Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Received 0 XXXXXXX 0000
Accepted 0 XXXXXXX 0000
Online 0 XXXXXXX 0000

Automated Protein Model Completion: an Inverse Kinematics Approach

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

[a]Joint Center for Structural Genomics, Stanford Synchrotron Radiation Laboratory, SLAC, 2575 Sand Hill Road, Menlo Park, CA 94025., and [b]Department of Computer Science, Stanford University, Stanford CA 94305.. Correspondence e-mail: adeacon@slac.stanford.edu

Rapid protein structure determination relies greatly on the availability of software that can automatically build a protein model into an experimental electron density map. In favorable cases, the programs *ARP/wARP*, *RESOLVE*, *MAID*, and *TEXTAL* are all capable of building over 90% of the final model. At medium-low resolution (2.3Å $\leq d <$ 2.9Å), only about 2/3 completeness is typically attained. Manual completion of these partial models is usually feasible, but is time-consuming, and easily leads to inaccuracies. Except for the N- and C-termini of the chain, the end points of each missing fragment are known from the initial model. Hence, loop fitting reduces to an *inverse kinematics* problem.

We have combined a fast inverse kinematics algorithm with real space, torsion angle refinement in a two stage approach to fit a poly-alanine chain to the electron density between two anchor points. The first stage aims to sample a large number of closing conformations, guided by the electron density. These candidate conformations are ranked according to density fit. Top-ranking conformations are then subjected to torsion angle, subchain refinement in the second stage. Optimization steps are projected onto the null space of the subchain, thus preserving rigid geometry and closure.

In a test set of 103 structurally diverse fragments, the algorithm closed gaps of 12 residues in length to within, on average, 0.52Å aaRMSD of the final, refined structure at a resolution of 2.8Å. In an initial, 51%-complete model built at 2.6Å, it closed a 14-residue gap to within 0.9Å aaRMSD, thus extending automation of model building towards lower resolution levels.

© 2004 International Union of Crystallography
Printed in Denmark – all rights reserved

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 fully 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 in the process of structure solution (Burley *et al.*, 1999; Adams *et al.*, 2003).

There have 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.*, 1997; Terwilliger, 2002). 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 to an initial model of a new protein structure in a few weeks.

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%. Accurately determining the atomic coordinates of mobile fragments in the molecule, for instance, remains a challenge. Such fragments often 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 easily leads to inaccuracies. This step alone 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, often a large portion of the molecule has been resolved, and *N*- and *C*-termini of a missing fragment in the

IUCr macros version 2.0β15: 2004/05/19

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 can thus be interpreted as an *inverse kinematics* (IK) problem (Manocha & Zhu, 1994; Manocha *et al.*, 1995): Given the position and orientation of the end point of a kinematic chain, determine the corresponding values of the joint angles.

Exploiting this observation, we have combined an inverse kinematics algorithm with real space, torsion angle refinement in a two stage approach to fit a poly-alanine chain to the electron density between two anchor points. The first stage aims to sample a large number of closing conformations, guided by the electron density. These candidate conformations are ranked according to density fit. Top-ranking conformations are then subjected to torsion angle, real space subchain refinement in the second stage. Optimization steps are projected onto the null space of the subchain, thus preserving rigid geometry and closure.

In a test set of 103 structurally diverse fragments, the algorithm closed gaps of 12 residues in length to within, on average, 0.52Å all-atom Root Mean Square Deviation (aaRMSD[1]) of the final, refined structure at a resolution of 2.8Å. The algorithm has also been tested and used to aid protein model completion in areas of poor experimental electron density, where the initial model was built using ARP/wARP or RESOLVE. At a resolution of 2.4Å, it closed a 10-residue gap to within 0.43Å aaRMSD of the final, refined structure. In an initial, 51%-complete model built at 2.6Å, it closed a 14-residue gap to within 0.9Å aaRMSD, thus extending automation of model building towards lower resolution levels.

2. Background

A variety of techniques have found successful application and widespread use in automated interpretation of electron density maps. The program *ARP/wARP* (Perrakis *et al.*, 1997), for instance, iteratively adds pseudo-atoms to a partial model that it subsequently refines in reciprocal space. The program *TEXTAL* (Ioerger & Sacchettini, 2003) employs local pattern recognition techniques to select regions in a database of previously determined structures similar to those in the unknown structure. Some automated systems targeting lower resolution levels, notably *RESOLVE* (Terwilliger, 2002) 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 to template matching techniques for a variety of reasons. The electron density may exhibit *multimodal disorder*, where 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 medium to low resolution levels ($2.3\text{Å} \leq d < 2.9\text{Å}$), the initial model resulting from these programs is typically a gapped polypeptide chain, and only about 2/3 completeness is attained. 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 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 (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.

The problem of fitting a protein backbone fragment between two anchor points is closely related to the inverse kinematics problem in robotics (Craig, 1989). It is known that for manipulators in a 3-D workspace there are a finite number of solutions to the IK problem when the number of DoFs does not exceed six. In the case of a 6R manipulator, which is the most relevant to protein fragments, an analytic solution exists and the number of unique solutions is at most sixteen (Raghavan & Roth, 1989).

Gō and Scheraga were the first to study analytical loop closure, limited to 6 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 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 discrete 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.*, 2004). Extracting candidate fragments from the PDB satisfying conditions on length and geometry started with (Jones & Thirup, 1986), and was further developed in (Fidelis *et al.*, 1994; van Vlijmen & Karplus, 1997; Du *et al.*, 2003). Various methods exist for optimization of candidate loops, such as molecular dynamics (Bruccoleri & Karplus, 1987; Fiser *et al.*, 2000; Zheng

[1] Square root of the averaged squared distances between the corresponding atoms $\{N_i, C\alpha_i, C\beta_i, C_i, O_i\}$. It is calculated after the loops are optimally aligned in 3-D.

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 terminals 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 Jr, 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.

This study combines the CCD loop closure algorithm with real space, torsion angle subchain refinement to aid model completion. The objective is to automatically fit a poly-alanine chain between two anchor residues, satisfying electron density constraints. Real space, least squares refinement offers the advantage of speed over otherwise superior reciprocal space, maximum likelihood refinement techniques. The algorithm assumes rigid peptide geometry with residue-dependent values for bond lenghts and bond angles taken from (Engh & Huber, 1991). Final chains will need to be refined using standard refinement programs such as CNS (Brunger *et al.*, 1998) or REFMAC (Murshudov *et al.*, 1997).

3. Methods

The algorithm proceeds in two stages: candidate generation and refinement. In the first stage, candidate loops are built using the CCD algorithm, while putting additional constraints on the DoFs to take the electron density and collision avoidance into account. Next, initial conformations are ranked according to density fit. Top-ranking initial conformations are refined 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) searches for the global minimum of the target function while maintaining loop closure. Each candidate is optimized 6 times and the best scoring loops are returned.

Deficient density information is compensated for by taking advantage of the loop closure constraint to guide the loop to its correct positioning in space. In the first stage, the closure constraint enables the generation of loops that lie within 2Å RMSD of the true solution. The approximate enforcement of the closure constraint during loop refinement prevents the search from diverging and limits the searched space to motions that preserve loop closure.

The input to the algorithm is given by the electron density, in most cases a $2mF_o - DF_c$ map, the partial model, and the amino acid sequence. The latter is needed to identify the missing residues.

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 will be denoted *stationary* anchors. The algorithm starts by constructing a protein chain $\mathcal{C}$ of length L in a random conformation, 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 adjusts each backbone dihedral angle in turn such that the distance between the three backbone atoms of the mobile anchor and the corresponding atoms of the stationary anchor are minimized.

For longer loops (9 or more residues), $\mathcal{C}$ is 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.

A total of 1000 starting conformations are calculated to start the procedure. Each is allowed 2000 iterations for closure up to a preset tolerance distance d_{closed}. Chains that did not close are discarded. A cross-correlation density score is calculated for all conformations, and the 99-th percentile (with a maximum of 6 chains) is passed on to stage two. Each of these is then subjected to 6 SA refinement cycles, the 2 top-scoring fragments of which are written to disk. As most chains close within 2000 iterations, this gives a total of 12 fragments. The program also writes a log file containing the full cross-correlation electron density score for each fragment.

3.1.1. Random Initial Conformations For each starting configuration, ω_i is considered to be a fixed, $N(180, 5.8)$ random variable for all i. Half of the starting configurations are obtained by adjusting each $(\phi, \psi)_i$ in turn to optimize agreement with the electon density while stereochemical constraints are observed. The remaining five hundred starting conformations are purely random, and obtained from sampling $(\phi, \psi)_i, i = 0 \ldots L-1$ angle pairs from PDB-derived distributions. A finite mixture of bivariate normal distributions was thereto 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.

3.1.2. Electron Density Constraints A change to the DoFs of a residue is calculated as follows: The CCD step proposes a distance minimizing dihedral angle ϕ_i for residue i, and based on ϕ_i, it proposes a minimizing ψ_i. (In our implementation, we change each DoF in turn, although this is not strictly necessary.) Thus, a proposed angle pair $(\phi, \psi)_i^p$ is obtained. To guide the loop, a heuristic electron density constraint has been added to the CCD algorithm. For each pair i, consider the set of atoms $\mathcal{A}_i$ that is subject to change by angle pair i, and not affected by changes in angle pair $i + 1$. Hence, $\mathcal{A}_i = \{C\beta_i, C_i, O_i, N_{i+1}, C\alpha_{i+1}\}$, where $C\beta_i$ is excluded whenever residue i is a Glycine. Electron density scores corresponding to trial positions in a square neighborhood $U_{(\phi,\psi)^p}$ about $(\phi,\psi)^p$ in conformation space are calculated. A simple local scoring function is adopted; the sum of the electron density values at atom center positions of $\mathcal{A}_i$. The angle pair $(\phi,\psi)_i$ is set to the trial position with maximum density score, i.e $(\phi,\psi)_i = \arg\max_{(\phi,\psi)\in U_{(\phi,\psi)_i^p}} S(\phi,\psi)$, where $S(\phi,\psi) = \sum_{A_j\in\mathcal{A}_i} \rho(A_j(c))$, and $\rho(A_j(c))$ denotes the value of the electron density at the center of atom A_j. At this point, overlaps of van der Waals surfaces of atoms in $\mathcal{A}_i$ and the rest of the protein structure are determined. If no overlaps occur, the new $(\phi,\psi)_i$ 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.

3.2. Stage 2: Refinement

A candidate fragment is refined by minimizing the least squares residuals between the observed density ρ^o and the density calculated from the model ρ^c. The target function sums the squared differences between the observed density and the calculated density at each grid point in some volume V around the fragment:

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

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 3-D Gaussians (Waasmaier & Kirfel, 1995). The factors S and k scale ρ^o to ρ^c and are computed once at initialization using the partial model.

3.2.1. Optimization with closure constraints Our method uses the redundant DoFs of the fragment to minimize the target function without breaking closure. 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 (Craig, 1989) of the fragment. 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:

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

Thus, $J(q) = df(q)/d(q)$ 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 Equation 2. We get:

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

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 get 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. 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 0.5Å 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 Figure 1. The inner loop performs 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 3-D Gaussian kernel. Since the convolution of a Gaussian with a

Gaussian can be computed analytically by summing the means and variances of the two functions, the computation of ρ^c does not require any convolution. The variance of the Gaussian function used to represent the density contribution of atoms is simply augmented by the variance of the desired kernel.

```
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

4. Results and Discussion

The performance of the algorithm was evaluated on a test set of 103 structurally diverse fragments at various resolution levels. Additionally, we tested its ability to close gaps at various lengths using experimental data and initial models provided by the JCSG. We furtermore 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 4, 8, 12, and 15 at each even numbered residue of a test structure, the protein TM1621 (PDB code 1O1Z, SCOP classification a/b). TM1621 consists of one chain, with 34% of the residues in 10 alpha helices, and 19% in 9 beta sheets. Diffraction data for this 234-residue protein structure had been collected at a resolution of 1.6Å. To evaluate the performance at various resolution levels, three $2mF_o - DF_c$ electron density maps were calculated 2.0, 2.5, and 2.8Å, using structure factors obtained from the PDB. Since the low resolution electron density was obtained by truncating a high resolution data set, the RMSDs in this section are not typical for their resolution levels.

At a resolution of 2.0Å, the algorithm successfully closed all 103 gaps of length 4 to within 1.0Å, and all length 8 gaps to within 0.85Å, as shown in Figure 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 1.0Å.

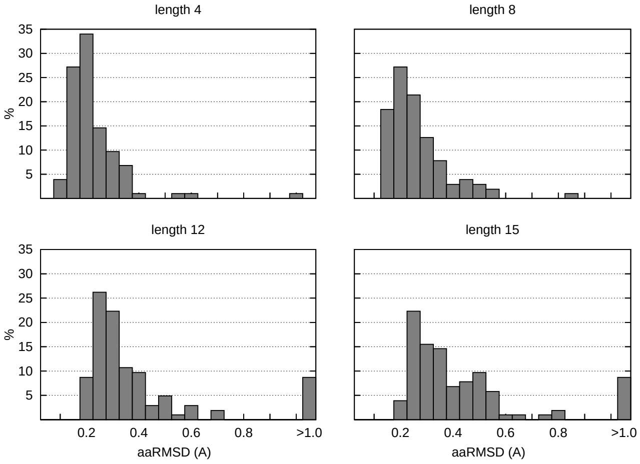

Figure 2
The aaRMSD-distribution of 103 fragments with lengths 4, 8, 12, and 15 residues of TM1621 at a resolution of 2.0Å. A total of 9% of 12-residue and 9% of 15-residue fragments have an aaRMSD $>$ 1.0Å.

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 2/3 of its residues are classified as such. A total of fourteen 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 0.3Å. Four percent of non-helical, 12-residue fragments were found to have an aaRMSD $>$ 1.0Å, compared to 36% of helical fragments. For 15-residue fragments, these numbers are 4% and 63% respectively.

At a resolution of 2.5Å, all gaps of length 4 and 8 were closed to within 1.0Å aaRMSD and 0.85Å aaRMSD resp., whereas four 12-residue fragments and twelf 15-residue fragments deviated by more than 1.0Å aaRMSD. The results are depicted in Figure 3. One percent of non-helical, 12-residue fragments were found to have an aaRMSD $>$ 1.0Å, compared to 21% of helical fragments. For 15-residue fragments, these numbers are 7% and 63% respectively.

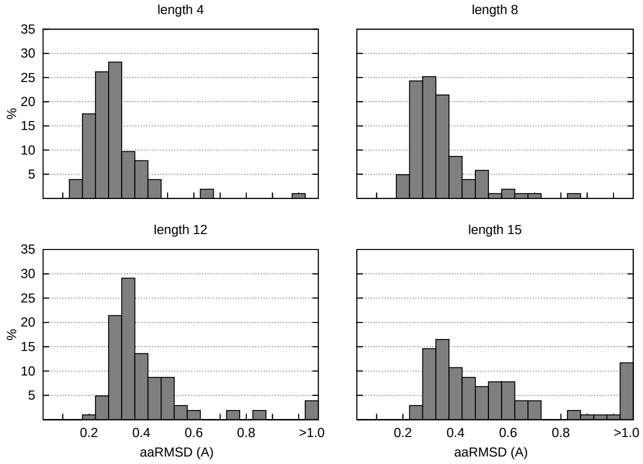

Figure 3
The aaRMSD-distribution of 103 fragments with lengths 4, 8, 12, and 15 residues of TM1621 at a resolution of 2.5Å. A total of 4% of fragments of length 12, and 12% of fragments of length 15 have an aaRMSD $>$ 1.0Å.

At a resolution of 2.8Å, all gaps of length 4 and 8 closed to within 1.05Å aaRMSD and 0.75Å aaRMSD resp.. Four 12-residue fragments and eighteen 15-residue fragments deviated by more than 1.0Å aaRMSD. The results are depicted in Figure 4. Two percent of non-helical, 12-residue fragments were found to have an aaRMSD $>$ 1.0Å, compared to 14% of helical fragments. For 15-residue fragments, these numbers are 12% and 88% respectively.

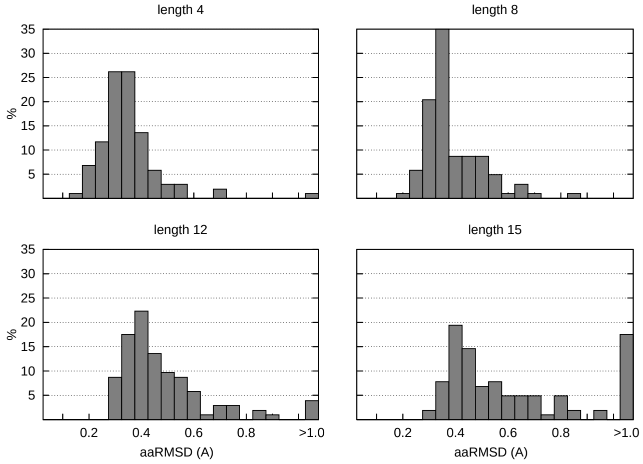

Figure 4
The aaRMSD-distribution of 103 fragments at lengths 4, 8, 12, and 15 residues of TM1621 at a resolution of 2.8Å. A total of 4% of fragments of length 12, and 17% of fragments of length 15 have an aaRMSD $>$ 1.0Å.

Table 1 summarizes the performance at three resolution levels.

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

	2.0Å			2.5Å			2.8Å		
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

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 minutes for short fragments to just under 3 hours 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 2.66GHz Intel P4 Xeon running RedHat 9. The source code was compiled using gcc 3.2.

Table 2
Average run times (in minutes) on a 2.66GHz Intel P4 Xeon at various fragment lengths and resolution levels. Average is calculated over 103 fragments.

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

An equivalent analysis on TM0423 (376 residues, PDB code 1KQ3, SCOP classification multi-domain a/b, 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 8 beta-sheets. The longest helix has length 17, and if a single Glycine classified as a hydrogen bonded turn is included, its length is 26.

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

	2.0Å			2.5Å			2.8Å		
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

Clearly, the algorithm performs more modestly when fitting longer fragments. In addition to an increasing median aaRMSD, a larger proportion of fragments deviates by more than 1.0Å as fragment length increases, particularly when a large number of residues are in alpha-helical 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.

Interestingly, lowering the resolution of the data only mildly affects performance. We believe that this is the true strength of the algorithm; lack of structured, well-defined electron density information is compensated by maintaining a closed conformation.

4.2. Missing Fragments

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

In one instance, the 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 anchoring 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 < 4 are extended to length 4 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 can not 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 beta-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 2.0Å. An initial model for the 206-residue hypothetical protein TM1586 was obtained from Xsolve, a fully automated crystallographic data processing and structure solution software suite under development at the JCSG (Wolf, 2004). At the time of processing this data, Xsolve only supported RESOLVE v2.06 for model building.

The model was obtained from MAD data collected at 2.0Å, and showed gaps in 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 gap 142-152 by one residue at each end, this gap was easily closed to within 0.5Å aaRMSD using an experimental map obtained with SOLVE v2.03 (Terwilliger & Berendzen, 1999). The gaps in between residues 86-98 and 107-117 proved to be more difficult. The extended RESOLVE model was combined with an ARP/wARP model, and a more complete model was obtained after various rounds of phase improvements. The *N*-terminus was now largely complete, with gaps remaining in between residues 13-23, 49-52, 89-99, and 105-113. 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 also trimmed back one residue at the *C*-terminus. The missing fragments were all located on one face of the molecule, and the density remained weak in this area. The map improved slightly after phases obtained from SHELXD (Schneider & Sheldrick, 2002) and autoSHARP were used. At this point, the remaining missing fragments were generated, which served as a starting point for subsequent manual refinement. The resulting structure was subsequently refined with REFMAC5. Table 4 shows the aaRMSD of fragments to this final, refined model.

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

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. Figure 5 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 0.8 σ, and that side-chain density is poorly defined.

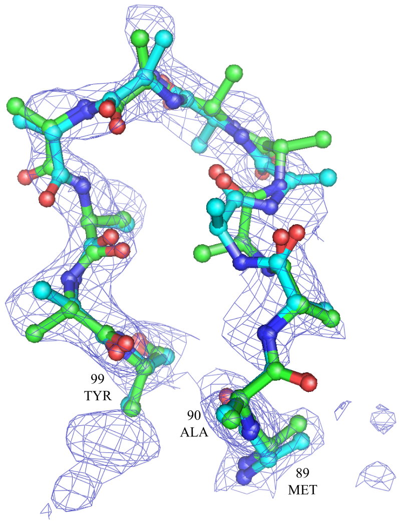

Figure 5
Residues 89-99 of TM1586. The fragment inserted into the model is shown in cyan, and the corresponding final, refined fragment in green. The aaRMSD between the two fragments is 1.01Å. The electron density map is shown contoured at 0.8 σ, and is discontinuous around the Alanine 90.

4.2.2. TM1742 at 2.4Å. MAD data for the 271-residue, putative Nagd protein TM1742 (PDB code 1VJR) was collected at a resolution of 2.4Å. An initial electron density map of good quality was obtained using the program SOLVE v2.03 (Terwilliger & Berendzen, 1999), at a resolution of 2.5Å. Iterative model building using Terwilliger's `resolve_build` script resulted in an 88% complete model, with gaps in between residues 17-25, 56-62, 129-132, 146-148, 229-231. Furthermore, the region in between residues 191-202 had been built incorrectly. The

IUCr macros version 2.0β15: 2004/05/19

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.

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

4.2.3. TM0542 at 2.6Å. MAD data for the 376-residue protein TM0542 (Malate Oxidoreductase) was collected at a resolution of 3.0Å, and a native data set was obtained at 2.6Å. An electron density map was calculated with phase extension using the program SOLVE. Iterative model building using RESOLVE 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 in 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, but for the first 76-residue one. Table 6 summarizes the results.

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

Fitting a poly-alanine 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. Figure 6 shows an example of a fragment where two consecutive residues are flipped. While the aaRMSD is relatively high at 0.9Å for this fragment, the $C\alpha$-trace is in excellent agreement with the manually built fragment. The flipped residues are easy to identify and correct for a trained crystallographer.

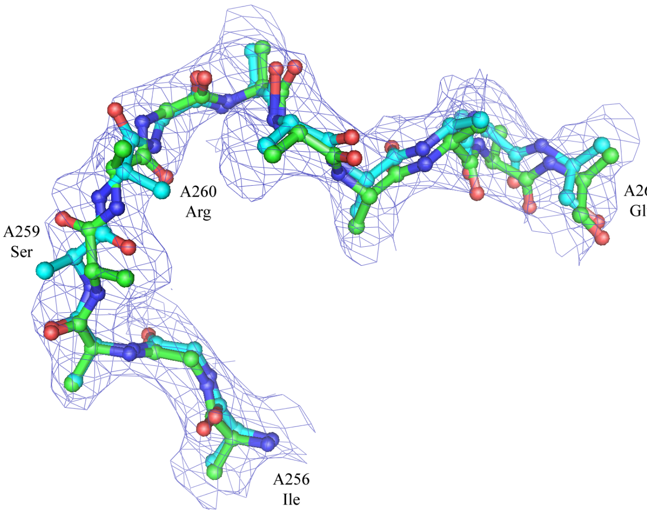

Figure 6
Residues A256-A267 of TM0542. The top-scoring fragment is shown in cyan, and the corresponding manually completed and refined fragment in green. The aaRMSD between the two fragments is 0.87Å. The fragment is largely correct, apart from residues A259 (Serine) and A260 (Arginine) being flipped. The electron density map is shown contoured at 1.0 σ.

4.3. Identifying alternative main-chain conformations

Binding of ligands to a protein or protein-protein interactions are typically facilitated by mobile regions in the macromolecule. Such flexible fragments sometimes crystallize in multimodal disordered substates, where 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 our method can be extended to support identification and refinement of multiple, distinct conformations at sub-atomic resolution.

A model for the 398-residue hypothetical protein TM0755 was determined from a 1.8Å MAD data set using ARP/wARP. The structure was completed manually, apart form a short fragment around residue A320. The electron density from residue A317 to A323 indicated that this fragment had crystallized in two distinct conformations. Furthermore, a structurally similar dioxygen reduction enzyme, Rubredoxin Oxygen: Oxidoreductase (pdb code 1e5d), binds a Flavin Mononucleotide at the corresponding residues, providing additional evidence for the presence of multiple conformations at this site.

While one conformation was clearly visible in the electron density, the main-chain trace of the alternative conformation was much less obvious. From residue A320 to A323 the density was particularly ambiguous; the alternative conformation was difficult to identify and initially not modelled. To model the fragment from residue A317 to A323 with our algorithm, it was decided to build two conformations at half occupancy each. The algorithm was slightly modified; 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 Serine A316, and the C-anchor ranged from Alanine A320 to

Histidine A323. In the final run, four out of the final twelf fragments adopted configuration 'A', another three adopted conformation 'B', and the remaining five fragments did not fit the density meaningfully. Figure 7(a) shows the two alternative conformations for residues A316-A323. Side chains were added manually to the poly-alanine chains. Figure 7(b) and (c) show residues A316 to A320 of both conformations in the electron density. The fit of Tyrosine A318 is particularly telling in each case.

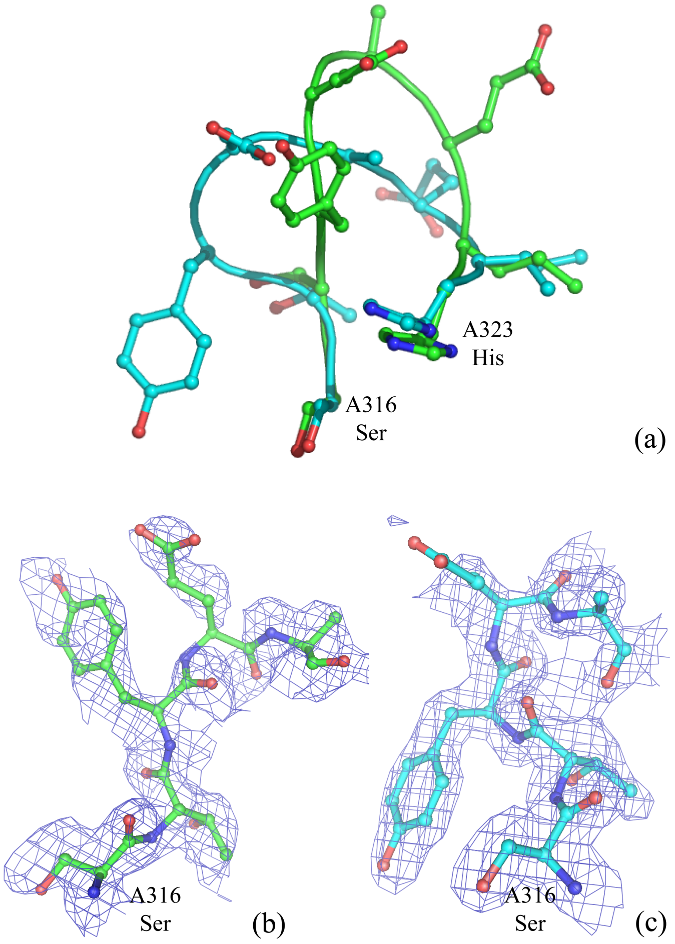

Figure 7
(a) The two alternative conformations for residues A316-A323 in the hypothetical protein TM0755. A total of 33% the final fragments output by the algorithm converged to conformation 'A', while another 25% assumed conformation 'B'. Side chains were added manually. (b) Residues A316-A320 of conformation 'A'. (c) Residues A316-A320 of conformation 'B'. For clarity, residues A321-A323 are omitted in figures (b) and (c).

5. Conclusion

Existing model building software sometimes fails to resolve parts of a protein, resulting in an initial structure with gaps. In this study we presented a two stage approach to model missing main-chain fragments, given the anchor points and an electron density map. IK techniques allowed us to enforce a closure constraint, 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, at lengths up to 12-15 residues.

Fitting a poly-alanine 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 alpha-helices.

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. Our methods can easily be extended to model alternative conformations, even at subatomic resolution, as was shown in Section 4.3. 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

The algorithm is actively being used in the 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.

Acknowledgements

Test structures used in this work were solved and deposited as part of the JCSG pipeline (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. SSRL operations is funded by DOE BES, and the SSRL Structural Molecular Biology program by DOE BER, NIH NCRR BTP and NIH NIGMS. IL is supported in part by a Siebel Fellowship. IL and J-CL are 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**(4), 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). In *Structural Bioinformatics*, pp. 75–87. Hoboken, N.J.: Wiley-Liss.

Badger, J. (2003). *Acta Cryst.* D**59**, 823–827.

van den Bedem, H., Miller, M. & Wolf, G. (2003). *Synchrotron Radiation News*, **16**, 15–19.

Bruccoleri, R. & Karplus, M. (1987). *Biopolymers*, **26**, 137–168.

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

Burdick, J. (1989). In *IEEE Int. Conf. Robot. Autom. (ICRA)*, vol. 1, pp. 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). *Nat. Genet.* **23**, 151–157.

Canutescu, A. & Dunbrack Jr, R. (2003). *Prot. Sci.* **12**, 963–972.

Chapman, M. (1995). *Acta Cryst.* A**51**(1), 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). *Prot. 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 nad control*. Addison-Wesley, 2nd ed.

Deane, C. & Blundell, T. (2000). *Proteins*, **40**(1), 135–144.

DePristo, M., de Bakker, P., Lovell, S. & Blundell, T. (2003). *Proteins*, **51(1)**, 41–55.

Diamond, R. (1971). *Acta Cryst.* A**27**(5), 436–452.

Du, P., Andrec, M. & Levy, R. (2003). *Prot. Engin.* **16**(6), 407–414.

Engh, R. & Huber, R. (1991). *Acta Cryst.* A**47**, 392–400.

Fidelis, K., Stern, P., Bacon, D. & Moult, J. (1994). *Prot. Engin.* **7**(8), 953–960.

Fine, R., Wang, H., Shenkin, P., Yarmush, D. & Levinthal, C. (1986). *Proteins*, **1**, 342–362.

Fiser, A., Do, R. & Sali, A. (2000). *Prot. Sci.* **9(9)**, 1753–73.

Gō, N. & Scheraga, H. (1970). *Macromolecules*, **3**, 178–186.

Ioerger, T. & Sacchettini, J. (2003). In *Methods in Enzymology.*, vol. 374, pp. 244–270. San Diego: Academic Press.

Jacobson, M., Pincus, D., Rapp, C., Day, T., Honig, B., Shaw, D. & Friesner, R. (2004). *Proteins.* **55**(2), 351–67.

Jones, T. & Kjeldgaard, M. (1997). In *Methods in Enzymology*, vol. 277, pp. 173–230. San Diego: Academic Press.

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**(1), 43–53.

Kirkpatrick, S., Gelatt, C. & Vecchi, M. (1983). *Science*, **220**(4598), 671–680.

Kolodny, R., Guibas, L., Levitt, M. & Koehl, P. (2004). Inverse kinematics in biology: the protein loop closure problem. Submitted to IJRR.

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/.

Lesley, S. A., Kuhn, P., Godzik, A., Deacon, A. M., Mathews, I., Kreusch, A., Spraggon, G., Klock, H. E., McMullan, D., Shin, T. & et. al. (2002). *Proc. Nat. Acad. Sci.* **99**(18), 11664–11669.

Levitt, D. (2001). *Acta Cryst.* D**57**, 1013–1019.

Li, Z. & Scheraga, H. (1987). *Proc. Natl. Acad. Sci.* **84**(19), 6611–6615.

Lovell, S., Davis, I., Arendall III, W., de Bakker, P., Word, J., Prisant, M., Richardson, J. & Richardson, D. (2003). *Proteins*, **50**(3), 437–450.

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**(1), 71–86.

McLachlan, G., Peel, D., Basford, K. & Adams, P. (1999). *J. Stat. Software*, **4**(2).

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. & Teller, E. (1953). *J. Chem. Phys.* **21**, 1087–1092.

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., Sixma, T., Wilson, K. & Lamzin, V. (1997). *Acta Cryst.* D**53**, 448–455.

Raghavan, M. & Roth, B. (1989). In *Int. Symp. Robot. Res.*, pp. 314–320. Tokyo.

Schneider, T. & Sheldrick, G. (2002). *Acta Cryst.* D**58**, 1772–1779.

Shenkin, P., Yarmush, D., Fine, R., Wang, H. & Levinthal, C. (1987). *Biopolymers*, **26**, 20532085.

Terwilliger, T. (2002). *Acta Cryst.* D**59**, 34–44.

Terwilliger, T. & Berendzen, J. (1999). *Acta Cryst.* .

van Vlijmen, H. & Karplus, M. (1997). *J. Mol. Biol.* **267**(4), 975–1001.

Waasmaier, D. & Kirfel, A. (1995). *Acta Cryst.* A**51**(3), 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.

Wolf, G., (2004). Personal Communication.

Yakey, J., S.M., L. & Kavraki, L. (2001). *IEEE Trans. Robot. Autom.* **17**(6), 951–959.

Zheng, Q., Rosenfeld, R., Vajda, S. & DeLisi, C. (1992). *J. Comput. Chem*, **14**, 556–565.

IUCr macros version 2.0β15: 2004/05/19