Automated Structure Solution with the PHENIX Suite

PETER H. ZWART,[A] PAVEL V. AFONINE,[A] RALF W. GROSSE-KUNSTLEVE,[A] LI-WEI HUNG,[B] THOMAS R. IOERGER,[C] AIRLIE J. MCCOY,[D] ERIK MCKEE,[C] NIGEL W. MORIARTY,[A] RANDY J. READ,[D] JAMES C. SACCHETTINI,[E] NICHOLAS K. SAUTER,[A] LAURENT C. STORONI,[D] THOMAS C. TERWILLIGER[F] AND PAUL D. ADAMS [A*]

*[a]Lawrence Berkeley National Laboratory, One Cyclotron Road, BLDG 64R0121, Berkeley, CA 94720, USA, [b]Biophysics Group, Mail Stop D454, Los Alamos National Laboratory, Los Alamos, NM 87545, USA, [c]Department of Computer Science, Texas A&M University, 301 H.R. Bright Building, 3112 TAMU, College Station, TX 77843, USA, [d]Department of Haematology, University of Cambridge, Cambridge Institute for Medical Research, Wellcome Trust/MRC Building, Hills Road, Cambridge, CB2 2XY, UK, [e]Department of Biochemistry and Biophysics, Texas A&M University, 103 Biochemistry/Biophysics Building, 2128 TAMU, College Station, TX 77843, USA, and [f]Los Alamos National Laboratory, Mailstop M888, Los Alamos, NM 87545, USA. *Corresponding Author; E-mail: PDAdams@lbl.gov; Fax: 510-486-5909, Phone: 510-486-4225.*

Running title: Automated structure solution with PHENIX

Abstract

Significant time and effort are often required to solve and complete a macromolecular crystal structure.The development of automated computational methods for the analysis, solution and completion of crystallographic structures has the potential to produce minimally biased models in a short time without the need for manual intervention. The PHENIX software suite is a highly

automated system for macromolecular structure determination that can rapidly arrive at an initial partial model of a structure without significant human intervention, given moderate resolution and good quality data. This achievement has been made possible by the development of new algorithms for structure determination, maximum-likelihood molecular replacement (PHASER), heavy-atom search (HySS), template and pattern-based automated model-building (RESOLVE, TEXTAL), automated macromolecular refinement (phenix.refine), and iterative model-building, density modification and refinement that can operate at moderate resolution (RESOLVE, AutoBuild). These algorithms are based on a highly integrated and comprehensive set of crystallographic libraries that have been built and made available to the community. The algorithms are tightly linked and made easily accessible to users through the PHENIX Wizards and the PHENIX GUI.

Key Words: Crystallography, Automated Structure Determination, PHENIX, Wizards, Strategies, Python, C++, Likelihood, Refinement.

1. Introduction

The world-wide efforts of many structural genomics projects, in particular the NIH Protein Structure Initiative have lead to many new technological advances in robotized cloning, sample expression and purification, screening of crystallization conditions ***(1)***, data collection at synchrotron sources ***(2)***, and structure solution ***(3)***. These have made high-throughput structure determination possible, and achievable for a number of the structures solved at structural genomics centers. More recently these technologies have started to be adopted by many structural biology research laboratories, where they are being applied to challenging systems such as large molecular complexes, and membrane proteins.

The demand for better software for crystallographic structure solution is increasing as it becomes possible for researchers to study more systems using high-throughput methods. This increased demand will need to be at least partially met by automated software for structure solution. Automation cannot rely on the availability of manual input from a trained crystallographer; the software must be able to make many complex decisions itself. Individual investigator research groups face a related problem as more biologists and biochemists make use of crystallography purely as a technique to better understand their biological system. Often there is insufficient time to obtain a very detailed expert crystallographic knowledge. A significant amount of this knowledge must therefore be built into the software. Furthermore, automated processes avoid possible subjective interpretation from manual interpretation of complex numerical data that can lead to delays or even inhibit reaching a high quality, final structure. Automated methods have the potential to produce minimally biased models in an efficient manner.

Current software packages such as SOLVE ***(4)***, SHARP ***(5)***, ACrS ***(6),*** SHELX-C/D/E ***(7,)*** CRANK ***(8),*** Elves ***(9)***, Auto-Rickshaw ***(10)*** and BnP ***(11)*** are capable of automatic structure solution using MAD, SAD, or other sources of experimental phases. Molecular replacement can be carried out in an automated fashion by software including Amore ***(12),*** PHASER ***(13)***, EPMR ***(14)***, and MOLREP ***(15)***. Model-building can be carried out automatically by several algorithms including those in ARP/wARP ***(16)***, RESOLVE ***(17, 18)***, TEXTAL ***(19)*** and MAID ***(20)***. Manual model building programs such as O ***(21)***, XtalView ***(22)***, COOT ***(23)*** and MAIN ***(24)*** have also incorporated an increasing number of tools that help automate complex tasks such as validation and model building. However, there still remain serious computational bottlenecks in structure determination. Truly automated structure solution is still limited to routine structures for which high quality experimental data are available, typically at 2.5 Å or better.

Current shortcomings of automated algorithms are unlikely to be overcome by simply combining current software packages into automated "pipelines". Rather, new algorithms must be developed and combined with new approaches to decision-making. The PHENIX software ***(25, 26)*** has been developed with the needs for automation and complex decision-making in mind.

2. Materials: The PHENIX Suite

PHENIX builds upon Python, the Boost.Python Library, and C++ to provide an environment for automation and scientific computing. Many of the fundamental crystallographic building blocks, such as data objects and tools for their manipulation are provided by the Computational Crystallography Toolbox (*cctbx*; ***(27)***). The computational tasks which perform complex crystallographic calculations are then built on top of this. Finally, there are a number of different user interfaces available in PHENIX. In order to facilitate automated operation there is the Project Data Storage (PDS) that is used to store and track the results of calculations.

2.1. User Interfaces

Different user interfaces are required depending on the needs of a diverse user community. There are currently three different user interfaces, each described below.

2.1.1. Command Line Interface

For a number of applications a command-line interface is most effective. This is particularly the case when rapid results are required, such as data quality assessment and twinning analysis, or substructure solution at the synchrotron beam line. Tools that facilitate the ease of use at the early stages of structure solution, such as data analyses (phenix.xtriage), substructure solution (phenix.hyss) and reflection file manipulations such as the generation of a test set, reindexing and merging of data (iotbx.reflection_file_converter) are available via simple command line

interfaces. Another major application that is controlled via the command line interface is phenix.refine.

To illustrate the command line interface, the command used to run the program that carries out a data quality and twinning analyses is:

```
phenix.xtriage my_data.sca [options]
```

Further options can be given on the command line, or can be specified via a parameter file:

```
phenix.xtriage my_parameters.def
```

A similar interface is used for macromolecular refinement:

```
phenix.refine my_model.pdb my_data.mtz
```

Although SCALEPACK and MTZ formats are indicated in the above example, reflection file formats such as D*TREK, CNS/XPLOR or SHELX can be used, as the format is detected automatically.

Help for all command line applications can be obtained by use of the flag `--help` :

```
phenix.refine --help
```

2.1.2. Tasks and Strategies

The PHENIX strategy interface provides a way to construct complex networks of tasks to perform a higher-level function (**Fig. 1**). For example, the steps required to go from initial data to a first electron density map in a SAD experiment can be broken down into well-defined tasks (available from the task window in the GUI) which can be reused in other procedures. Instead of requiring the user to run these tasks in the correct order they are connected together by the software developer, and can thus be run in an automated way. However, because the connection

between tasks is dynamic they can be reconfigured or modified, and new tasks introduced as necessary if problems occur. This provides the flexibility of user input and control, while still permitting complete automation when decision-making algorithms are incorporated into the environment. The tasks and their connection into strategies rely on the use of plain text task files written using the Python scripting language. This enables the computational algorithms to be used easily in a non-graphical environment. The PHENIX GUI permits strategies to be visualized and manipulated. These manipulations include loading a strategy distributed with PHENIX, customizing and saving it for future recall.

Current tasks and strategies available include:

- Density modification; Carries out a single run of RESOLVE.
- Substructure solution; Runs phenix.hyss ***(28)***.
- Molecular replacement; Computes rotation and translation functions with PHASER.
- Model building; Using TEXTAL or RESOLVE.

2.1.3. Wizards

The decision-making in strategies is local, with decisions being made at the end of each task to determine the next path in the network. Crystallographers typically make decisions in a very similar way during structure solution; a program is run, the outputs manually inspected and a decision made about the next step in the process. By contrast, a wizard provides a user interface that can make more global decisions, by considering all of the available information at each step in the process. Wizards are designed to lead the users through the process of setting up a desired task, making automatic decisions when possible, but prompting the user for additional information when necessary. The wizard interface uses the same graphical environment as the

strategies, but consists of only a single input/output area (**Fig. 2**).

Currently available wizards perform the following tasks:

- Structure solution using experimental phasing approaches such as SAD/MAD and SIR
- Structure solution via molecular replacement
- Automated model building, structure completion and refinement of structures
- Automated ligand building

2.2. Common Crystallographic Computations

The following paragraphs are a brief description of a number of common tasks that can be performed within the PHENIX framework.

2.2.1. Automated Structure Solution Using Experimental Phasing Techniques

Structure solution via SAD, MAD or SIR(AS) can be carried out with the AutoSol wizard. The AutoSol wizard performs heavy atom location, phasing, density modification and initial model building in an automated manner.

The heavy atoms are located with substructure solution engine also used in phenix.hyss ***(29***), a dual space method similar to SHELXD ***(7)*** and Shake and Bake ***(30)***. Phasing is carried out with PHASER for SAD cases and with SOLVE for MAD and SIR(AS) cases. Subsequent density modification is carried out with RESOLVE. The hand of the substructure is determined automatically on the basis of the quality of the resulting electron density map. It is noteworthy that the whole process is not necessarily linear but that the wizard can decide to step back and (for instance) try another set of heavy atoms if appropriate.

In the resulting electron density map, a model is built (currently limited to proteins). Further

model completion can be carried out via the AutoBuild wizard. The AutoBuild wizard iterates model building and density modification with refinement of the model in a scheme similar to other iterative model building methods, for example ARP/wARP ***(16)***.

2.2.2. Automated Structure Solution Via Molecular Replacement

Structure solution via molecular replacement is facilitated via the AutoMR wizard. The AutoMR wizard guides the user through setting up all necessary parameters to run a molecular replacement job with PHASER.

The molecular replacement carried out by PHASER uses likelihood based scoring function ***(13, 31)***, improving the sensitivity of the procedure and the ability to obtain reasonable solutions with search models that have a relatively low sequence similarity to the crystal structure being determined. Besides the use of likelihood based scoring functions, structure solution is enhanced by detailed bookkeeping of all search possibilities when searching for more then a single copy in the asymmetric unit or when there the choice of space group is ambiguous.

When a suitable molecular replacement solution is found, the AutoBuild wizard is invoked and rebuilds the molecular replacement model given the sequence of the model under investigation.

2.2.3. Automated Model Building

Automated model building given a starting model or a set of reasonable phases can be carried out by the AutoBuild wizard. A typical AutoBuild job combines density modification, model building, macromolecular refinement and solvent model updates ('water picking') in an iterative manner.

Various modes of building a model are available. Depending on the availability of a molecular model, model building can be carried by locally rebuilding an existing model (*rebuild in place*)

or by building in the density without any information of an available model. The rebuilding in place model building is a powerful building scheme that is used by default for molecular replacement models that have a high sequence similarity to the sequence of the structure that is to be built.

A fundamental feature of the AutoBuild wizard is that it builds various models, all from slightly different starting points. The dependency of the outcome of the model building algorithm on initial starting conditions provides a straightforward mechanism to obtain a variety of plausible molecular models. It is not uncommon that certain sections of a map are built in one model, while not in another. Combining these models allows the AutoBuild wizard to converge faster to a more complete model than when using a single model building pass for a given set of phases.

Dedicated loop fitting algorithms are used to close gaps between chain segments. This feature, together with the water picking and side chain placement, typically results in highly complete models of high quality that need minimal manual intervention before they are ready for deposition.

2.2.4. Refinement

The refinement engine used in the AutoBuild and AutoSol wizards can also be run from the command line with the phenix.refine command. The phenix.refine program carries out likelihood based refinement and has the possibility to refine positional parameters, individual or grouped atomic displacement parameters, individual or grouped occupancies. The refinement of anisotropic displacement parameters (individual or via a TLS parameterization ***(32, 33)***) is also available. Positional parameters can be optimized using either traditional gradient-only based optimization methods, or via simulated annealing protocols ***(34, 35)***.

The command line interface allows the user to specify which part of the model should be refined in what manner. It is in principle possible to refine half of the molecule as a rigid group with grouped B values, whereas the other half of the molecule has a TLS parameterization. The flexibility of specifying the level of parameterization of the model is especially important for the refinement of low resolution data or when starting with severely incomplete atomic models. Another advantage of this flexibility in refinement strategy is that a user can perform a complex refinement protocol that carries out simulated annealing, isotropic B refinement and water picking in 'one go'.

Another main feature of phenix.refine is the way in which the relative weights for the geometric and ADP restraints with respect to the X-ray target are determined. Considerable effort has been put into devising a good set of defaults and weight determination schemes that results in a good choice of parameters for the data set under investigation. Defaults can of course be overwritten if the user chooses to.

Besides being able to handle refinement against X-ray data, phenix.refine can refine against neutron data or against X-ray and neutron data simultaneously.

2.2.5. Ligands

Automated fitting of ligands into the electron density is facilitated via the LigandFit wizard. The ligand building is performed by finding an initial fit for the largest rigid domain of the ligand and extending the remaining part of the ligand from this initial 'seed'. Besides being able to fit a known ligand into a difference map, the LigandFit wizard is capable to identify ligands on the basis of the difference density only. In the latter scheme, density characteristics for ligands occurring frequently in the PDB ***(36, 37)*** are used to provide the user with a range of plausible ligands.

Stereo chemical dictionaries of ligands whose chemical description is not available in the supplied monomer library ***(38)*** for the use in restrained macromolecular refinement can be generated with the *electronic ligand builder and optimization workbench* (eLBOW). eLBOW generates a 3D geometry from a number of chemical input formats including MOL2 or PDB files and SMILES strings ***(39)***. SMILES is a compact, chemically dense description of a molecule that contains all element and bonding information and optionally other stereo information such as chirality. To generate a 3D geometry from an input format that contains no 3D geometry information, eLBOW uses a Z-Matrix formalism in conjunction with a table of bond lengths calculated using the Hartree-Fock method with a 6-31G(d,p) basis set to obtain a Cartesian coordinate set. The geometry is then optionally optimized using the semi-empirical quantum chemistry method AM1. The AM1 optimization provides chemically meaningful and accurate geometries for the class of molecule typically complexed with proteins. eLBOW outputs the optimized geometry and a standard CIF restraint file that can be read in by phenix.refine and can also be used for real space refinement during manual model building sessions in the program *COOT* ***(23)***. An interface is also available to use eLBOW within COOT.

2.2.6. Twinned Data

The presence of twinning can severely delay structure solution, model completion and refinement if not explicitly taken into account. Detection of twinning on the basis of intensity statistics only is facilitated via the program phenix.xtriage. This command line driven program analyses an experimental data set and provides diagnostics that aid in the detection of other common idiosyncrasies such as the presence of pseudo translational symmetry or certain data processing problems. Other sanity checks, such as a Wilson plot sanity check ***(40)*** and an algorithm that tries to detect the presence of ice rings from the merged data are performed as

well.

If twin laws are present for the given unit cell and space group, a Britton plot ***(41)*** is computed, an H-test ***(42, 43)*** is performed and a likelihood based method is used to provide an estimate of the twin fraction. Twin laws are deduced from first principles for each data set, avoiding the danger of over-looking twin laws by incomplete lookup tables. If a model is available, more efficient twin detection tools are available. The RvsR statistic ***(44)*** is particularly useful in the detection of twinning in combination with pseudo rotational symmetry. This statistic is computed by phenix.xtriage if calculated data is supplied together with the observed data. A more direct test for the presence of twinning is by refinement of the twin fraction given an atomic model. The command line utility phenix.twin_map_utils provides a straightforward way to refine a twin fraction given an atomic model and an X-ray data set and also produces 'detwinned' $2F_o$-F_c and gradient maps. The implementation of least-squares targets for refinement of twinned data will be available in phenix.refine in the near future.

The routines in Xtriage can also detect the presence of higher intensity symmetry than specified by the space group of the data. If higher intensity symmetry is detected, the user is advised to consider reprocessing the data.

3. Methods: Worked Examples

A few examples are given here to highlight many of the points mentioned in the previous sections. The results shown here have been generated with PHENIX version 1.26b-d2 (December 2006).

3.1. Structure solution via S-SAD Phasing

An X-ray data set of insulin measured at a wavelength of 1.54 A, was input to the AutoSol

wizard for substructure solution and phasing (*see* **Note 1**). Seven anomalous sites are found and refined with PHASER. Phasing is carried out for both choices of the hand. The quality of the electron density is used to determine the correct hand. The solution that produces the best map is used for further density modification and model building.

Initial phasing of the sites produces a map with a mean figure of merit equal to 0.38. The experimental phases are of such a quality that large aromatic side chains such as tyrosine can be recognized in the map even before any density modification is applied (**Fig. 3**). The AutoSol wizard generated an almost complete model lacking only five N-terminal residues in weak density. Subsequent manual building of these residues in COOT and automated placement of waters and additional refinement by phenix.refine resulted in R-values (work/free) of 18%/20%.

3.2. Molecular Replacement

To illustrate a typical structure solution with molecular replacement, the structure of epsin ***(45)*** was solved using the structure of 1XGW as a search template. The sequence identity of the search template is 54% over the length of the alignment. The X-ray data extend to 1.84 Å. A Matthews analysis suggested 1 molecule per asymmetric unit with an approximate solvent content of 44%.

The rotation function reveals two significant peaks with Z-scores of 5.7 and 6.2. The subsequent translational search results in two solutions with Z-scores of 10 and 13. After rigid-body refinement a single, unique, solution is produced. This solution can be used as a starting point in automated model building. Maps with virtually no model bias can be obtained (*see* **Note 2**) in a subsequent run of the AutoBuild wizard.

The space group of this particular data set is $P3_121$. The AutoMR wizard can be instructed to

search in space groups with the same point groups as well. In this particular case, the translation function only gives a satisfactory result in space group $P3_121$. Possible solutions in $P3_221$ and in P321 have low Z-scores and multiple clashes with symmetry related copies.

The molecular replacement shown in this example is relatively straightforward. Molecular replacement attempts with low similarity search templates and multiple copies in the unit cell can be particularly challenging (*see* **Note 3**), but are often successful with the likelihood algorithms implemented in PHASER.

3.3. Ligand Building

The ligand building capabilities of the LigandFit wizard are illustrated by fitting NADH and cholic acid into the structure of SS_LADH ***(46)***. The X-ray data extends to 1.54 Å, and the difference density is rather clear.

Atomic models for NADH and cholic acid were constructed from smiles strings using elbow.builder. The command used to obtain the cholic acid model is

```
elbow.builder --smiles="CC(CCC(O)=O)C1CCC2C3C(O)CC4CC(O)CCC4(C)C3CC(O)C12C" --opt
```

SMILES strings can be constructed with molecular editors such as JME (http://www.molinspiration.com/jme/) or can be obtained directly from MSDChem ***(47)***.

The automated ligand building procedure uses a protein model (without ligands) and the X-ray data to compute a difference map in which the ligand is built. Two copies of NADH and two copies of cholic acid were built. The quality of the model is shown in **Fig. 4**.

3.4. Refinement

A typical refinement with phenix.refine is initiated with the following command:

```
phenix.refine my_data.mtz my_model.pdb
```

The refinement program will try to determine which columns in the MTZ file to use for refinement and which column contains the test flags for cross validation purposes.

An example of the application of TLS refinement and its effects on the R-values is illustrated by refinement of the synaptotagmin structure ***(48)***.The available X-ray data extended to 3.2 Å and is over 97% complete. Standard refinement (positional parameters and individual atomic displacement parameters (ADPs)) results in R-values of 24.6% and 27.7% for the work and test set, respectively. At this resolution, ADPs are often refined in groups by applying constraints to the ADP values for all atoms within a residue. The refinement of ADPs in this manner resulted in R-values of 24.7% and 28.9% for the work and test set, respectively. The application of a TLS model to the atomic displacement parameters that models the displacement of rigid groups within a crystal reduced the R-values to 22.7% and 25.9% for the work and test set respectively. An ORTEP diagram of the anisotropic ADPs is shown in **Fig. 5**.

If only a TLS parameterization is used to model the ADPs (*see* **Note 4**), local variations in ADPs due to increased or decrease flexibility cannot be taken into account. A more complete ADP model includes both a TLS and individual ADP parameterization. The refinement of both the TLS parameters and the individual ADPs result in R-values of 20.7% and 24.4% for the work and test set respectively.

The command that is needed to perform this last refinement is straightforward:

```
phenix.refine scale.hkl synaptotagmin.pdb tls.param
```

Note that besides the experimental data and the atomic model and extra parameter file is specified. This parameter file has the following content:

```
refinement.refine {
  strategy = *individual_sites *individual_adp *tls
```

```
  adp {
    tls = "(chain A and resid :421)"
    tls = "(chain A and resid 422:430)"
    tls = "(chain A and resid 431:)"
  }
}
```

The line containing the keyword *strategy* specifies that the positional parameters for individual sites should be refined, as well as a TLS model and individual ADPs. The TLS groups are defined by the *adp* scope. In this case, 3 TLS domains are specified within chain A.

Although most parameters for the refinement are set automatically, defaults (such as weights) can be set manually if desired (*see* **Notes 5, 6**)

3.5. Twinning

The deposited X-ray dataset of PDB ID 1GH7 was analyzed by phenix.xtriage for the presence of twinning. A single twin law was found (-h-k, k ,-l). Analyses of the intensity statistics indicates that the data is twinned:

Statistics independent of twin laws

```
  - <I^2>/<I>^2 : 1.795
  - <F>^2/<F^2> : 0.843
  - <|E^2-1|>   : 0.658
  - <|L|>, <L^2>: 0.396, 0.219
       Multivariate Z score L-test: 8.104
       The multivariate Z score is a quality measure of the given
       spread in intensities. Good to reasonable data is expected
       to have a Z score lower than 3.5.
       Large values can indicate twinning, but small values do not
       necessarily exclude it.

The results of the L-test indicate that the intensity statistics are
significantly different then is expected from good to reasonable, untwinned
data.
As there are twin laws possible given the crystal symmetry, twinning could be
the reason for the departure of the intensity statistics from normality.
It might be worthwhile carrying out refinement with a twin specific target
function.
```

An H-test (42, 43) and Britton analyses (41) indicate a twin fraction of approximately 7%.

Refinement of the twin fraction and bulk solvent and scaling parameters reveals that the data is 16% twinned, a fact overseen during the original structure solution ***(49)***.

4. Notes

1. The effect of the data quality on the ability to solve the substructure

The quality of the anomalous signal has a large impact on the ability to solve the substructure. The AutoSol wizard analyses the anomalous signal in a dataset by computing either a correlation coefficient between the anomalous differences or a signal to noise ratio for SAD data. On the basis of these statistics, resolution limits for substructure solution are chosen.

The quality of the anomalous data can be checked manually with iotbx.reflection_statistics. It computes correlation coefficients between anomalous differences and a statistic known as the measurability ***(50)*** for SAD data sets. Correlation coefficients larger then 30% indicate significant anomalous signal in a MAD data set. For SAD datasets, measurabilities larger than 6% indicate the presence of significant anomalous signal.

Although the AutoSol wizard analyses the signal to noise level of the anomalous data and makes appropriate resolution cut offs, it can be worthwhile running phenix.hyss with various resolution cutoffs if the AutoSol wizard fails to find the substructure with weak anomalous diffraction data.

2. Bias removal in molecular replacement maps

The presence of bias in molecular replacement phases can make the interpretation of the electron density difficult or misleading. This bias can be removed by computing a *Full Omit* map in the AutoBuild wizard. The Full Omit procedure is reminiscent of the composite omit maps of CNS ***(51)*** but provides a means to remove nearly all the bias, at the cost of computing time.

3. Difficult molecular replacement problems

Not all structures can be solved by molecular replacement. Certain strategies can however be adopted to push its capabilities to the boundaries of what is possible. Careful editing of the input model by removing non conserved, flexible loops can make a big difference. Breaking a flexible model down into multiple rigid domains that can be used in a multi-copy search can be a vital ingredient for a successful structure solution. Other suggestions are available from the program documentation.

4. Interpreting the result of a TLS refinement

By default, the ADPs written by phenix.refine are the total ADPs rather than residual ADPs, (which can be negative). This convention allows for easy viewing of the results of structure refinement in molecular graphics programs.

5. Definition of NCS restraints in refinement

In order to increase the data to parameter ratio during refinement, multiple copies of the protein within the asymmetric unit can be restrained to have a similar conformation. These NCS restraints can be set up automatically by phenix.refine, or defined manually by the user.

6. Weight optimization in restrained macromolecular refinement

The weight that determines the relative contribution of the X-ray target with respect to the restraint terms is determined automatically. The procedure works well in most cases, but a manual optimization of this weight can be used if necessary. Changing the weight manually can be performed via the following command:

```
phenix.refine my_data.mtz my_model.pdb wxc_scale=5
```

Rerunning the refinement job with various values for the weight *wxc_scale* and a careful monitoring of the free R-value will give an indication of a suitable value for the weight. The

same manual manipulation of the weighting for isotropic ADP restraints can be achieved with the *wxu_scale* parameter.

5. Acknowledgements

PHENIX can be downloaded from http://www.phenix.online.org/, and is freely available to non-profit researchers. The open source crystallographic library (the CCTBX) is available from http://cctbx.sourceforge.net/.

We gratefully acknowledge the financial support of NIH/NIGMS through grants 5P01GM063210, 5P50GM062412, 5R01GM071939, and the PHENIX industrial consortium. This work was supported in part by the US Department of Energy under Contract No. DE-AC02-05CH11231.

6. References

1. Page, R., Grzechnik, S. K., Canaves, J. M., Spraggon, G., Kreusch, A., Kuhn, P., Stevens, R. C., and Lesley, S. A. (2003) Shotgun crystallization strategy for structural genomics: an optimized two-tiered crystallization screen against the Thermotoga maritima proteome. *Acta Crystallogr. D Biol. Crystallogr.* **59,** 1028-1037.
2. Snell, G., Cork, C., Nordmeyer, R., Cornell, E., Meigs, G., Yegian, D., Jaklevic, J., Jin, J., Stevens, R. C., and Earnest, T. (2004) Automated Sample Mounting and Alignment System for Biological Crystallography at a Synchrotron Source. *Structure* **12,** 537-545.
3. Adams, P. D., Grosse-Kunstleve, R.W., and Brunger, A.T. (2003) Computational aspects of high-throughput crystallographic macromolecular structure determination. *Methods Biochem. Anal.* **44,** 75-87.

4. Terwilliger, T. C., and Berendzen, J. (1999) Automated MAD and MIR structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **55,** 849-861.

5. de la Fortelle, E., Bricogne, G. (1997) Maximum-likelihood heavy atom parameter refinement in the MIR and MAD methods. *Methods Enzymol* **276**, 472-494.

6. Brunzelle, J. S., Shafaee, P., Yang, X., Weigand, S., Ren, Z., and Anderson, W. F. (2003) Automated crystallographic system for high-throughput protein structure determination. *Acta Crystallogr. D Biol. Crystallogr.* **59,** 1138-1144.

7. Schneider, T. R., and Sheldrick, G. M. (2002) Substructure solution with SHELXD. *Acta Crystallogr. D Biol. Crystallogr.* **58,** 1772-1779.

8. Ness, S. R., de Graaff, R. A., Abrahams, J. P., and Pannu, N. S. (2004) CRANK: new methods for automated macromolecular crystal structure solution. *Structure* **12,** 1753-1761.

9. Holton, J., and Alber, T. (2004) Automated protein crystal structure determination using ELVES. *Proc. Natl. Acad. Sci. U S A* **101,** 1537-1542.

10. Panjikar, S., Parthasarathy, V., Lamzin, V. S., Weiss, M. S., and Tucker, P. A. (2005) Auto-Rickshaw: an automated crystal structure determination platform as an efficient tool for the validation of an X-ray diffraction experiment. *Acta Crystallogr. D Biol. Crystallogr.* **61,** 449-457.

11. Weeks, C., Blessing, R., Miller, R., Mungee, R., Potter, S., Rappleye, J., Smith, G., Xu, H. and Furey, W. (2002). Towards automated protein structure determination: BnP, the SnB-PHASES interface. *Zeitschrift Kristallographie* **217**, 686-693, 2002.

12. Navaza, J. (1994) AMoRe: an automated package for molecular replacement doi:10.1107/S0108767393007597. *Acta Crystallogr. A* **50,** 157-163.

13. McCoy, A. J., Grosse-Kunstleve, R. W., Storoni, L. C., and Read, R. J. (2005) Likelihood-enhanced fast translation functions. *Acta Crystallogr. D Biol. Crystallogr.*

61, 458-464.

14. Kissinger, C. R., Gehlhaar, D. K., and Fogel, D. B. (1999) Rapid automated molecular replacement by evolutionary search. *Acta Crystallogr. D Biol. Crystallogr.* **55,** 484-491.

15. Vagin, A., and Teplyakov, A. (2000) An approach to multi-copy search in molecular replacement. *Acta Crystallogr. D Biol. Crystallogr.* **56,** 1622-1624.

16. Perrakis, A., Morris, R., and Lamzin, V. S. (1999) Automated protein model building combined with iterative structure refinement. *Nat. Struct. Biol.* **6,** 458-463.

17. Terwilliger, T. (2003) Automated side-chain model building and sequence assignment by template matching. *Acta Crystallogr. D Biol. Crystallogr.* **59,** 45-49.

18. Terwilliger, T. (2003) Automated main-chain model building by template matching and iterative fragment extension. *Acta Crystallogr. D Biol. Crystallogr.* **59,** 38-44.

19. Holton, T., Ioerger, T. R., Christopher, J. A., and Sacchettini, J. C. (2000) Determining protein structure from electron-density maps using pattern matching. *Acta Crystallogr. D Biol. Crystallogr.* **56,** 722-734.

20. Levitt, D. G. (2001) A new software routine that automates the fitting of protein X-ray crystallographic electron-density maps. *Acta Crystallogr. D Biol. Crystallogr.* **57,** 1013-1019.

21. Jones, T. A., Zou, J.-Y., Cowan, S. W., and Kjeldgaard, M. (1991) Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47,** 110-119.

22. McRee, D. E. (1999) XtalView/Xfit--A versatile program for manipulating atomic coordinates and electron density. *J. Struct. Biol.* **125,** 156-165.

23. Emsley, P., and Cowtan, K. (2004) Coot: model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.* **60,** 2126-2132.

24. Main, D. (1992), Technische Universitaet Muenchen, Muenchen.

25. 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., and Terwilliger, T. C. (2002) PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallogr. D Biol. Crystallogr.* **58,** 1948-1954.

26. Adams, P. D., Gopal, K., Grosse-Kunstleve, R. W., Hung, L.-W., Ioerger, T. R., McCoy, A. J., Moriarty, N. W., Pai, R. K., Read, R. J., Romo, T. D., Sacchettini, J. C., Sauter, N. K., Storoni, L. C., and Terwilliger, T. C. (2004) Recent developments in the PHENIX software for automated crystallographic structure determination. *J. Synchr. Radiat.* **11,** 53-55.

27. Grosse-Kunstleve, R. W., Sauter, N. K., Moriarty, N. W., and Adams, P. D. (2002) The Computational Crystallography Toolbox: crystallographic algorithms in a reusable software framework. *J. Appl. Crystallogr.* **35,** 126-136.

28. Grosse-Kunstleve, R. W., and Adams, P. D. (2003) Substructure search procedures for macromolecular structures. *Acta Crystallogr. D Biol. Crystallogr.* **59,** 1966-1973.

29. Grosse-Kunstleve, R. W., and Brunger, A. T. (1999) A highly automated heavy-atom search procedure for macromolecular structures. *Acta Crystallogr. D Biol. Crystallogr.* **55,** 1568-1577.

30. Weeks, C. M., and Miller, R. (1999) Optimizing Shake-and-Bake for proteins. *Acta Crystallogr. D Biol. Crystallogr.* **55,** 492-500.

31. Read, R. (2001) Pushing the boundaries of molecular replacement with maximum likelihood. *Acta Crystallogr. D Biol. Crystallogr.* **57,** 1373-1382.

32. Schomaker, V., and Trueblood, K. N. (1968) On the rigid-body motion of molecules in crystals. *Acta Crystallogr. B* **24,** 63-76.

33. Winn, M. D., Isupov, M. N., and Murshudov, G. N. (2001) Use of TLS parameters to

model anisotropic displacements in macromolecular refinement. *Acta Crystallogr. D Biol. Crystallogr.* **57,** 122-133.

34. Brunger, A. T., Adams, P. D., and Rice, L. M. (1999) Annealing in crystallography: a powerful optimization tool. *Prog. Biophys. Mol. Biol.* **72,** 135-155.

35. Rice, L. M., and Brunger, A. T. (1994) Torsion angle dynamics: reduced variable conformational sampling enhances crystallographic structure refinement. *Proteins* **19,** 277-290.

36. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) The Protein Data Bank. *Nucleic Acids Res.* **28,** 235-242.

37. Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Jr., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., and Tasumi, M. (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112,** 535-542.

38. Vagin, A. A., Steiner, R. A., Lebedev, A. A., Potterton, L., McNicholas, S., Long, F., and Murshudov, G. N. (2004) REFMAC5 dictionary: organization of prior chemical knowledge and guidelines for its use. *Acta Crystallogr. D Biol. Crystallogr.* **60,** 2184-2195.

39. Weininger, D. (1988) SMILES 1. Introduction and Endoding Rules. *J. Chem. Inf. Comput. Sci.* **28,** 31.

40. Morris, R. J., Zwart, P. H., Cohen, S., Fernandez, F. J., Kakaris, M., Kirillova, O., Vonrhein, C., Perrakis, A., and Lamzin, V. S. (2004) Breaking good resolutions with ARP/wARP. *J. Synchr. Radiat.* **11,** 56-59.

41. Fisher, R. G., and Sweet, R. M. (1980) Treatment of diffraction data from crystals twinned by merohedry. *Acta Crystallogr. A* **36,** 755-760.

42. Yeates, T. O. (1988) Simple statistics for intensity data from twinned specimens. *Acta Crystallogr A* **44 (Pt 2),** 142-144.

43. Yeates, T. O. (1997) Detecting and overcoming crystal twinning. *Methods Enzymol* **276,** 344-358.

44. Lebedev, A. A., Vagin, A. A., and Murshudov, G. N. (2006) Intensity statistics in twinned crystals with examples from the PDB. *Acta Crystallogr. D Biol. Crystallogr.* **62,** 83-95.

45. Hyman, J., Chen, H., Di Fiore, P. P., De Camilli, P., and Brunger, A. T. (2000) Epsin 1 undergoes nucleocytosolic shuttling and its eps15 interactor NH(2)-terminal homology (ENTH) domain, structurally similar to Armadillo and HEAT repeats, interacts with the transcription factor promyelocytic leukemia Zn(2)+ finger protein (PLZF). *J. Cell Biol.* **149,** 537-546.

46. Adolph, H. W., Zwart, P., Meijers, R., Hubatsch, I., Kiefer, M., Lamzin, V., and Cedergren-Zeppezauer, E. (2000) Structural basis for substrate specificity differences of horse liver alcohol dehydrogenase isozymes. *Biochemistry* **39,** 12885-12897.

47. Golovin, A., Oldfield, T. J., Tate, J. G., Velankar, S., Barton, G. J., Boutselakis, H., Dimitropoulos, D., Fillon, J., Hussain, A., Ionides, J. M., John, M., Keller, P. A., Krissinel, E., McNeil, P., Naim, A., Newman, R., Pajon, A., Pineda, J., Rachedi, A., Copeland, J., Sitnov, A., Sobhany, S., Suarez-Uruena, A., Swaminathan, G. J., Tagari, M., Tromm, S., Vranken, W., and Henrick, K. (2004) E-MSD: an integrated data resource for bioinformatics. *Nucleic Acids Res.* **32,** D211-216.

48. Sutton, R. B., Ernst, J. A., and Brunger, A. T. (1999) Crystal structure of the cytosolic C2A-C2B domains of synaptotagmin III. Implications for Ca(+2)-independent snare complex interaction. *J. Cell Biol.* **147,** 589-598.

49. Carr, P. D., Gustin, S. E., Church, A. P., Murphy, J. M., Ford, S. C., Mann, D. A., Woltring, D. M., Walker, I., Ollis, D. L., and Young, I. G. (2001) Structure of the complete extracellular domain of the common beta subunit of the human GM-CSF, IL-3, and IL-5 receptors reveals a novel dimer configuration. *Cell* **104,** 291-300.

50. Zwart, P. (2005) Anomalous signal indicators in protein crystallography. *Acta Crystallogr. D Biol. Crystallogr.* **61,** 1437-1448.

51. Brunger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T., and Warren, G. L. (1998) Crystallography & NMR System: A New Software Suite for Macromolecular Structure Determination. *Acta Crystallogr. D Biol. Crystallogr.* **54,** 905-921.

52. Potterton, L., McNicholas, S., Krissinel, E., Gruber, J., Cowtan, K., Emsley, P., Murshudov, G. N., Cohen, S., Perrakis, A., and Noble, M. (2004) Developments in the CCP4 molecular-graphics project. *Acta Crystallogr. D Biol. Crystallogr.* **60,** 2288-2294.

53. Merritt, E. A. (1999) Expanding the model: anisotropic displacement parameters in protein structure refinement. *Acta Crystallogr. D Biol. Crystallogr.* **55,** 1109-1117.

Figure Legends

Fig. 1. Example of the PHENIX strategy interface, showing a substructure and phasing strategy for MAD data. The tasks are connected by lines, which indicate the flow of program execution dependant upon the outcome of each task (in this case the two possible outcomes are OK or Fail, the latter being implicit).

Fig. 2. Example of the PHENIX wizard interface. Shown is the AutoBuild cycle invoked after a potential molecular replacement solution has been found.

Fig. 3. (**A**) Experimental S-SAD phases generate by PHASER during the execution of the AutoSol wizard for cubic insulin, before any density modification. The density is of a quality that it can be readily interpreted. The figure was prepared with CCP4MG ***(52)***. (**B**) The electron density map corresponding to the refined cubic insulin model after manual model completion and refinement with phenix.refine. The figure was prepared with CCP4MG ***(52)***.

Fig. 4. The difference density and the model of NADH build in an automated manner by the LigandFit Wizard. The figure was prepared with CCP4MG ***(52)***.

Fig. 5. An ORTEP-style diagram with anisotropic displacement parameters shown as three-dimensional ellipsoids, color coded by the magnitude of the total displacement for each atom

(darker grey indicates a higher total ADP for an atom). The lower domain of the protein clearly shows significant rigid-body displacements, which are well modeled by the TLS formulism with a small number of parameters. This figure was made using RASTEP ***(53)***.

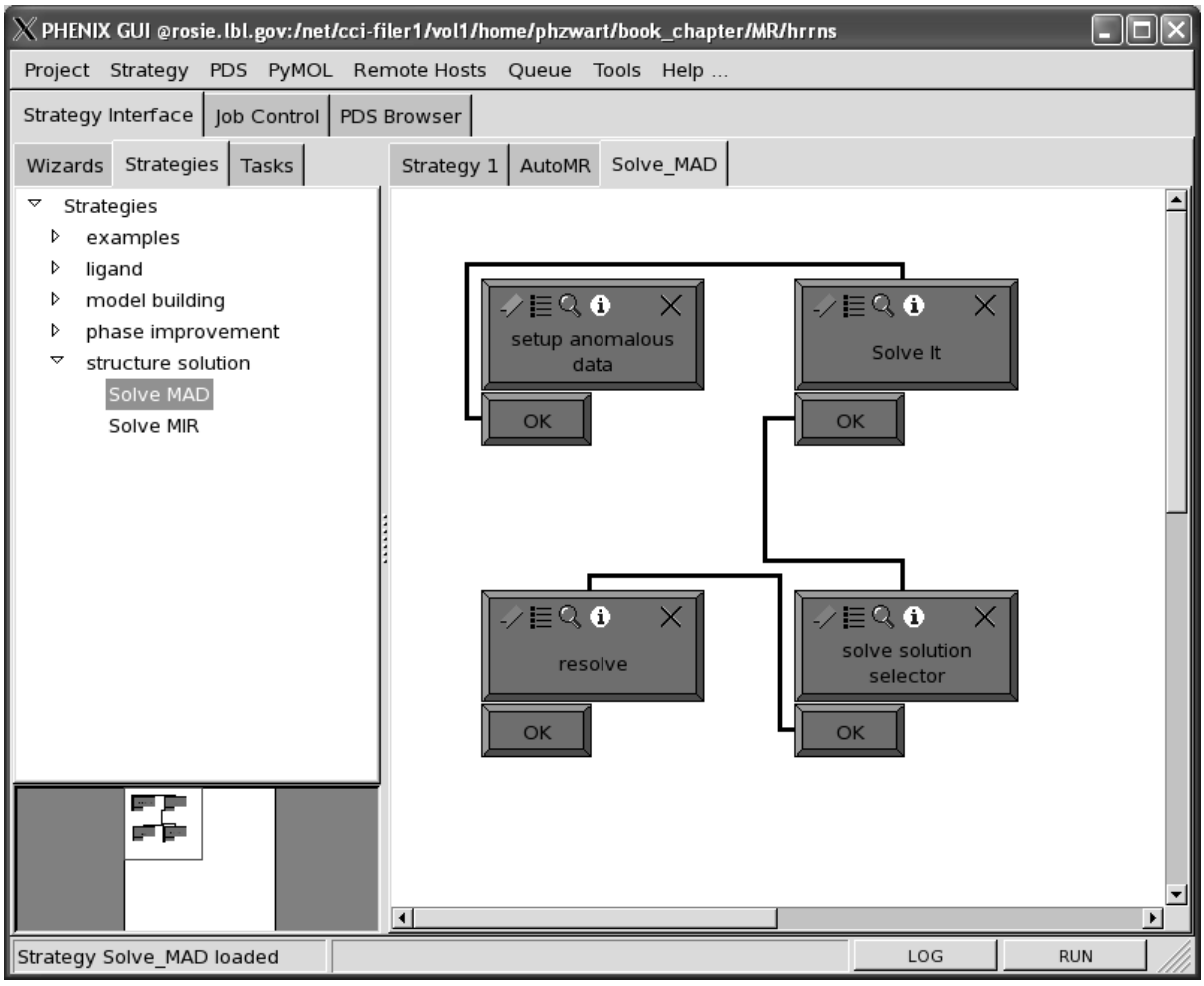
PHENIX GUI @rosie.lbl.gov:/net/cci-filer1/vol1/home/phzwart/book_chapter/MR/hrrns
Project
Strategy
PDS
PyMOL
Remote Hosts
Queue
Tools
Help ...
Strategy Interface
Job Control
PDS Browser
Wizards
Strategies
Tasks
Strategy 1
AutoMR
Solve_MAD
Strategies
examples
ligand
model building
phase improvement
structure solution
Solve MAD
Solve MIR
setup anomalous data
OK
Solve It
OK
resolve
OK
solve solution selector
OK
Strategy Solve_MAD loaded
LOG
RUN

Figure 1

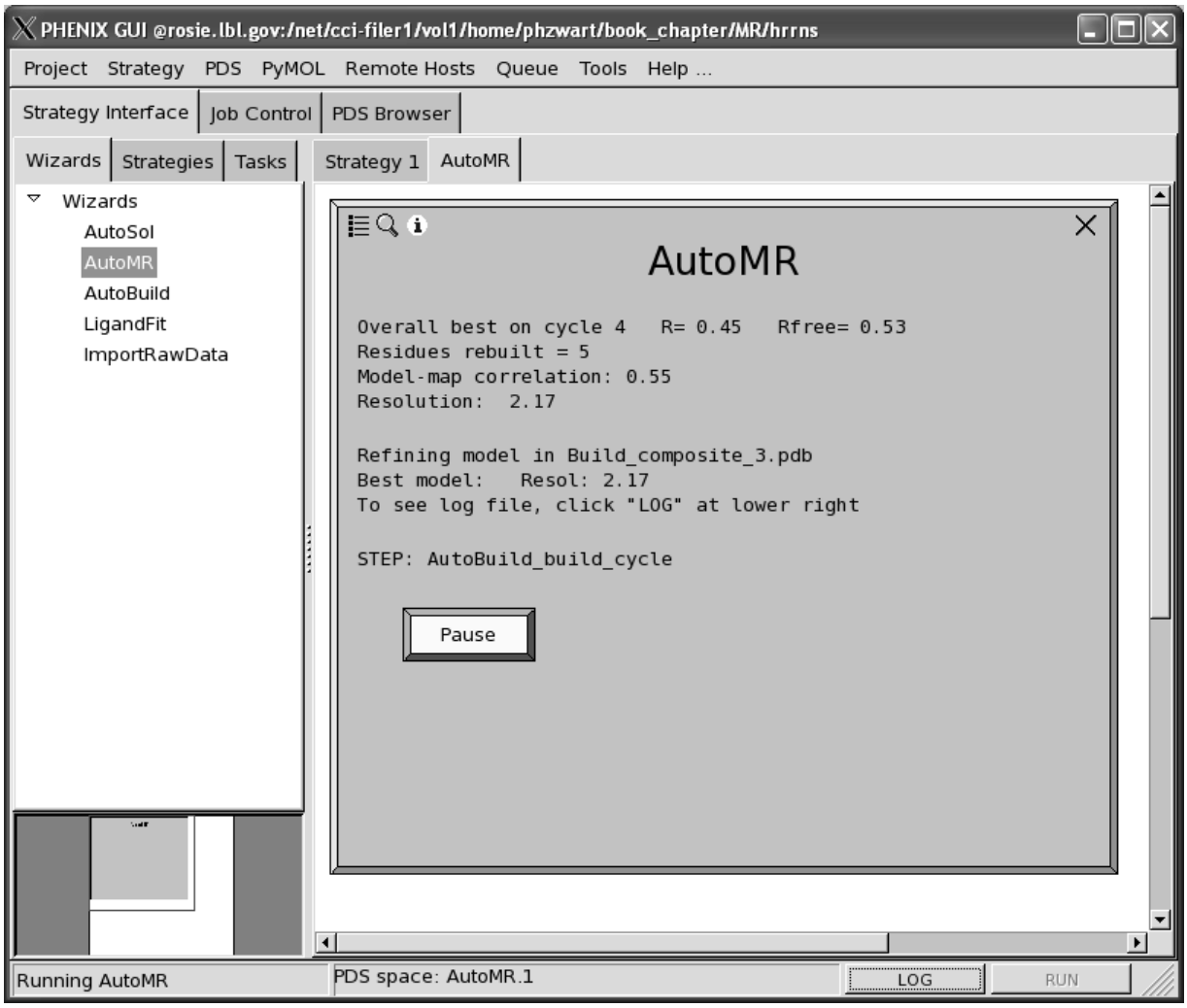
PHENIX GUI @rosie.lbl.gov:/net/cci-filer1/vol1/home/phzwart/book_chapter/MR/hrrns
Project Strategy PDS PyMOL Remote Hosts Queue Tools Help ...
Strategy Interface Job Control PDS Browser
Wizards Strategies Tasks
Strategy 1 AutoMR
Wizards
AutoSol
AutoMR
AutoBuild
LigandFit
ImportRawData
AutoMR
Overall best on cycle 4 R= 0.45 Rfree= 0.53
Residues rebuilt = 5
Model-map correlation: 0.55
Resolution: 2.17
Refining model in Build_composite_3.pdb
Best model: Resol: 2.17
To see log file, click "LOG" at lower right
STEP: AutoBuild_build_cycle
Pause
Running AutoMR
PDS space: AutoMR.1
LOG
RUN

Figure 2

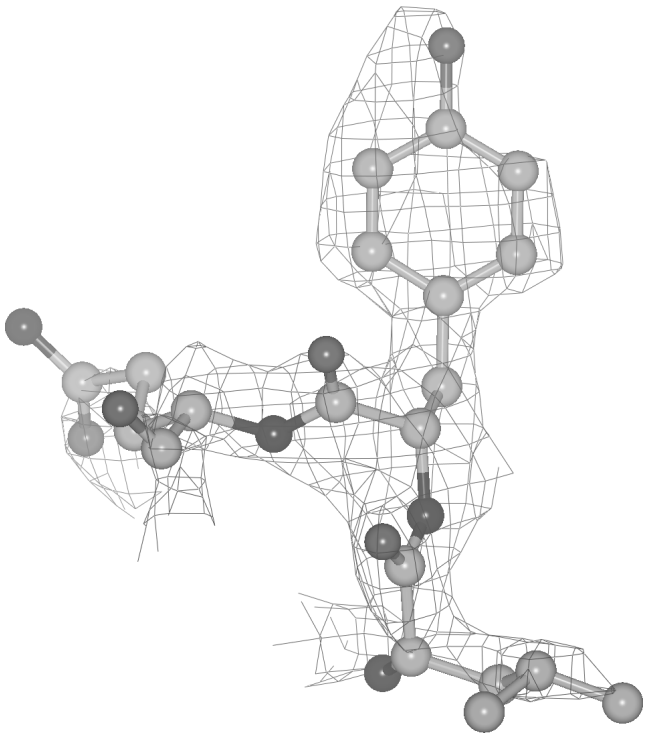

Figure 3A

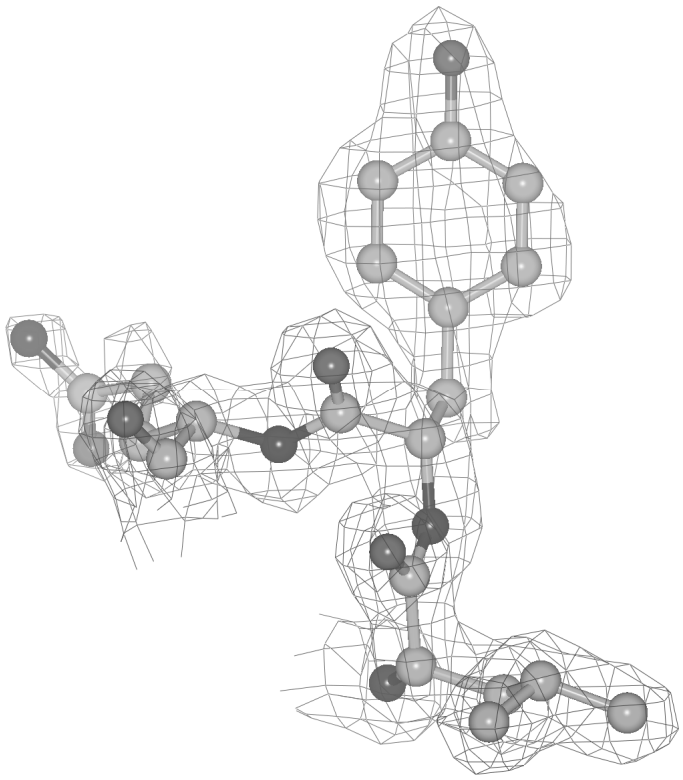

Figure 3B

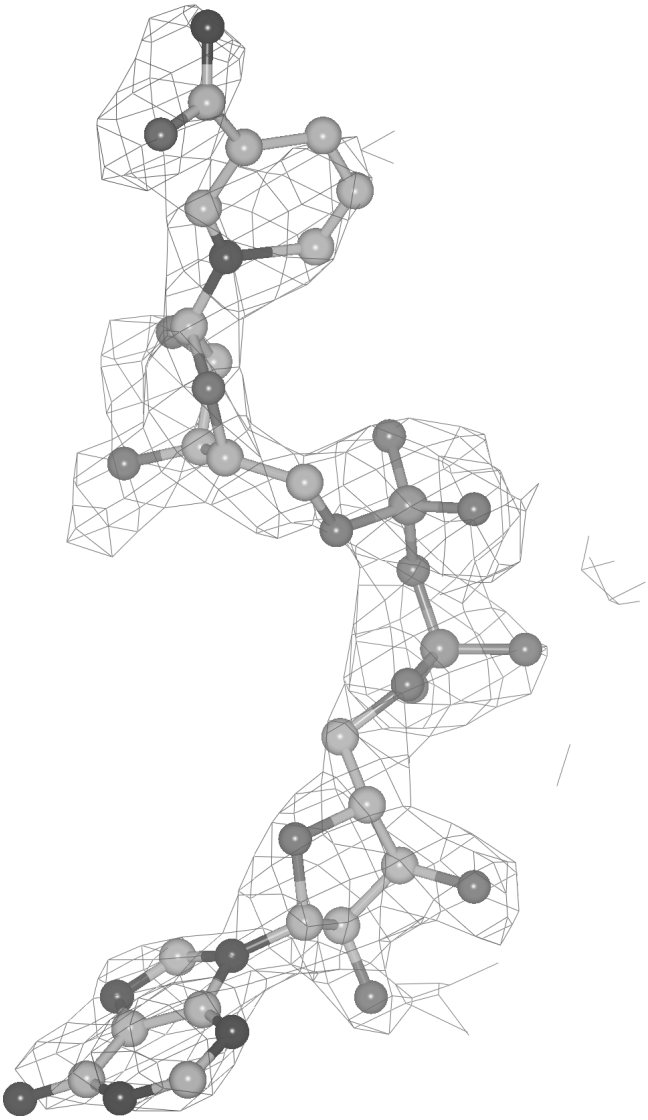

Figure 4

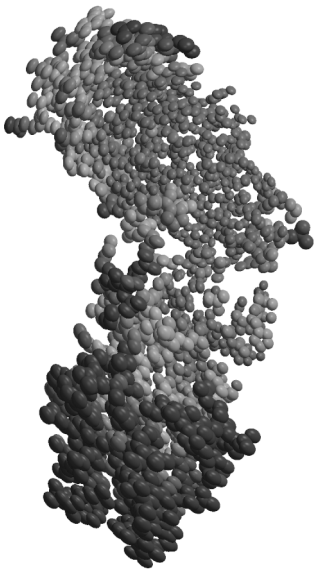

Figure 5