, S. K., Almo, S. C., Bonanno, J. B., Capel, M., Chance, M. R., Gaasterland, T., Lin, D., Sali, A., Studier, F. W. & Swaminathan, S. (1999). *Nature Genet.* **23**, 151–157.
Canutescu, A. & Dunbrack, R. Jr (2003). *Protein Sci.* **12**, 963–972.
Chapman, M. (1995). *Acta Cryst.* A**51**, 69–80.
Cohen, A. E., Ellis, P. J., Miller, M. D., Deacon, A. M. & Phizackerley, R. P. (2002). *J. Appl. Cryst.* **35**, 720–726.
Collura, V., Higo, J. & Garnier, J. (1993). *Protein Sci.* **2**, 1502–1510.
Coutsias, E., Seok, C., Jacobson, M. & Dill, K. (2004). *J. Comput. Chem.* **25**, 510–528.
Cowtan, K. D., (2004). *Clipper Libraries.* http://www.ysbl.york.ac.uk/cowtan/clipper/clipper.html.
Craig, J. (1989). *Introduction to Robotics: Manipulation and Control*, 2nd ed. Reading, MA, USA: Addison–Wesley.
Deane, C. & Blundell, T. (2000). *Proteins*, **40**, 135–144.
DePristo, M., de Bakker, P., Lovell, S. & Blundell, T. (2003). *Proteins*, **51**, 41–55.
Diamond, R. (1971). *Acta Cryst.* A**27**, 436–452.
Du, P., Andrec, M. & Levy, R. (2003). *Protein Eng.* **16**, 407–414.
Engh, R. & Huber, R. (1991). *Acta Cryst.* A**47**, 392–400.
Fidelis, K., Stern, P., Bacon, D. & Moult, J. (1994). *Protein Eng.* **7**, 953–960.
Fine, R., Wang, H., Shenkin, P., Yarmush, D. & Levinthal, C. (1986). *Proteins*, **1**, 342–362.
Fiser, A., Do, R. & Sali, A. (2000). *Protein Sci.* **9**, 1753–1773.
Gō, N. & Scheraga, H. (1970). *Macromolecules*, **3**, 178–186.
Ioerger, T. & Sacchettini, J. (2003). *Methods Enzymol.* **374**, 244–270.
Jacobson, M., Pincus, D., Rapp, C., Day, T., Honig, B., Shaw, D. & Friesner, R. (2004). *Proteins*, **55**, 351–367.
Jones, T. & Kjeldgaard, M. (1997). *Methods Enzymol.* **277**, 173–230.
Jones, T. & Thirup, S. (1986). *EMBO J.* **5**, 819–822.
Jones, T., Zou, J.-Y. & Cowtan, S. (1991). *Acta Cryst.* A**47**, 110–119.
Khatib, O. (1987). *Int. J. Robot. Autom.* **RA-3**, 43–53.
Kirkpatrick, S., Gelatt, C. & Vecchi, M. (1983). *Science*, **220**, 671–680.
Kolodny, R., Guibas, L., Levitt, M. & Koehl, P. (2005). Submitted.
Korostelev, A., Bertram, R. & Chapman, M. S. (2002). *Acta Cryst.* D**58**, 761–767.
Krissinel, E. (2004). *CCP4 Coordinate Library Project.* http://www.ebi.ac.uk/keb/cldoc/.
La Fortelle, E. de & Bricogne, G. (1997). *Methods Enzymol.* **276**, 472–494.
Lesley, S. A. *et al.* (2002). *Proc. Natl Acad. Sci. USA*, **99**, 11664–11669.
Levitt, D. (2001). *Acta Cryst.* D**57**, 1013–1019.
Li, Z. & Scheraga, H. (1987). *Proc. Natl Acad. Sci. USA*, **84**, 6611–6615.
Lovell, S., Davis, I., Arendall, W. III, de Bakker, P., Word, J., Prisant, M., Richardson, J. & Richardson, D. (2003). *Proteins*, **50**, 437–450.
McLachlan, G., Peel, D., Basford, K. & Adams, P. (1999). *J. Stat. Software*, **4**(2).
Manocha, D. & Zhu, Y. (1994). *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **2**, 285–293.
Manocha, D., Zhu, Y. & Wright, W. (1995). *Comput. Appl. Biosci.* **11**, 71–86.
Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. & Teller, E. (1953). *J. Chem. Phys.* **21**, 1087–1092.
Morris, R., Perrakis, A. & Lamzin, V. (2002). *Acta Cryst.* D**58**, 968–975.
Moult, J. & James, M. (1986). *Proteins*, **1**, 146–163.
Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.
Oldfield, T. (2001). *Acta Cryst.* D**57**, 82–94.
Perrakis, A., Morris, R. & Lamzin, V. (1999). *Nature Struct. Biol.* **6**, 458–463.
Raghavan, M. & Roth, B. (1989). *International Symposium on Robotics Research*, pp. 314–320. Tokyo, Japan.
Read, R. (1986). *Acta Cryst.* A**42**, 140–149.
Schneider, T. & Sheldrick, G. (2002). *Acta Cryst.* D**58**, 1772–1779.
Shenkin, P., Yarmush, D., Fine, R., Wang, H. & Levinthal, C. (1987). *Biopolymers*, **26**, 2053–2085.
Terwilliger, T. (2003). *Acta Cryst.* D**59**, 38–44.
Terwilliger, T. & Berendzen, J. (1999). *Acta Cryst.* D**56**, 849–861.
Vlijmen, H. van & Karplus, M. (1997). *J. Mol. Biol.* **267**, 975–1001.
Vonrhein, C., Blanc, E., Roversi, P. & Bricogne, G. (2005). In preparation.
Waasmaier, D. & Kirfel, A. (1995). *Acta Cryst.* A**51**, 416–431.
Walsh, M., Dementieva, I., Evans, G., Sanishvili, R. & Joachimiak, A. (1999). *Acta Cryst.* D**55**, 1168–1173.
Wang, L. & Chen, C. (1991). *IEEE Trans. Robot. Autom.* **7**, 489–499.
Wedemeyer, W. & Scheraga, H. (1999). *J. Comput. Chem.* **20**, 819–844.
Wilson, M. & Brunger, A. (2000). *J. Mol. Biol.* **301**, 1237–1256.
Yakey, J., LaValle, S. M. & Kavraki, L. (2001). *IEEE Trans. Robot. Autom.* **17**, 951–959.
Zheng, Q., Rosenfeld, R., Vajda, S. & DeLisi, C. (1992). *J. Comput. Chem.* **14**, 556–565.