Iterative model-building, structure refinement, and density modification with the PHENIX AutoBuild Wizard

Thomas C. Terwilliger[a*], Ralf W. Grosse-Kunstleve[b], Pavel V. Afonine[b], Nigel W. Moriarty[b], Peter Zwart[b], Li-Wei Hung[a], Randy J. Read[c], Paul D. Adams[b*]

[a]*Los Alamos National Laboratory, Mailstop M888, Los Alamos, NM 87545, USA* [b]*Lawrence Berkeley National Laboratory, One Cyclotron Road, Bldg 64R0121, Berkeley, CA 94720, USA.* [c]*Department of Haematology, University of Cambridge, Cambridge CB2 0XY, UK.*

* Email: terwill@lanl.gov or PDAdams@lbl.gov

Running title: The PHENIX AutoBuild Wizard

Abstract The PHENIX AutoBuild Wizard is a highly automated tool for iterative model-building, structure refinement and density modification using RESOLVE or TEXTAL model-building, RESOLVE statistical density modification, and phenix.refine structure refinement. Recent advances in the AutoBuild Wizard and phenix.refine include automated detection and application of NCS from models as they are built, extensive model completion algorithms, and automated solvent molecule picking. Model completion algorithms in the AutoBuild Wizard include loop-building, crossovers between chains in different models of a structure, and side-chain optimization. The AutoBuild Wizard has been applied to a set of 48 structures at resolutions ranging from 1.1 Å to 3.2 Å, resulting in a mean R-factor of 0.24 and a mean free R factor of 0.29. The R-factor of the final model is dependent on the quality of the starting electron density, and relatively independent of resolution.

Keywords: Model building; model completion; macromolecular models; Protein Data Bank; structure refinement; PHENIX

1. Introduction

Iterative model-building and refinement is a powerful approach to obtaining a complete and accurate macromolecular model. The approach consists of cycles of building an atomic model based on an electron density map for a macromolecular structure, refining the structure, using the refined structure as a basis for improving the map, and building a new model. This type of approach has been carried out in a semi-automated fashion for many years, with manual model-building iterating with automated refinement (Jensen, 1997). More recently, with the development first of ARP/wARP (Perrakis et al., 1999), and later other procedures including RESOLVE iterative model-building and refinement (Terwilliger,

2003b), RAPPER (DePristo et al., 2005) and hip-hop refinement Ondráček, (2005), the entire process has become highly automated.

Despite the high degree of sophistication and automation of these procedures, there remain many improvements to be made, particularly in the automation of the process at low resolutions, in the completion of models, and in model editing and validation. The AutoBuild Wizard has been developed as a part of the PHENIX project (Adams et al., 2002) as a second-generation tool for iterative model-building, density modification and refinement with these needs in mind. Here we describe current features of the AutoBuild Wizard and the application of the Wizard to a set of structures from a library of experimentally phased structures.

2. Methods

2.1. Iterative model-rebuilding, density modification and refinement

The purpose of the AutoBuild Wizard is to provide a highly automated system for model rebuilding and completion. The Wizard design allows the user to specify data files and parameters through an interactive GUI, or alternatively through keyworded scripts. The AutoBuild Wizard begins with datafiles with structure factor amplitudes and uncertainties, along with either experimental phase information or a starting model, typically from molecular replacement, carries out cycles of model-building and structure refinement alternating with model-based density modification, and produces a relatively complete atomic model.

The AutoBuild Wizard has been designed for ease of use combined with maximal user control, with as many parameters set automatically by the Wizard as possible, but maintaining parameters accessible to the user through a GUI and through keyword-based scripts. The Wizard uses the input/output routines of the *cctbx* library (Grosse-Kunstleve, Sauter, & Adams, 2004) allowing data files of many different formats so that user data need not be converted to any particular format before using the Wizard. Use of the phenix.refine refinement package (Afonine, Grosse-Kunstleve & Adams, 2005b) in the AutoBuild Wizard allows a high degree of automation of refinement so that the neither the user nor Wizard is required to specify parameters for refinement. The phenix.refine package automatically includes a robust bulk solvent model and automatically places solvent molecules (Afonine, Grosse-Kunstleve & Adams, 2005a).

The five core modules in the AutoBuild Wizard are (1) building a new model into an electron density map, (2) rebuilding an existing model, (3) refinement, (4) iterative model-building beginning from experimental phase information, and (5) iterative model-building beginning from a model. These five procedures are described in the next sections.

The standard procedures available in the AutoBuild Wizard that are based on these modules include (a) model-building and completion starting from experimental phases, (b) rebuilding a model from scratch, with or without experimental phase information, and (c) rebuilding a model in place, maintaining connectivity and sequence register. In cases where the starting point is a set of experimental phases and structure factor amplitudes, normally procedure (a) is carried out, and then the resulting model is rebuilt with procedure (b). In cases where the starting point is a model (e.g., from molecular replacement) and experimental structure factor amplitudes, procedure (c) is normally carried out if the starting model differs less than about 5% in sequence from the desired model, and otherwise procedure (b) is used.

2.2. Building a model into an electron density map

The AutoBuild Wizard has a multi-step procedure for building an initial model into an electron density map. In this procedure, several models are built, refined, and recombined with each other to create new models. If a model is available from a previous step or is provided by the user, this model can be recombined with the other models as well. After each stage of building there is a single "best" model, and any number of additional models that have been constructed up to that point.

Initial models are scored based on the number of residues built (N_{built}), the number of residues assigned to the sequence (N_{placed}), and the number of chains in the model (N_{chains}). A large number of chains typically indicates that there are many places where chain connectivity is broken. The score (Q) is calculated as $Q = N_{built} + N_{placed} - 2 * N_{chains}$. Once a model is obtained with an R-factor below a pre-set threshold (typically 0.40), then low R-factors are used instead of high Q-score for identifying the best model.

The model-building process begins with (1) building several models with RESOLVE (Terwilliger, 2003a) into the electron density map. Optionally a model can be built with TEXTAL (Ioerger & Sacchettini, 2003), or with both RESOLVE and TEXTAL. The RESOLVE model-building procedure uses a convolution-based search for helices and strand fragments in the map, and this search gives results that depend on the precise orientations of helix and strand templates that are used in the search. Consequently a relatively diverse set of models can be created by simply varying the parameters of this convolution search. Typically three models are built in the first step of the AutoBuild model-building procedure. The best model is refined with phenix.refine as described below, including automatic placement of waters and the use of NCS if present, and all models (refined and unrefined) are used in the next step.

The models created in step (1) above are then combined (2) into a single model using the RESOLVE "extend-only" model-building procedure. In this procedure a model or models are cut into overlapping segments (typically 10 residues long) and are extended by as far as

possible into the electron density by RESOLVE model-building. The resulting set of overlapping segments are then combined into one or more chains by scoring the segments based on length and fit to density, and iteratively extending the highest-scoring segment by joining another segment to it, crossing over in a place where two or more sequential C_α-atoms in the two segments superimpose within a small distance (typically 1 Å).

Once a "best" single model is obtained from step (2) above, attempts are made to improve this model by (a) rebuilding in the region outside the current model, and (b) by using two methods to try to fit loops. The rationale for rebuilding in the region outside the current model is that the thresholds for fit of a segment of a model being built are set based on the overall rmsd of the map in the region containing the macromolecule. If there are some parts of the molecule that are more poorly defined, then these parts might never be built as the density is not high enough in that region. By masking off the region of the molecule that has been built already, the thresholds can be more reasonably determined for the remaining region containing the macromolecule. Additionally by focusing on a small region of the map where no model has been built, an extensive search for helices and strands can be carried out in a reasonable amount of computing time. A partial model containing segments of the model that can be built outside the region containing the current model is then added to the current set of working models, and is recombined with the other working models as in step (2) above.

Two methods are used to attempt to build loops. One method is to identify all pairs of C-termini and N-termini of existing chains that are near each other (typically within 15 Å), and to try to extend the C-terminus of the first chain and the N-terminus of the second chain in a way that leads to at least one amino acid overlapping with a low rmsd of main-chain atoms (typically 1 Å). All such connecting segments that are found are then added as if they were another partial model to the current set of working models as in step (2) above. The second method for building loops is to use the sequence alignment of the current best model to identify short segments that are missing from the chain, and to use the above method to try and fill in the loop. This method differs from the case where the sequence alignment is unknown in that the precise number of amino acids in the loop is known. Once a set of loops has been built, a new model is created by grafting these loops on to the current best model, creating a new model with the loops built. This model is then recombined with the other working models as in step (2) above.

2.3. Rebuilding an existing model

The AutoBuild Wizard has two procedures for rebuilding a model. One is to build a model from scratch exactly as described above, except recombining the best parts of the model to be rebuilt with the new model during that building process. The second procedure for rebuilding

a model is quite different; this is the “rebuild-in-place” procedure in which an existing model is rebuilt in segments without adding or deleting residues.

The rebuild-in-place procedure has the advantage that no parts of the model are “lost” in rebuilding, but the disadvantage that no new model is built. It is best suited to situations where the model is essentially complete and close to correct, yet significant local main-chain corrections need to be made to improve the model. The rebuild-in-place procedure is based on the loop-fitting algorithms described above, combined with a procedure for recombination of two chains that have different conformations but are aligned and have the same residues. The rebuild-in-place option is well suited to the rebuilding of high-sequence similarity models derived from molecular replacement.

In the first step of the rebuild-in-place procedure, the rebuild-in-place method in RESOLVE is used to sequentially rebuild overlapping segments of the model. A segment, typically 6 residues long, is removed from the model. Then the loop-fitting algorithm described above is used to rebuild this segment, maintaining the identities of the residues in the loop and the length of the loop. During the loop-fitting process, the orientations of the residues at the two ends of the resulting gap are varied slightly by randomizing the coordinates of the main-chain atoms of these residues by a small distance (typically an rmsd of 0.5 Å). As the loop residue positions are generated by extending from the last amino acid in the chain, this randomization has the effect of introducing diversity into the loops that are created. If a new loop conformation can be found, it is used to replace the existing loop. If no acceptable conformation is found, the existing loop is maintained. The process is repeated, offsetting the loop-building by 5 residues at a time, until the entire model (except the very ends of each chain) has been rebuilt. In the second step of rebuilding in place, the model created by rebuilding overlapping segments is recombined with the original model, taking the best-fitting segments of each model. This crossover process is carried out by aligning the two models, identifying all the places where crossover can occur as corresponding C_α atoms that are within a small distance of each other (typically 0.5 Å), and choosing whichever model has the higher local map correlation for each segment of the model from one crossover point to the next. Once a recombined model is obtained, side-chains are rebuilt using a highly-curated rotamer library (Lovell et al., 2000) instead of the rotamer libraries originally created for RESOLVE model-building (Terwilliger, 2003a).

2.4. Refinement with phenix.refine

A complete description of the phenix.refine program will be published elsewhere; here we outline the features used in the automated model building procedures in the AutoBuild Wizard. Depending on the quality of the initial electron density map, the models undergoing refinement may be quite incomplete and contain significant coordinate and/or displacement

parameter errors. Therefore, methods described here have been designed to be fault tolerant, a necessary requirement of an automated procedure. Firstly, a robust automatic bulk solvent correction and anisotropic scaling procedure is used to account for the scattering from disordered solvent in the crystal and correct for any anisotropic diffraction (Afonine, Grosse-Kunstleve & Adams, 2005a). Coordinate refinement is performed by LBFGS minimization (Liu & Nocedal, 1989) of the target function E_{xyz} w.r.t. atomic coordinates, while keeping all other parameters fixed. E_{xyz} can be a Least-Squares target (LS; Afonine, Grosse-Kunstleve & Adams, 2005a), an amplitude-based Maximum-Likelihood target (ML; Afonine, Grosse-Kunstleve & Adams, 2005a) or a Phased Maximum-Likelihood target (MLHL; Pannu *et al.*, 1998). In the refinement of Atomic Displacement Parameters (ADP) the target E_{adp} is minimized w.r.t. isotropic *ADP*s while all other model parameters are fixed. E_{adp} is defined as:

$$E_{adp} = \sum_{i=1}^{N_{atoms}} \left[\sum_{j=1}^{M_{atoms}} \frac{1}{r_{ij}^{k}} \frac{\left(B_i - B_j\right)^2}{B_i + B_j} \right] \quad (1)$$

Here N_{atoms} is the total number of atoms in the model, the inner sum is extended over all M_{atoms} in the sphere of radius R around atom i, r_{ij} is a distance between two atoms i and j, B_i and B_j are the corresponding isotropic ADPs and k is user-defined constant. By default, R and k are fixed at 5.0Å and 1.0, respectively, but they can also be refined. This target function makes use of the following ideas:

- A bond is almost rigid, therefore the ADPs of bonded atoms are similar (Hirshfeld, 1976);
- ADPs of spatially close (non-bonded) atoms are similar (Schneider, 1996);
- The bond rigidity, and therefore the difference between the ADPs of bonded atoms, is related to the absolute values of the ADPs. Atoms with higher ADPs can have larger differences (Ian Tickle, CCP4 Bulletin Board).

We have implemented a completely automated protocol for updating the ordered solvent model during the refinement process. If requested by the user (and by default in the AutoBuild Wizard), waters are updated (added and removed) in each macro cycle. In the same macro cycle, the complete structure including the updated water structure is subject to coordinate and ADP refinement. Updating the ordered solvent model involves the following steps:

- Elimination of waters present in the initial model based on user-defined cutoff criteria by ADP, occupancy and inter-atomic distances (water-water, macromolecule-water).
- Location of peaks in a cross-validated likelihood-weighted difference map (Read, 1986; Urzhumtsev *et al.*, 1996).
- Confirmation of peaks found in the previous step using a 2*m*Fobs-αFcalc map.

- Elimination of peaks in regions occupied by the macromolecule as determined by the current bulk-solvent mask.
- Elimination of peaks too close to each other (the default minimum distance is 2.0 Å; the strongest peak of two close peaks is retained).
- Elimination of peaks too close to macromolecular atoms (the default minimum distance is 1.8 Å).
- Elimination of peaks too far away from macromolecular atoms (the default maximum distance is 6.0 Å).
- Elimination of peaks based on the evaluation of tabulated empirical distance distributions derived from the analysis of high-resolution models in the PDB (Afonine, Grosse-Kunstleve & Adams, 2005a). Distance distributions between water oxygen and macromolecular C, N and O atoms are tabulated. Only peaks with a good fit to at least one distance distribution are retained.

It is not uncommon for macromolecular crystal structures to have more than one copy of a molecule in the asymmetric unit, generating some form of non-crystallographic symmetry (NCS). This symmetry is exploited in the model building procedure, and can also be used in the refinement of the structures in phenix.refine. Briefly, the sequence of the input model is subject to pairwise sequence alignment (Needleman & Wunsch, 1970; Smith & Waterman, 1981) to identify similar molecules in the model. If any relationships are found least-squares superposition of the structures is performed (Kearsley, 1989) and the coordinate deviation calculated. If the root-mean-square deviation between the coordinates is below a user-specified tolerance (default: 3.0 Å) NCS restraints are applied to the related coordinates during structure refinement. The default NCS restraints are very tight (0.1 Å for both main-chain and side-chain NCS-related pairs).

2.5. Iterative model-building beginning from an experimental map

The AutoBuild Wizard has one procedure for initial iterative model-building beginning from an experimental electron density map, and a second procedure for iterative rebuilding and completion of an initial model. These procedures are based in part on the “build” and “rebuild” procedures in the RESOLVE model-building script (Terwilliger, 2003b), though they contain additional steps as described above.

The procedure for model-building from an experimental electron density map consists of cycles of two basic steps. These steps are (a) using experimental phase information and any additional phase information available in statistical density modification (Terwilliger, 2000) to create a new working electron density map, and (b) building and refining a model based on this new map as described in Section 2.2 above.

For density modification, several sources of additional phase information are used when available. One is any non-crystallographic symmetry information (NCS) as implemented in RESOLVE (Terwilliger, 2002). NCS is deduced from coordinates of heavy-atom sites if available, and also directly from the current atomic model of the macromolecule as described above if the sequence has been aligned. A second source of information is the presence of recognizable local patterns of density in the electron density map (Terwilliger, 2003c), and a third is the presence of density matching a helical or strand template in the map (Terwilliger, 2001). A fourth source of information consists of any partial models of the macromolecule that have been built. For the purpose of identifying patterns in the electron density map, a composite omit map is produced each cycle in which model information is excluded from the omit region (Terwilliger, 2003b).

The approach used to carry out density modification in this "build" procedure has several steps. First, electron density information from local patterns of density and helical and strand locations are combined. Both the identification of local patterns of density and identification of helical and strand fragment procedures result in a pseudo-electron density map with density that has some information about the true electron density map. Relative weights for these maps are chosen such that the weighted average pseudo-electron density map has the highest possible correlation with the current working map. The resulting pseudo-electron density map is then used as a target for statistical density modification in the same way that NCS and model-based information is incorporated, except that the uncertainty associated with this target map is arbitrarily set to a very high value (typically the rmsd of the current electron density map) so as not to overly emphasize this information (Terwilliger, 2003b, Terwilliger, 2003c).

The phase probability distributions obtained are then used as prior phase information in a second density modification step that includes model information as well as any NCS and solvent flattening information (Terwilliger, 2003b). The models obtained in any previous cycles are used to calculate a composite target model map (Terwilliger, 2003b), and the target model map is scaled to match the working map as closely as possible, including only grid points near the positions of atoms in the model (typically within 2.5 Å). The rmsd between the working map and this target map at these grid points is used as the uncertainty for the values in the target map in statistical density modification (Terwilliger, 2003b). The map obtained from this statistical density modification procedure is then used for model-building.

2.6. Iterative density modification, model-building, and refinement beginning from a model

The AutoBuild Wizard procedure for iterative model-building beginning from a partial model is similar to the procedure starting from experimental phase information, but there are

differences resulting from the fact that the phases available are coming from a partial model. Due to model bias, the methods for identifying local patterns of density and for finding helices and strands used when starting from experimental phase information are not effective and are therefore skipped. Additionally, the starting map used in the final density modification step comes from the model and not from experimental phases.

The procedure for density modification beginning from a model uses model-based phase probabilities as the starting point for density modification. A composite target map is calculated from any models available from previous cycles, just as in the procedure described in section 2.4. This map is then used as a target for statistical density modification, using the same procedure for calculating uncertainties in the target density that was used for the incorporation of pattern-based density information in section 2.4 (i.e., simply using the rmsd of the map as the uncertainty). The resulting phases and map are then used in statistical density modification including NCS and solvent flattening, yielding a density-modified, model-based electron density map. This map is then used as a starting point for density modification that includes model information as well as any NCS and solvent flattening information as described in section 2.4. The prior phase probabilities for this density modification step consist only of any experimental phase information that is available (so there is no prior phase information in cases where the rebuilding is done with no experimental phase information).

3. Results and Discussion

3.1. Outline of AutoBuild Wizard operation

A schematic of the operation of the AutoBuild Wizard for a case where experimental phase information is available is shown in Fig. 1. The Wizard begins with experimental structure factor amplitudes and estimates of crystallographic phases, optionally encoded as Hendrickson-Lattman coefficients (Hendrickson & Lattman, 1979). The phase information is improved by using statistical density modification to improve the correlation of NCS-related density in the map (if present) and to improve the match of the distribution of electron densities in the map with those expected from a model map (Terwilliger, 2000). This improved map is then used to build and refine an atomic model. In subsequent cycles, the models from previous cycles are used as a source of phase information in statistical density modification, iteratively improving the quality of the map used for model-building. Additionally, during the first few cycles additional phase information is obtained by detecting and enhancing (1) the presence of commonly-found local patterns of density in the map, and (2) the presence of density in the shape of helices and strands. The final model obtained is analyzed for residue-based map correlation (Branden & Jones, 1990) and density at the

coordinates of individual atoms, and an analysis including a summary of atoms and residues that are in strong, moderate, or weak density and out of density is provided.

3.2. Application of AutoBuild Wizard to model-building beginning from structure factor amplitudes and experimental estimates of phases

We have developed and tested the AutoBuild Wizard by using it to build atomic models for structures in the PHENIX structure library where experimental phase information (MIR, MAD, or SAD) was available. In each case the structure had been solved previously and an atomic model was available. The PHENIX AutoSol Wizard was used to (re)-solve the structure, and the AutoBuild Wizard was then used with default settings to iteratively build and refine a model.

Figure 2A illustrates the R-factors and free R-factors obtained in this test on 48 MAD, SAD, and MIR structures at resolutions ranging from 1.1 Å to 3.2 Å. The median R-factor for these 48 structures is 0.22, and the median free R-factor is 0.28; the corresponding means are 0.24 and 0.29, respectively. Somewhat surprisingly, the R-factors and free R-factors do not have a strong dependence on resolution. They do, however, have a strong dependence on the quality of the starting density-modified electron density map. This is illustrated in Figure 2B, which shows the R- and free R-factors from Fig. 2A plotted as a function of the correlation coefficient of this map with a model map calculated from the known structure. Fig. 2C shows the same data as Fig. 2A, except that only the structures built beginning with the highest-quality starting maps (with map correlation to that of the known structure at least 0.85) are shown. Fig. 3C also shows little relationship between R-factor and resolution. Taken together, the data in Fig. 2 indicate that a key determinant of the overall correctness of the models produced by the AutoBuild wizard, as assessed by R- and free R-factors, is the quality of the starting density-modified experimental map, and that the resolution of the structure has a much smaller effect.

Figures 3A and 3B illustrate the completeness of the models obtained, as a function of the resolution of the data and of the quality of the starting density-modified electron density map. The median percentage of residues built is 95% and the median percentage of residues assigned to sequence is 90% (means of 90% and 78%, respectively). The percentage of residues built depends more on the quality of the starting map than on the resolution of the data, though neither of these variables very closely correlates with the completeness of the models. Fig. 3C illustrates that the completeness of the models is somewhat related to the resolution of the data in cases where a high-quality (map CC > 0.85) starting density-modified map was available, but only weakly so.

It seemed likely that the resolution of the data would have a significant influence on the details of the atomic model, even if the overall correctness of the model as measured by R-

factors and completeness was not strongly resolution-dependent. Figures 4A and 4B show the rms differences between the coordinates of atoms in the AutoBuild models and those of the models previously obtained for the same structures. In Fig. 4A, these are plotted as a function of the resolution of the data, and in Fig. 4B they are plotted as a function of the quality of the starting electron density maps. Surprisingly, there is not a strong relationship between the resolution of the data and the rmsd between the models obtained and those obtained previously for these structures. The median value of the rmsd of main-chain atoms for structures based on data from 1.1-1.9 Å is 0.57 Å, while the corresponding value for structures based on data from 2.0-3.2 Å is 0.47 Å. There is a weak correlation (Fig. 4B) between the ability of the AutoBuild Wizard to reproduce the previously-obtained structural models and the quality of the starting map. When only structures beginning with a high-quality map (map CC > 0.85) are considered (Fig. 4C), there is a weak relationship between resolution of the data and the rmsd between the models built by the AutoBuild Wizard and the previously-built models.

3.3. Application of AutoBuild Wizard to model-rebuilding starting from molecular replacement solutions

The AutoBuild wizard was applied to structure rebuilding of a model derived from molecular replacement. A number of different criteria can be applied to estimate the success of molecular replacement; correlation coefficients for the MR solution, and free R-values after an initial round of refinement are two commonly used approaches. A more stringent test is the application of model rebuilding using automated methods; for example ARP/wARP (Perrakis et al., 1999) or the PHENIX AutoBuild wizard described here. If a molecular replacement solution can be rebuilt without manual intervention, yielding a new model that has reasonable chemical structure while also showing differences from the starting model, it can be reasonably concluded that the MR solution is correct. To test this hypothesis we performed molecular replacement and subject the resulting structure to automated model rebuilding. The experimental data, to a maximal resolution of 2.4 Å, for α_{2u}-globulin (accession code 2A2U; Chaudhuri *et al.*, 1999) was obtained from the Protein Data Bank. A single monomer of the mouse urinary protein structure (accession code 1JV4; Kuser et al., 2001) was used as a search model. Molecular replacement, searching for the four molecules in the asymmetric unit, was performed using PHASER (Storoni, McCoy & Read, 2004; McCoy *et al.*, 2005) within the PHENIX AutoMR wizard. A clear solution for all four molecules was found. From this solution three other models were created: one monomer was removed to generate a 75% complete model, two monomers were removed to generate a 50% complete model, and the whole tetramer was randomly rotated and translated to generate an incorrectly placed but complete model. Each model was input to the AutoBuild wizard and the success monitored

by the final free R-value and the number of residues built (Table 1). When the MR solution is correct and complete, or correct and 75% complete it is possible to arrive at a close to complete model with the correct amino acid sequence after automated building with the PHENIX AutoMR wizard. When the MR solution is incorrect it is not possible to rebuild the model, as indicated by the R-factors and the number of residues built. When the model is correct but incomplete (50%), or complete and partially (50%) incorrect [you didn't introduce this case above!], automated building is unable to recover the missing or incorrectly place parts owing to the large initial phase error from the input coordinates.

4. Conclusions

The AutoBuild Wizard has been developed as a highly automated tool for building and refining macromolecular structures. This procedure can be equally well applied to phases derived from isomorphous/anomalous and molecular replacement methods. In the case of molecular replacement the success of automated model building is a strong indicator of the correctness and completeness of the molecular replacement solution.

We have found that the AutoBuild Wizard can yield highly complete and well-refined models, with half of the structures in our sample built to at least 95% completeness and the worst built to 58% completeness. Somewhat surprisingly, the final R-factors and free R-factors depended little on the resolution of the data and much more strongly on the quality of the starting density-modified electron density map. These results are encouraging for the prospects of generating even more complete models at moderate resolution.

There remain many aspects of model completion that are not yet fully implemented in the AutoBuild Wizard. These include building model for regions that are poorly ordered and those that are well-ordered but containing multiple conformations. Other aspects that are not implemented are the validation of models, the editing of models to remove segments unlikely to be correct, and automated placement of ligands. The extension of automated model-building and refinement to resolutions lower than about 3.2 Å presents challenges in model-building as well, although recent developments suggest that this difficulty may be surmountable (deMaio et al., 2006).

Table 1 Results of the AutoBuild Wizard applied to molecular replacement models derived for α_{2u}-globulin using data to 2.4 Å. R-factors were calculated using phenix.refine. The deposited α_{2u}-globulin structure contains 688 amino acid residues arranged in four chains of 172 residues each. Calculation of R-factors for the deposited model (2A2U), using all available data and the deposited cross-validation set, after one cycle of refinement with phenix.refine produced R- and free R-factors of 0.196 and 0.235 respectively.

Figure 1 Outline of AutoBuild Wizard operation beginning from experimental phase information.

Figure 2 R-factors (closed diamonds) and free R-factors (open triangles) of final models obtained with the AutoBuild Wizard for 48 structures from the PHENIX structure library beginning with experimental phases obtained with the AutoSol Wizard. A. R- and free R-rfactors as a function of resolution of the data used in modelling. B. R- and free R-factors as a function of the correlation coefficient of the starting density-modified experimental map. C. R-factors as in A, except that only structures with a correlation coefficient of the starting density-modified experimental map of greater than 0.85 are included.

Figure 3 Completeness of main-chain model (closed diamonds) and assignment of residues to sequence (open triangles) of final models in Fig. 2. A. Completeness as a function of resolution of the data used in modelling. B. Completeness as a function of the correlation coefficient of the starting density-modified experimental map to the model map. C. Completeness as a function of resolution as in A, except that only structures with a correlation coefficient of the starting density-modified experimental map of greater than 0.85 are included.

Figure 4 Main-chain (closed diamonds) and side-chain (open triangles) rmsd of final models in Fig. 2 compared to refined models previously obtained for the same structures. A. Rmsd as a function of resolution of the data used in modelling. B. Rmsd as a function of the correlation coefficient of the starting density-modified experimental map. C. Rmsd as in A, except that only structures with a correlation coefficient of the starting density-modified experimental map of greater than 0.85 are included.

Acknowledgements The authors would like to thank the NIH for generous support of the PHENIX project (1P01 GM063210). This work was partially supported by the US Department of Energy under Contract No. DE-AC02-05CH11231. RJR is supported by a Principal Research Fellowship from the Wellcome Trust (UK). The algorithms described here are available in the PHENIX software suite (http://www.phenix-online.org).

References

Adams, P.D., Grosse-Kunstleve, R.W., Hung, L.-W., Ioerger, T.R., McCoy, A.J., Moriarty, N.W., Read, R.J., Sacchettini, J.C., Sauter, N.K., Terwilliger, T.C. (2002). Acta Cryst. D58, 1948-1954.

Afonine, P.V., Grosse-Kunstleve, R.W, Adams, P.D. (2005a). Acta Cryst. D61, 850-855.

Afonine, P.V., Grosse-Kunstleve, R.W, Adams, P.D. (2005b). CCP4 newsletter, July 2005. Contribution 8.

Branden, C.I., Jones, T. A. (1990). Nature 343, 687-689.

Chaudhuri, B. N., Kleywegt, G. J., Bjorkman, J., Lehman-McKeeman, L. D., Oliver, J. D. & Jones, T. A. (1999). Acta Cryst. D55, 753-762.

DiMaio F., Shavlik J., Phillips G.N. (2006). Bioinformatics 22, e81-e89.

DePristo, M.A., de Bakker, P.I.W., Johnson, R.J.K., Blundell, T.L. (2005). Structure 13, 1311-1319.

Grosse-Kunstleve, R.W., Sauter, N.K., Adams, PD. (2004). IUCr Computing Commission Newsletter 3, 22-31.

Hendrickson, W.A. & Lattman, E.E. (1979). Acta Cryst. B26, 136-143.

Hirshfeld, F.L. (1976). *Acta Cryst.* A**32**, 239-244

Ioerger, T.R, Sacchettini, J.C. (2003). Methods Enzymol. 374, 244-270.

Jensen, L. H. (1997). *Methods Enzymol.* 277, 353-366.

Kearsley, S.K. (1989). Acta Cryst. A45, 208-210.

Kuser, P. R., Franzoni, L., Ferrari, E., Spisni, A. & Polikarpov, I. (2001). Acta Cryst. D57, 1863-1869.

Liu, D.C. & Nocedal, J. (1989). *Mathematical Programming*, **45**, 503-528.

Lovell, S.C., Word, J.M., Richardson, J.S., Richardson, D.C. (2000). Proteins: Struct. Func. Genet. 40, 389-408.

McCoy, A.J., Grosse-Kunstleve, R.W., Storoni, L.C. & Read, R.J. (2005). *Acta Cryst* D**61**, 458-464

Needleman, S. & Wunsch, C. (1970). J. Mol. Biol. 48(3), 443-53.

Ondráček J. (2005). Acta Cryst A61, C163.

Pannu, N. S., Murshudov, G. N., Dodson, E. J. & Read, R. J. (1998). *Acta Cryst*. **D54**, 1285-1294.

Perrakis, A., Morris, R., Lamzin, V.S. (1999). Nature Struct. Biol. 6, 458-463.

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

Schneider, T. (1996). *Proceedings of the CCP4 Study Weekend*. SERC Daresbury Laboratory, Daresbury, U.K., 133-144.

Smith, T. F. & Waterman M. S. (1981). J. Mol. Biol. 147, 195-197.

Storoni, L.C., McCoy, A.J. & Read, R.J. (2004). *Acta Cryst* D**60**, 432-438

Terwilliger, T. C. (2000). Acta Cryst. D56, 965-972.

Terwilliger T. C. (2001). Acta Cryst. D57, 1755-1762.

Terwilliger, T. C. (2002). Acta Cryst. D58, 2082-2086.

Terwilliger. T. C. (2003a) Acta Cryst. D59, 38-44.

Terwilliger, T. C. (2003b). Acta Cryst. D59, 1174-1182.

Terwilliger, T. C. (2003c) Acta Cryst. D59, 1688-1701.

Urzhumtsev, A., Skovoroda, T.P. & Lunin, V.Y. (1996). *J.Appl.Cryst*., **29**, 741-744.

Table 1

Starting Model	Final R-factors (R-free/R)	Residues Built	RMSD (Å) (Number of Matching Residues)
Correct, 100% complete	0.232/0.196	628	0.3 (628)
Correct, 75% complete	0.252/0.205	567	0.2 (528)
Correct, 50% complete	0.472/0.410	265	0.45 (58)
Correct 50%, Incorrect 50%			
Incorrect, 100% complete	0.524/0.456	20	N/A (0)

Figure 1

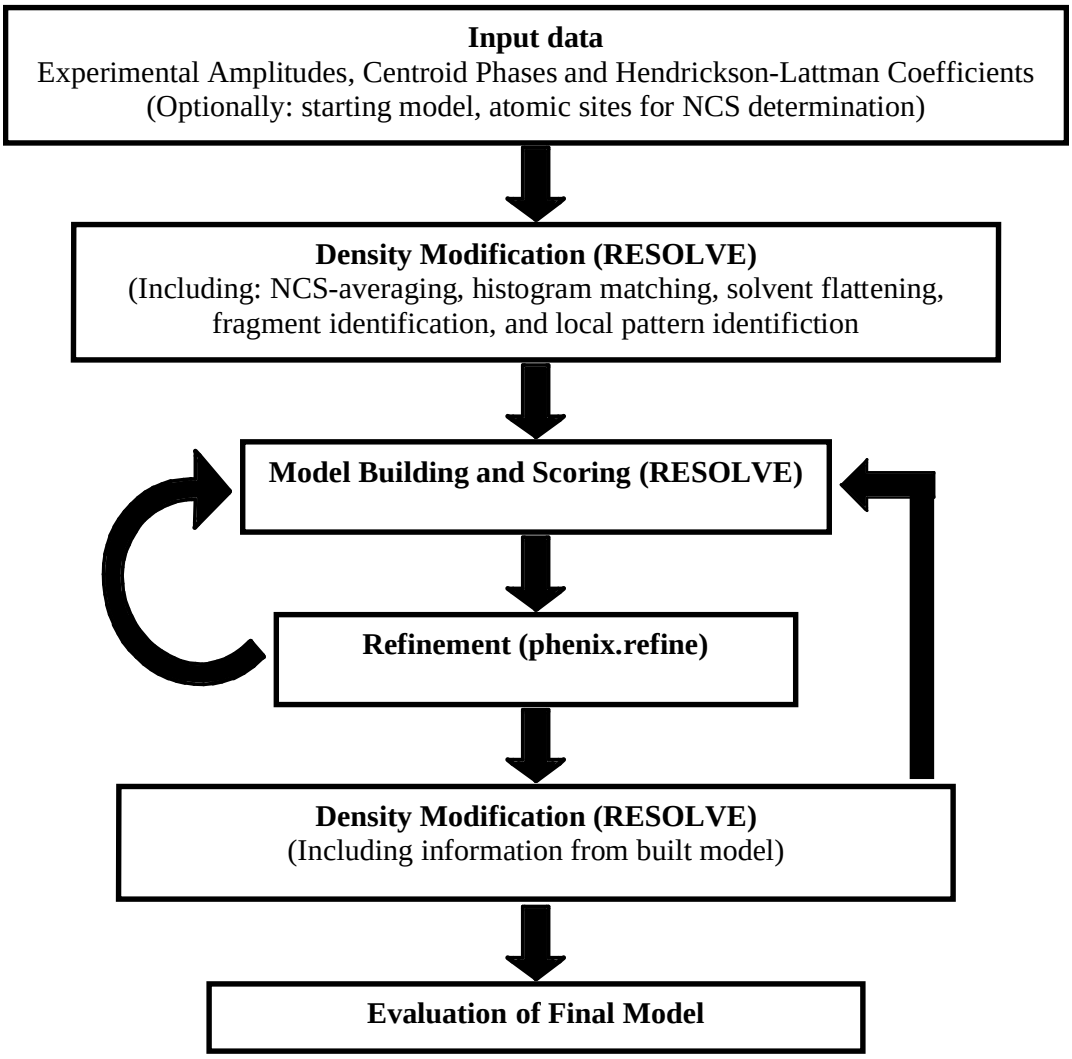

Input data
Experimental Amplitudes, Centroid Phases and Hendrickson-Lattman Coefficients
(Optionally: starting model, atomic sites for NCS determination)
Density Modification (RESOLVE)
(Including: NCS-averaging, histogram matching, solvent flattening,
fragment identification, and local pattern identifiction
Model Building and Scoring (RESOLVE)
Refinement (phenix.refine)
Density Modification (RESOLVE)
(Including information from built model)
Evaluation of Final Model

Figure 2A

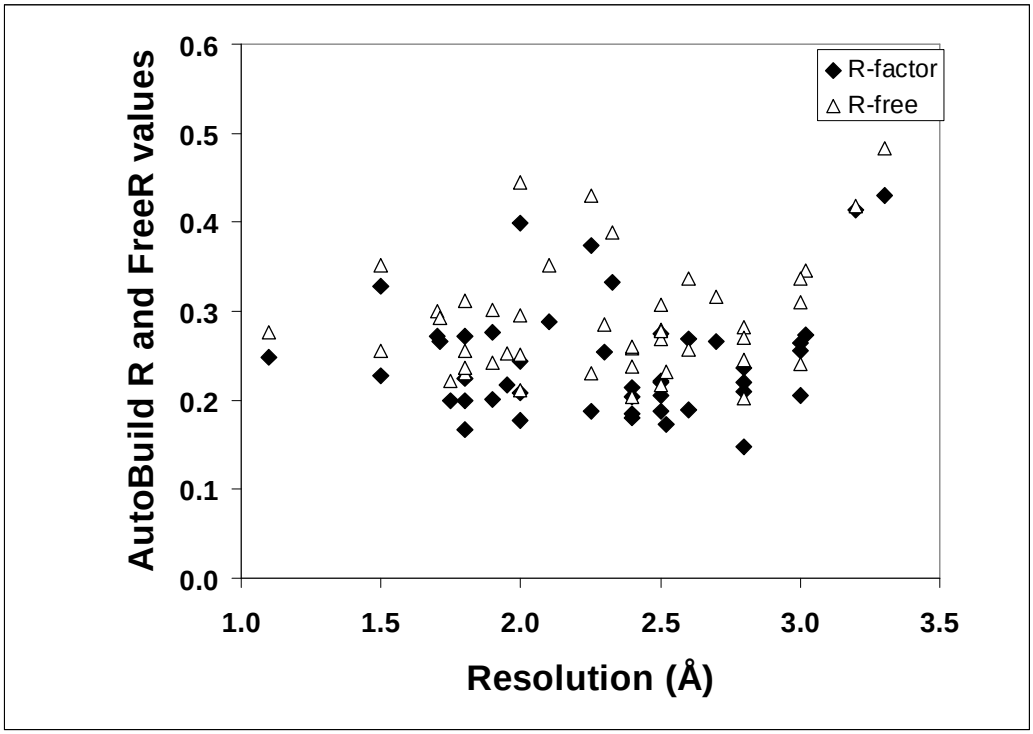

AutoBuild R and FreeR values
0.6
0.5
0.4
0.3
0.2
0.1
0.0
1.0
1.5
2.0
2.5
3.0
3.5
Resolution (Å)
R-factor
R-free

Figure 2B

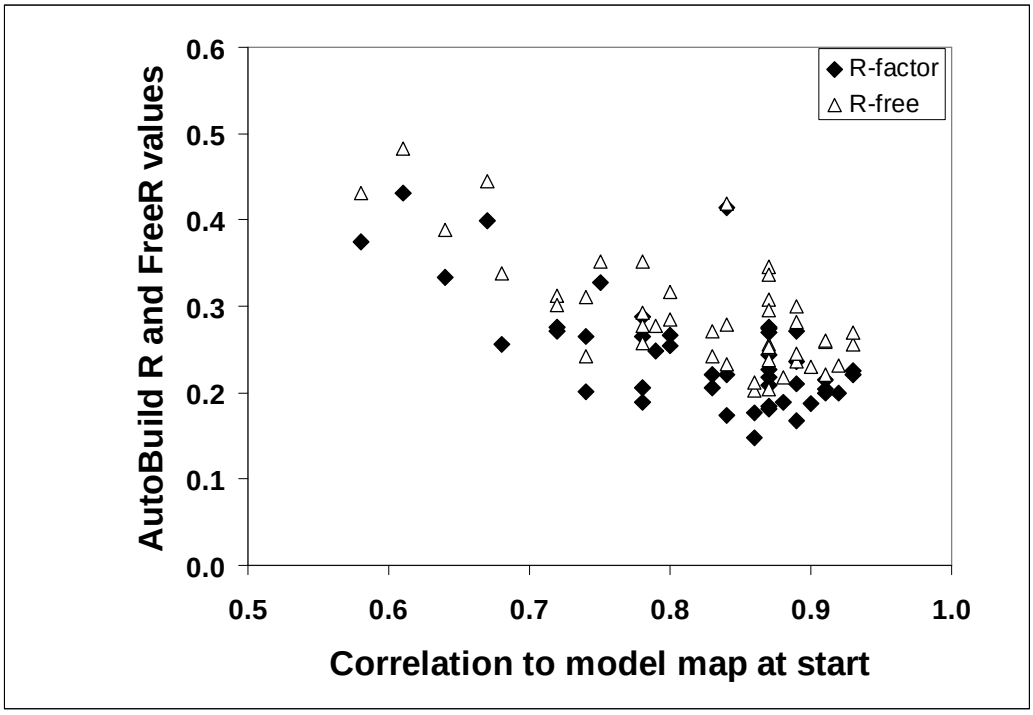

AutoBuild R and FreeR values
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.5
0.6
0.7
0.8
0.9
1.0
Correlation to model map at start
R-factor
R-free

Figure 2C

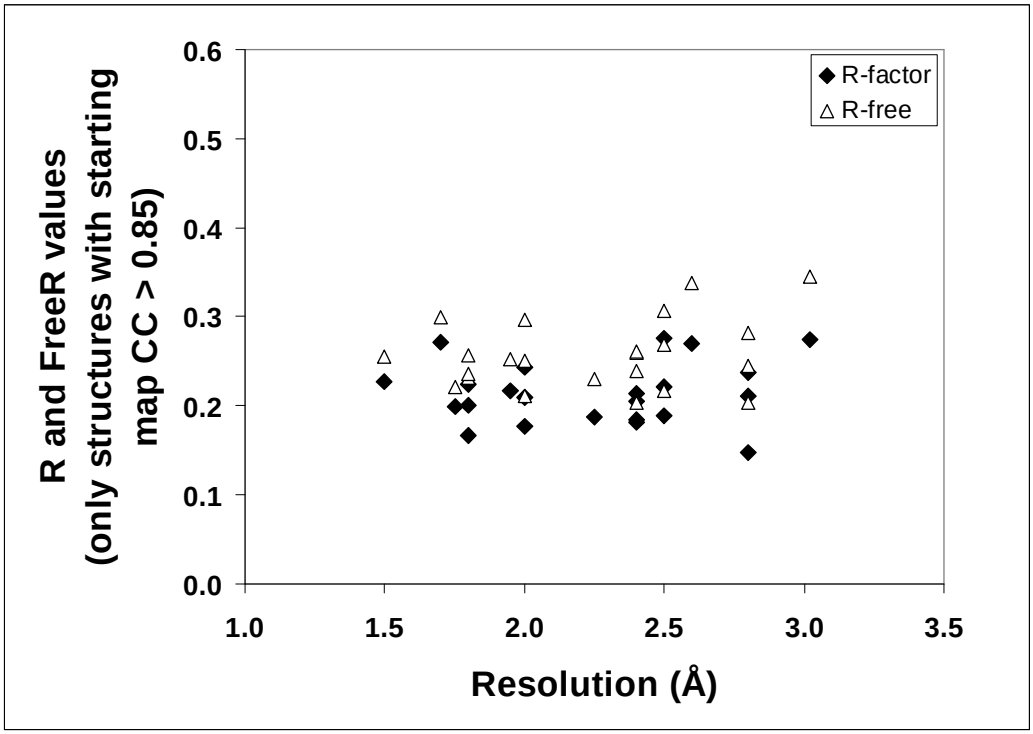

R and FreeR values
(only structures with starting map CC > 0.85)
0.6
0.5
0.4
0.3
0.2
0.1
0.0
1.0
1.5
2.0
2.5
3.0
3.5
Resolution (Å)
R-factor
R-free

Figure 3A

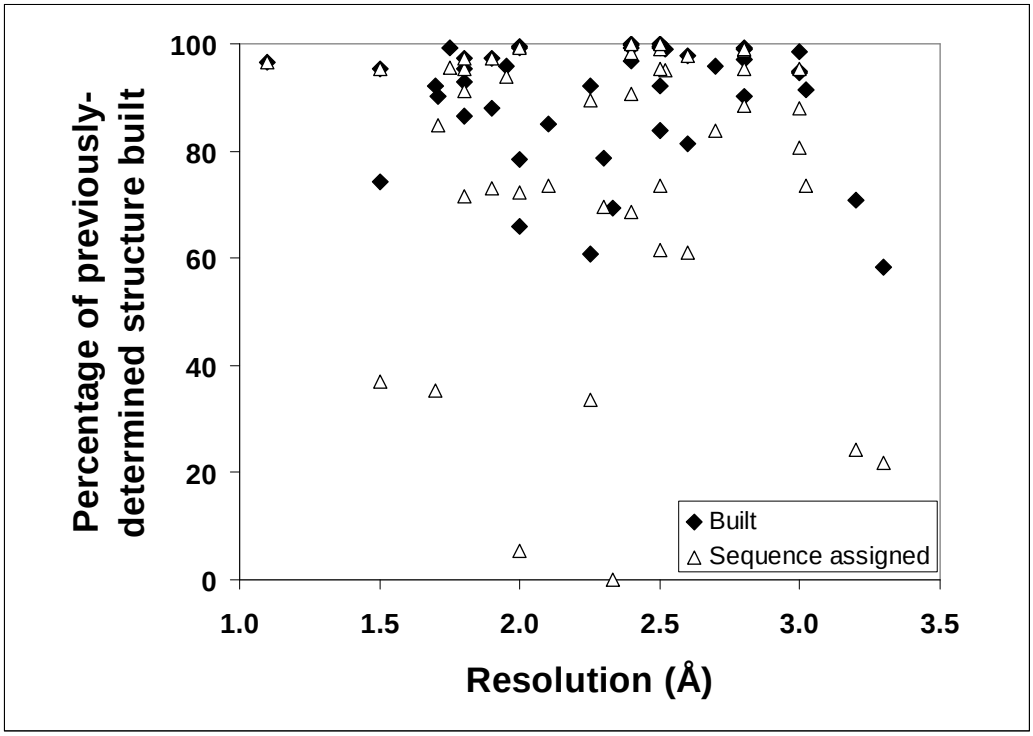

Percentage of previously-determined structure built
100
80
60
40
20
0
1.0
1.5
2.0
2.5
3.0
3.5
Resolution (Å)
Built
Sequence assigned

Figure 3B

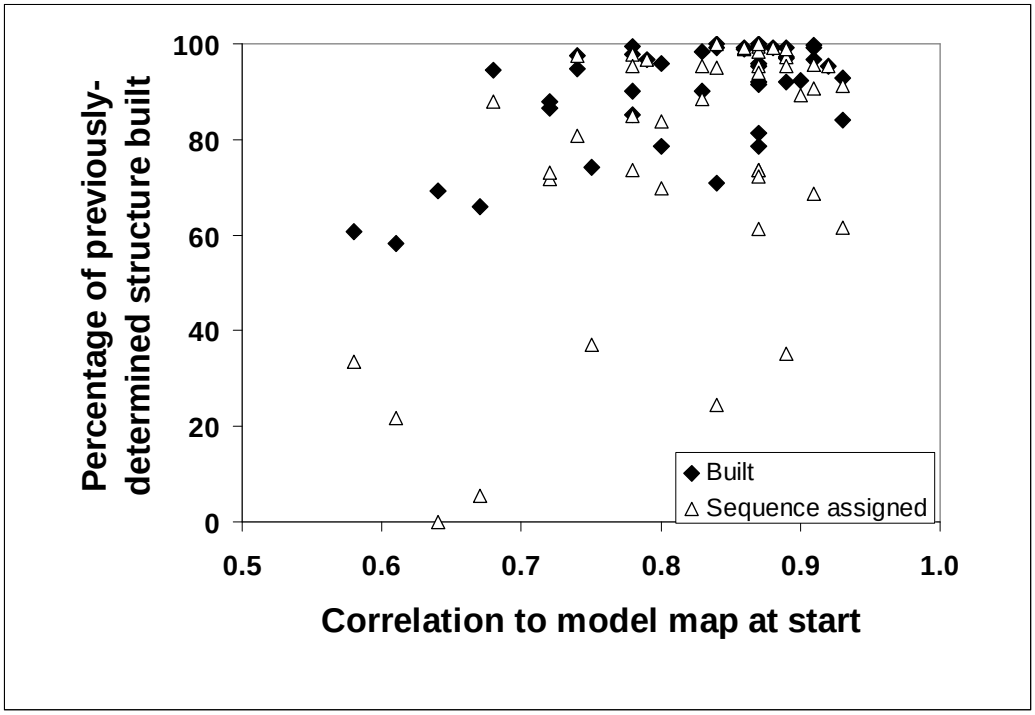

Percentage of previously-determined structure built
100
80
60
40
20
0
0.5
0.6
0.7
0.8
0.9
1.0
Correlation to model map at start
Built
Sequence assigned

Figure 3C

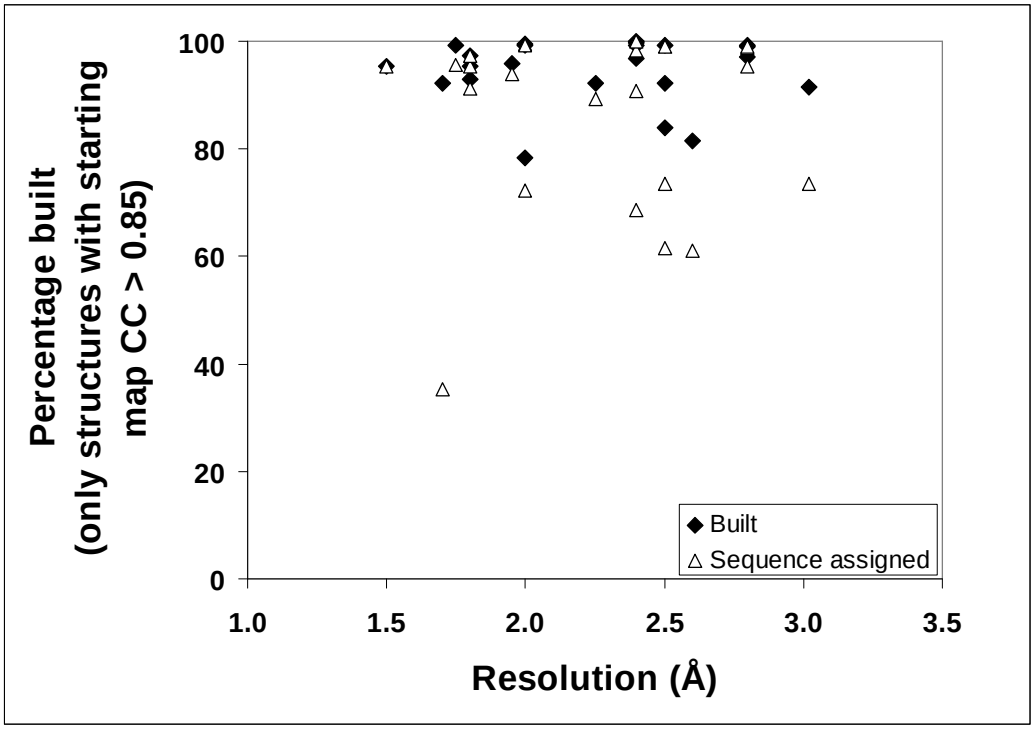

Percentage built
(only structures with starting
map CC > 0.85)
100
80
60
40
20
0
1.0
1.5
2.0
2.5
3.0
3.5
Resolution (Å)
Built
Sequence assigned

Figure 4A

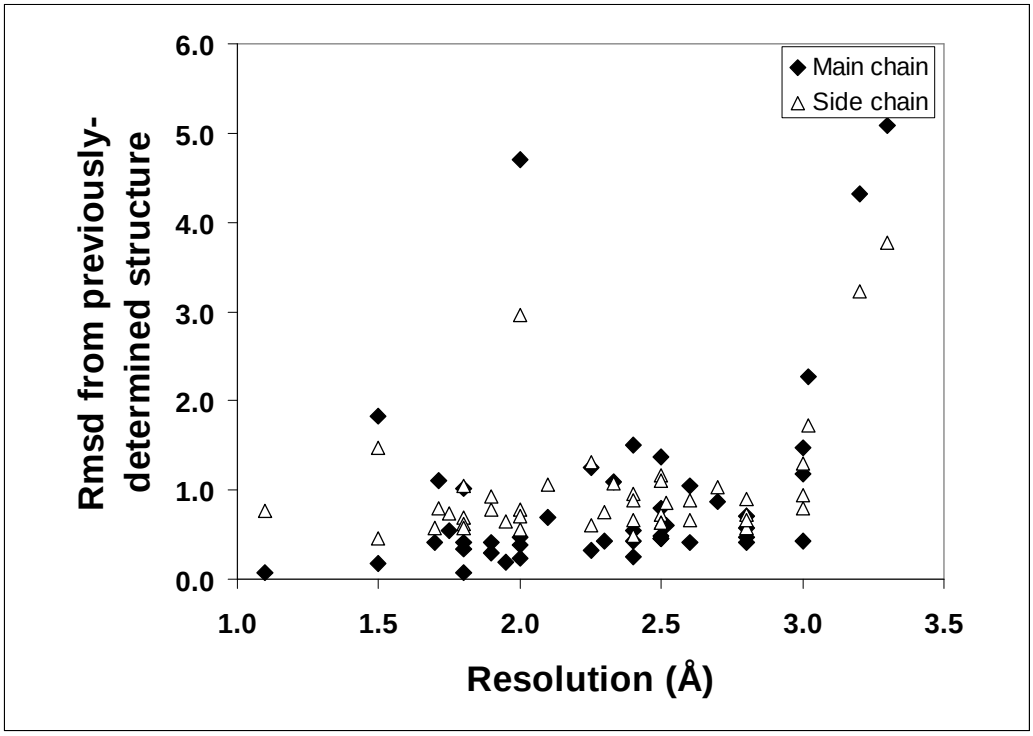

Main chain
Side chain
Rmsd from previously-determined structure
6.0
5.0
4.0
3.0
2.0
1.0
0.0
1.0
1.5
2.0
2.5
3.0
3.5
Resolution (Å)

Figure 4B

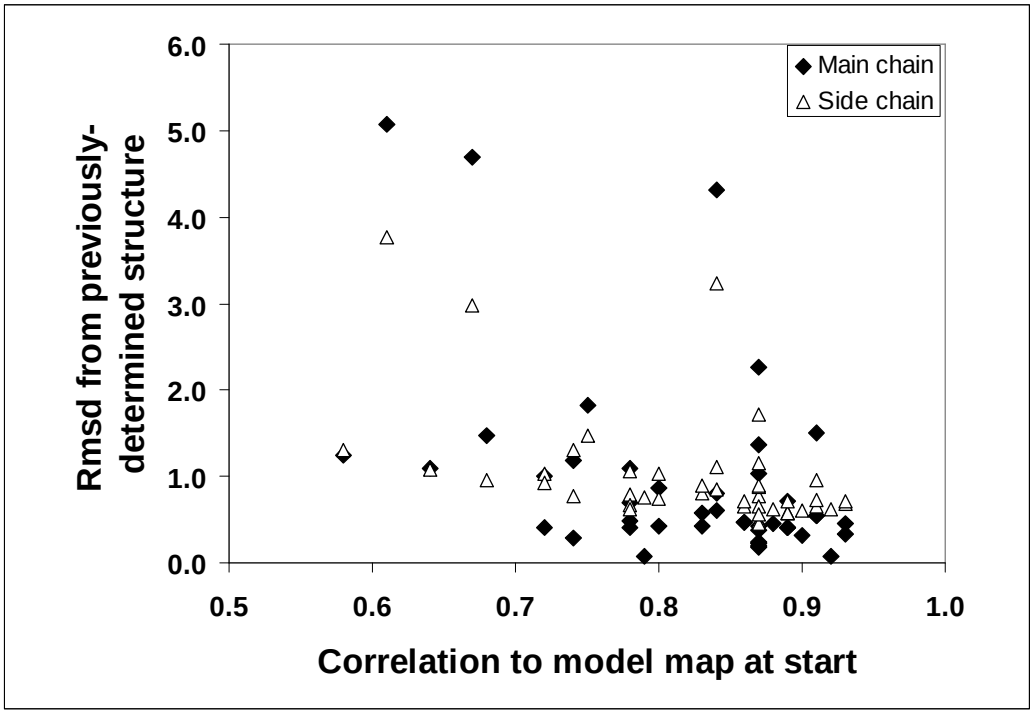
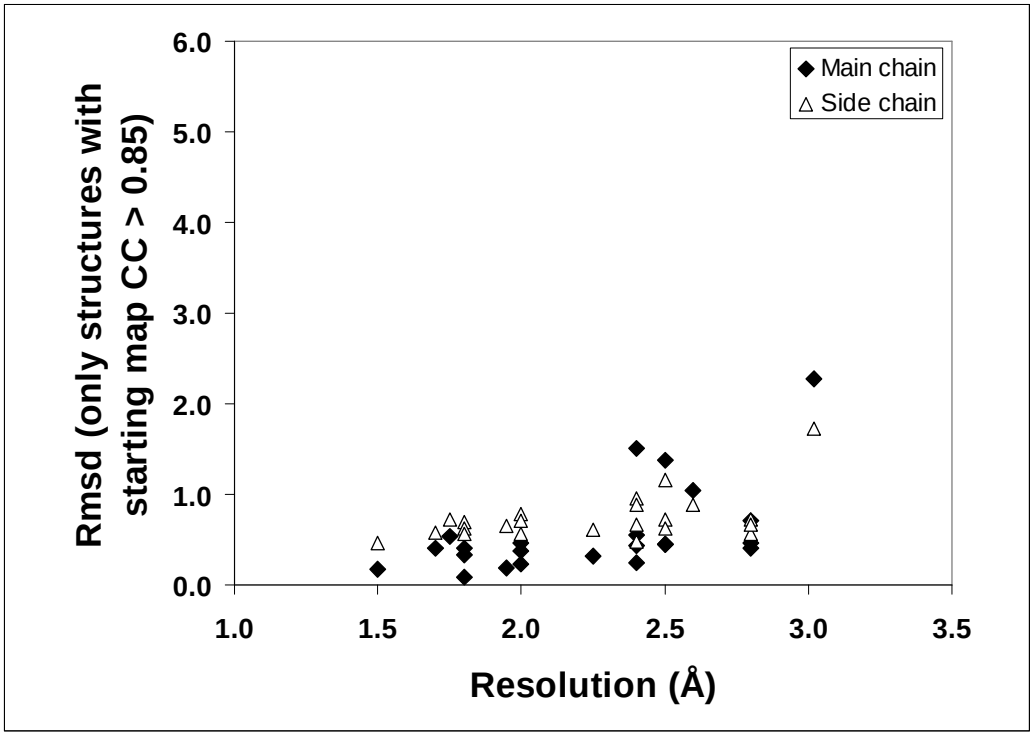

Main chain
Side chain
Rmsd from previously-determined structure
6.0
5.0
4.0
3.0
2.0
1.0
0.0
0.5
0.6
0.7
0.8
0.9
1.0
Correlation to model map at start

Figure 4C


Rmsd (only structures with starting map CC > 0.85)
Resolution (Å)
Main chain
Side chain
6.0
5.0
4.0
3.0
2.0
1.0
0.0
1.0
1.5
2.0
2.5
3.0
3.5