

Preface

Dependencies and requirements. The HKL-3000 program package combines all of the functions of HKL-2000 with a number of other macromolecular crystallographic computer programs. Please note that for the third-party programs, you must obtain the software and licenses (if needed) separately from their respective authors. The programs used are the CCP4 suite, the ARP/wARP suite, the SHELX-97 suite, the SOLVE/RESOLVE suite, Coot, PyMOL and Buccaneer. If a specific version of a third-party software is required, it is noted during the HKL-3000 installation process.

Scope. This manual describes the structure solution features of HKL-3000. The data collection and reduction features of HKL-3000 that were inherited from HKL-2000 are not described in this text. Please refer to the HKL-2000 tutorial for more information on these features.

Conventions of this manual. The term “page” refers to the area displayed whenever a tab along the top of the main window is pressed. If a page itself contains tabs, the contents of those areas are referred to as “subpages.” The names of pages and subpages are always shown in **boldface** (e.g. the **Project** page). References to specific elements (buttons, checkboxes, etc.) of the interface (e.g. click the “Refine” button) are shown in double quotes. The names of files and directories are shown in `fixed width`.

HKL-3000, HKL-2000, and the specific third-party programs used are described by the references listed below.

HKL-3000: Minor, W., Cymborowski, M., Otwinowski, Z. & Chruszcz, M. (2006). Acta Cryst. D62, 859-866.

HKL-2000: Otwinowski, Z. & Minor, W (1997). Methods Enzymol. 276, 307-326.

ARP/wARP: Perrakis, A., Morris, R. & Lamzin, V. S. (1999). Nature Struct. Biol. 6, 458-463.

BUCCANEER: Cowtan, K. (2006). Acta Cryst. D62, 1002-1011. Cowtan, K. (2008). Acta Cryst. D64, 83-89.

CCP4: Collaborative Computational Project, Number 4 (1994). Acta Cryst. D50, 760-763.

COOT: Emsley, P. & Cowtan, K. (2004). Acta Cryst. D60, 2126-2132.

DM: Cowtan, K. (2001). Acta Cryst. D57, 1435-1444. Cowtan, K. & Main, P. (1998). Acta Cryst. D54, 487-493. Cowtan, K. D. & Zhang, K. Y. (1999). Prog. Biophys. Mol. Biol. 72, 245-270.

MLPHARE: Otwinowski, Z. (1991). Proceedings of the CCP4 Study Weekend. Isomorphous Replacement and Anomalous Scattering, edited by W. Wolf, P. R. Evans & A. G. W. Leslie, pp. 80-86. Warrington: Daresbury Laboratory.

MOLREP: Vagin, A. A. & Teplyakov, A. (1997). J. Appl. Cryst. 30, 1022-1025. Vaguine, A. A., Richelle, J., Wodak, S. J. (1999). Acta Cryst. D55, 191-205.

PARROT: Zhang, K. Y., Cowtan, K., Main, P. (1997). Methods Enzymol. 277, 53-64.

PROFESS: program by Kevin Cowtan, supported through CCP4

PyMOL: The PyMOL Molecular Graphics System, Schrödinger, LLC.

REFMAC: Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D53, 240-255.

RESOLVE: Terwilliger, T. C. (2004). *J. Synchrotron Rad.* 11, 49-52.

SHELXC: Sheldrick, G.M. (2008). *Acta Cryst.* A64, 112-122.

SHELXD: Schneider, T. R. & Scheldrick, G. M. (2002). *Acta Cryst.* D58, 1772-1779.

SHELXE: Scheldrick, G. M. (2002). *Z. Kristallogr.* 217, 644-650.

SOLVE: Terwilliger, T. C. & Berendzen, J. (1999). *Acta Cryst.* D55, 849-861.

TRUNCATE: French, G. S. & Wilson, K. S. (1978). *Acta Cryst.* A34, 517-525.

Overview

This manual describes the structure solution features of HKL-3000, which combines the data collection and reduction functions of HKL-2000 with tools for data analysis, substructure solution, phasing, and model building for macromolecular crystallography. It merges different crystallographic software into a complete structure solution pipeline for both SAD/MAD and molecular replacement (MR) phasing.

HKL-3000 is built on years of experience solving many structures (including both straightforward and difficult cases) by macromolecular crystallography. Based on that experience, the default algorithms and parameters for each step of the process were chosen to be optimal for the majority of structure solution problems. For most straightforward cases, you should find that the default settings will not need to be changed to solve your structure. However, many procedures of HKL-3000 have an “Advanced Mode” dialog that permits you to change the default parameters (and in some cases choose the algorithm/programs used) as needed to solve more difficult structures.

Like HKL-2000, the interface of HKL-3000 is organized into a set of pages, which can be accessed by clicking on the appropriate tab along the top of the main window. If you are familiar with HKL-2000, note that some new pages have been added, which contain the new functionality present in HKL-3000. In general, the order of the page tabs mirrors the natural work flow of the diffraction structure determination process, so typically you will work through the pages of the program from left to right. These tabs are:

- **Project** – This page is used to define the parameters (such as amino acid sequence) of the protein under study.
- **Collect** – If your instance of HKL-3000 is configured to interact with an X-ray diffraction system, this page is used to define the data sets to be collected. Its tab will be inactive otherwise.
- **Data** – This page is used to select, organize and manage the diffraction image files that contain the currently collected data sets for a given crystal.
- **Summary** – This page lists the collection parameters of the currently selected data sets.
- **Index** – This page is used for indexing, Bravais lattice assignment and refinement of a small subset of diffraction images.
- **Strategy** – Once a representative subset of images has been indexed on the **Index** page, this tab provides tools to predict the completeness, redundancy and number of overlaps for specified data collection settings.
- **Integrate** – Once collection on a full dataset has begun, this page is used to integrate all of the peaks on the diffraction images using the initial refinement settings on the **Index** page.
- **Scale** – This page takes the list of integrated peaks for each diffraction image, identifies identical reflections on different frames, and scales all of the reflections into a single reflection dataset.
- **Structure** – This page contains the functionality to solve structures by either SAD/MAD or molecular replacement.
- **Publication** – This page is used to prepare CIF files for publication of small molecule structures,. Its tab will be inactive unless the small molecule diffraction module is enabled.

- **Macros** – This page may be used by advanced users to add specialized macros during indexing, refinement, integration and scaling.
- **Credits** – This page lists the authors of the HKL-2000/HKL-3000 system.
- **Copyrights** – This page lists the copyright notices for the programs and libraries included with the HKL-2000/HKL-3000 package.

As this manual focuses on structure solution by SAD/MAD and molecular replacement, many of the tabbed pages of the program are beyond the scope of this text (but are described elsewhere, such as in the HKL-2000 manual). This text does focus in some detail, however, on the functions implemented by the **Project** and **Structure** tab pages of HKL-3000.

The SAD/MAD pathway

Defining project parameters

Working with HKL-3000 to solve structures by SAD/MAD begins by clicking the **Project** tab of the program's main window. All structure solution and refinement work associated with a particular protein requires some basic information about the protein and the data being solved. This information is stored in a "project," and is entered on the **Project** page. Three sets of information are necessary to begin SAD/MAD structure solution with HKL-3000: a set of integrated, merged and scaled reflections, the primary polypeptide sequence(s), and the identity of the atom(s) producing the anomalous diffraction signal.

The screenshot shows the HKL-3000 v702c2.db027.ph104.sm036 software interface. The title bar indicates the package is licensed to Iwona Minor at HKL Research, Inc. The main menu includes File, Options, Site Configuration, Crystal Information, Report, and Help. The Project tab is selected, showing a sidebar with tabs: Project, Collect, Data, Summary, Index, Strategy, Integrate, Scale, Structure, Publication, Macros, Credits, and Copyrights.

Project

Name: project1
 Crystal: crystal1
 Experimenter: anna
 Date: Apr 21, 2011

Buttons: Load, Save, Edit Project, Edit Crystal, New Project, New Crystal, Evaluate Model

Principal Component: Protein

Protein Data

Name:
 Number of Residues:
 Organism:
 Description:
 Molecular Weight:
 NCBI accession code:
 Swiss-Prot accession code:
 No. of Homologues:
 Last check in PDB:

Phasing Method: SAD/MAD
 Source of Anomalous Signal:

Ala (A):	Gly (G):	Met (M):	Ser (S):
Cys (C):	His (H):	Asn (N):	Thr (T):
Asp (D):	Ile (I):	Pro (P):	Val (V):
Glu (E):	Lys (K):	Gln (Q):	Trp (W):
Phe (F):	Leu (L):	Arg (R):	Tyr (Y):

Anomalous Wizard

Scaled Data

Structure	File Name:	Cell:	Space Group:	Wavelength:	Mode:	Order:
<input checked="" type="checkbox"/>	/home/anna/data/Lysozyme_Marcin/data/proc01/output.sca	78.918 78.918 36.838 90.000 90.000 90.000	P43212	1.54178	Peak	1

Buttons: Add File, Remove File, Change Mode, Change Order

To create a new project, click the blue “New Project” button, which creates a dialog window titled “Setup Project”. In addition to specifying some optional parameters of the project, such as its name and a description, the dialog allows you to specify the polypeptide sequence in the box labeled “Sequence.”

Sequences may be automatically obtained from a number of different sources. By selecting “File” in the pulldown list, the “Load” button appears. Clicking this button opens a file selection dialog, where you may select a text file from the file system containing the sequence

The other options for the “Obtain Sequence from” pulldown allow you to download a sequence from a variety of online resources:

- NCBI – The NCBI GenBank Protein database (<http://www.ncbi.nlm.nih.gov/protein>)
- Swiss-Prot – The UniProtKB Knowledgebase (<http://www.uniprot.org>)
- MCSG – The Midwest Center for Structural Genomics database (<http://www.mcsg.anl.gov/>)
- CSGID – The Center for Structural Genomics of Infectious Diseases database (<http://www.csgid.org>)

- NYSGRG – The New York Structural Genomics Research Consortium database (<http://www.nysgrc.org>)
- PDB – The Protein Data Bank (<http://www.pdb.org/>)

Type the identifier for the appropriate database in the text box and click “Download”.



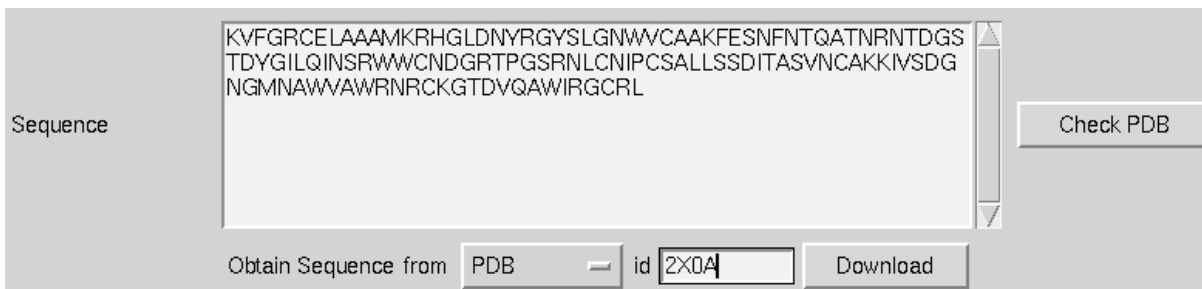
Obtain Sequence from PDB id Download

If you use the PDB, the program can check the id for you after clicking on the “Check PDB” button located to the right of the “Sequence” box.



Sequence Check PDB

When the sequence is found from one of the databases, it is automatically filled in the “Sequence” box.

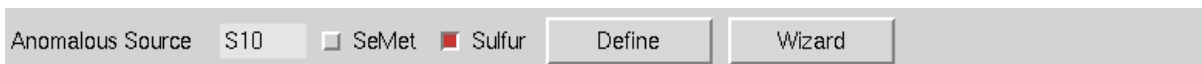


Sequence Check PDB

Obtain Sequence from PDB id Download

You may also cut and paste the sequence directly into the text box.

The next step is to identify the source of the anomalous signal. HKL-3000 identifies this source as a text string with the atomic symbol for the anomalously scattering element followed by the expected number of atoms of that element in the asymmetric unit. You may explicitly enter in the source string by clicking “Define,” or selecting the element on a periodic table by clicking “Wizard” and following the prompts. If you are phasing using sulfur or selenomethionine (SeMet), clicking the appropriate button will select the appropriate element and automatically determine the number of atoms using the sequence.



Anomalous Source ☐ SeMet ☒ Sulfur Define Wizard

You may complete and close the “Setup Project” dialog by clicking “Done.” The completed information now appears in the “Project” window. After saving project information, you may edit it later by clicking the blue “Edit Project” button.

HKL-3000 v702c2.db027.ph104.sm036 Package Licensed to Iwona Minor at HKL Research, Inc.

File Options Site Configuration Crystal Information Report Help

Project Collect Data Summary Index Strategy Integrate Scale Structure Publication Macros Credits Copyrights

Project

Name: project1
 Crystal: crystal1
 Experimenter: anna
 Date: Apr 22, 2011

Load Save
 Edit Project Edit Crystal
 New Project New Crystal
 Evaluate Model

Principal Component: Protein

Protein Data

Name:
 Number of Residues: 129
 Organism:
 Description:
 Molecular Weight: 14313.65
 NCBI accession code:
 Swiss-Prot accession code:
 No. of Homologues: 618
 Last check in PDB:

KVFGRCLEAAAMKRHGLDNYRGYSLGNWVCAAKFESNFTQA
 TNRINTDGSTDYGILQINSRWWCNDGRTPGSRNLCNIPCSALL
 SSDITASVNCACKIVSDGNGMINAWAWNRCKGTDVQAWIRG
 CRL

Phasing Method: SAD/MAD
 Source of Anomalous Signal: S10

Ala (A): 12	Gly (G): 12	Met (M): 2	Ser (S): 10
Cys (C): 8	His (H): 1	Asn (N): 14	Thr (T): 7
Asp (D): 7	Ile (I): 6	Pro (P): 2	Val (V): 6
Glu (E): 2	Lys (K): 6	Gln (Q): 3	Trp (W): 6
Phe (F): 3	Leu (L): 8	Arg (R): 11	Tyr (Y): 3

Anomalous Wizard

Scaled Data

Structure	File Name:	Cell:	Space Group:	Wavelength:	Mode:	Order:
Select	/home/anna/data/Lysozyme_Marcin/data/proc01/output.sca	78.918 78.918 36.838 90.000 90.000 90.000	P43212	1.54178	Peak	1

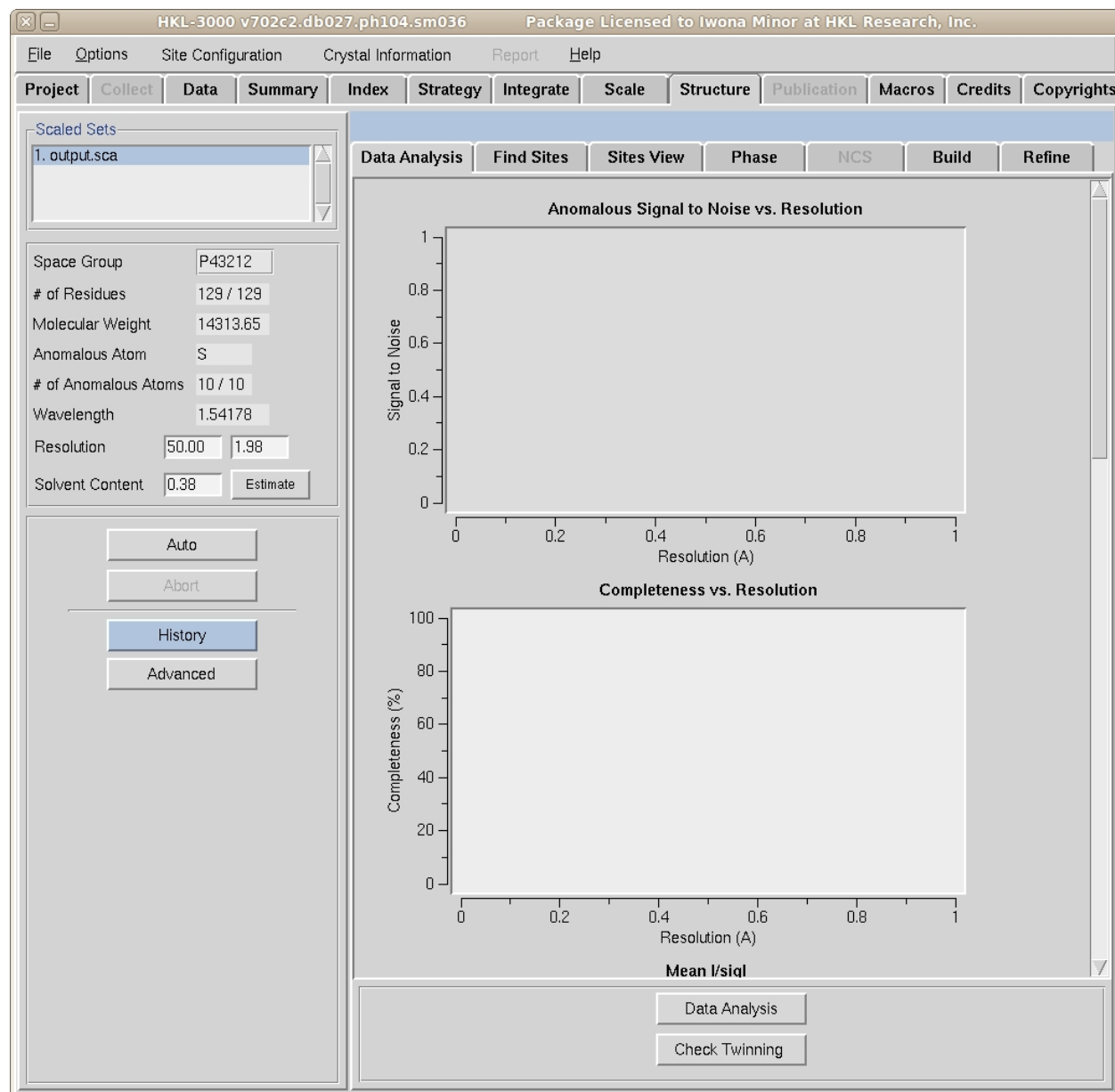
Add File
 Remove File
 Change Mode
 Change Order

At the bottom of the **Project** page is an area labeled “Scaled Data.” This box is used to manage the data files containing the integrated, scaled and merged reflections. The scaled reflections come from files ending in *.sca* as produced by the *Scalepack* program or by the **Scale** tab of HKL-2000/3000. (Please note that HKL-2000/3000 will typically produce two * *.sca* reflection files after a scaling run; one with a user-specified name, and a file called *scalepack.sca*. The *scalepack.sca* file is used internally by HKL-2000/3000 and should not be chosen.)

Each blue box in the “Scaled Data” list represents one *.sca reflection file. Add new data files by clicking “Add File.” To remove a file or change the mode (i.e. what component of the anomalous diffraction the file represents, such as peak, inflection, etc.), select the desired file with the “Select” button, and click “Remove File” or “Change Mode” respectively. Finally, the reflection files will be processed in the order they are listed; click “Change Order” to modify this order.

Data analysis

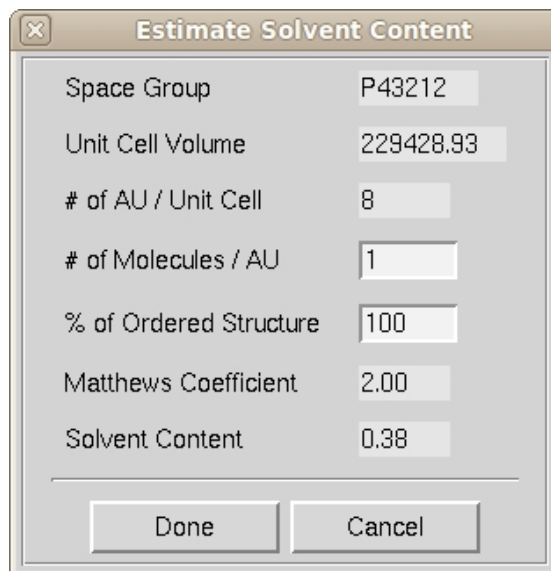
After defining the general project parameters and selecting the reflection files, select the **Structure** tab along the top of the main window. A typical view of the Structure page before running data analysis is shown below. Note that the **Structure** page contains a number of subpages: **Data Analysis**, **Find Sites**, **Sites View**, **Phase**, **NCS**, **Build** and **Refine**. These subpages are also navigated by a set of tabs. Note as well that the contents of a subpage may not fit within the window, but the rest of the subpage may be viewed using the scrollbar on the right side of the window.



All previously collected information is summarized in the sidebar on the left. The numbers of amino acid residues in a chain and in an asymmetric unit (AU) are calculated based on the sequence (the two numbers are separated by a slash). The molecular weight is the weight of a single polypeptide in Daltons. In the case where Se or S atoms are the source of anomalous signal, the numbers of anomalous atoms

in a chain and in the asymmetric unit are calculated automatically by the program (again separated by a slash). If the source is selenomethionine (SeMet), an N-terminal SeMet will be omitted from the automatic count, as it is presumed that the initial SeMet residue is cleaved or disordered.

HKL-3000 will automatically estimate the number of molecules in the AU, the Matthews coefficient and the solvent content. HKL-3000 calculates the probable solvent content using a formula proposed by Kantardjieff and Rupp (*Kantardjieff, K. A., Rupp B. (2003). Protein Science 12, 1865-1871*) for different numbers of molecules in the AU and chooses the value with the most reasonable solvent content. (Typically, the solvent content in protein crystals falls in the range of 30-70%). You may override the number of molecules in the AU by clicking the “Estimate” button, which opens a new window.



The image shows a dialog box titled "Estimate Solvent Content". It contains several input fields and two buttons at the bottom. The fields are: Space Group (P43212), Unit Cell Volume (229428.93), # of AU / Unit Cell (8), # of Molecules / AU (1), % of Ordered Structure (100), Matthews Coefficient (2.00), and Solvent Content (0.38). The buttons are "Done" and "Cancel".

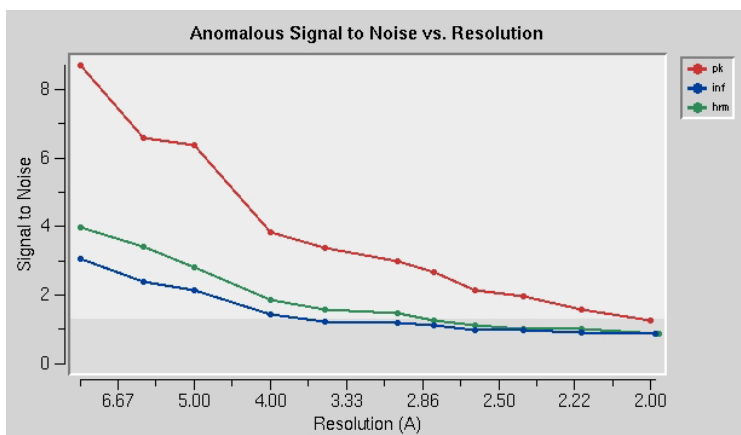
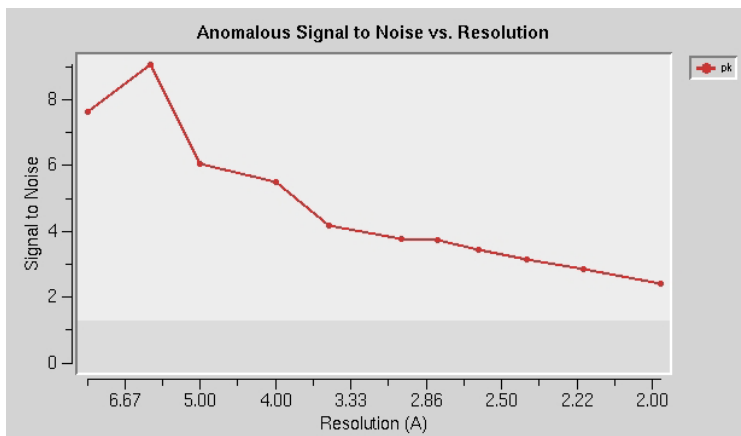
Field	Value
Space Group	P43212
Unit Cell Volume	229428.93
# of AU / Unit Cell	8
# of Molecules / AU	1
% of Ordered Structure	100
Matthews Coefficient	2.00
Solvent Content	0.38

There you can edit two fields: the number of molecules per AU and a percentage of each molecule predicted to be ordered. Changing the number of molecules will recalculate the solvent content and the Matthews coefficient. If these values appear in red, this suggests that the calculated values are unusual and possibly incorrect. Changing the percentage of the molecule with ordered structure does not change the solvent content and the Matthews coefficient, but does change the number of residues sent to the model building procedures. For example, a value of 90% of ordered structure indicates that if there are 200 residues in the project sequence, the model building programs will be instructed to look for 180 residues (times the number of molecules) in the AU.

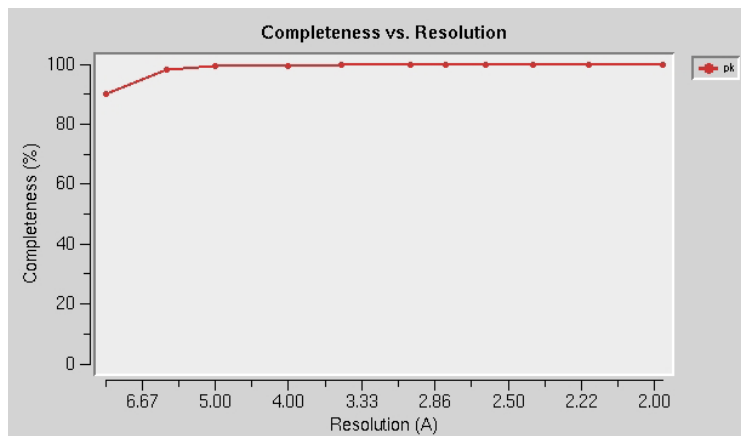
Data analysis of the reflection data is provided by the SHELXC and TRUNCATE programs, which is started by clicking the “Data Analysis” button. The results are presented on 2 (or 3) plots: “Anomalous Signal to Noise vs. Resolution,” “Completeness vs. Resolution” and “Wavelength Pairs” (MAD data only). In the example plots shown below, the upper plot of each pair shows the SAD results, and the lower plot the MAD results.

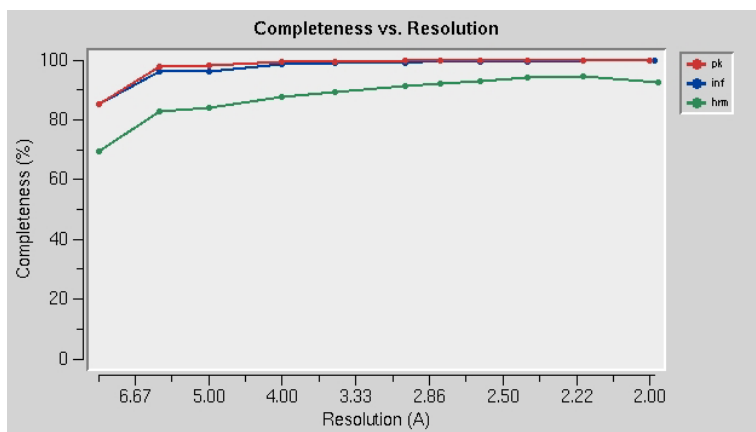
The “Anomalous Signal to Noise vs. Resolution” plot shows the presence or absence of an anomalous signal in different resolution shells as measured by the signal-to-noise ratio (SNR) as calculated by SHELXC. The gray shaded region on the graph shows SNR values below 1.3, where the anomalous signal is probably insignificant. In the SAD case presented below there is a strong anomalous signal for the data in resolution shells. In the MAD case, the anomalous signal for each dataset is calculated separately, on lines marked “pk” (peak wavelength), “inf” (inflection wavelength), “hrm” (high energy remote wavelength) and “lrm” (low energy remote wavelength). In this case, the peak anomalous signal is

significant for almost all resolution shells, while the inflection and low energy signals are significant below 3.5 Å or so.

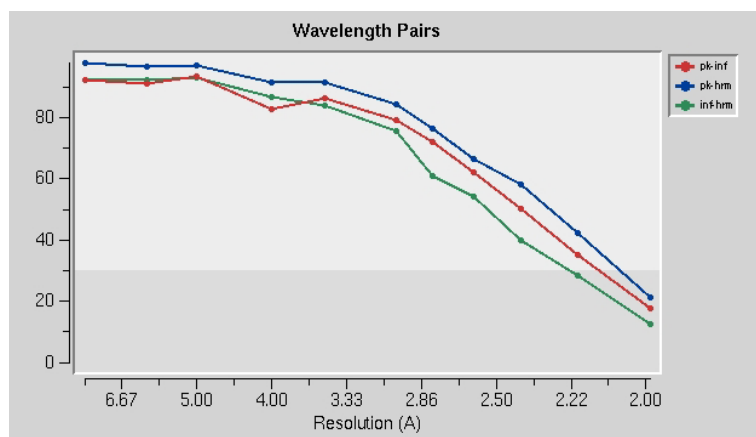


The “Completeness vs. Resolution” data represents completeness as a function of resolution shell. (As above, for MAD data, each dataset is plotted separately and labeled as above.) In the examples below, there is some incompleteness of the lowest resolution data. This may be caused by “overloads” of the lowest resolution, and most strongly diffracting, reflections. In this case, the diffraction experiment may not have been performed optimally. Complete and accurate low resolution data are important for phase solution and improvement.

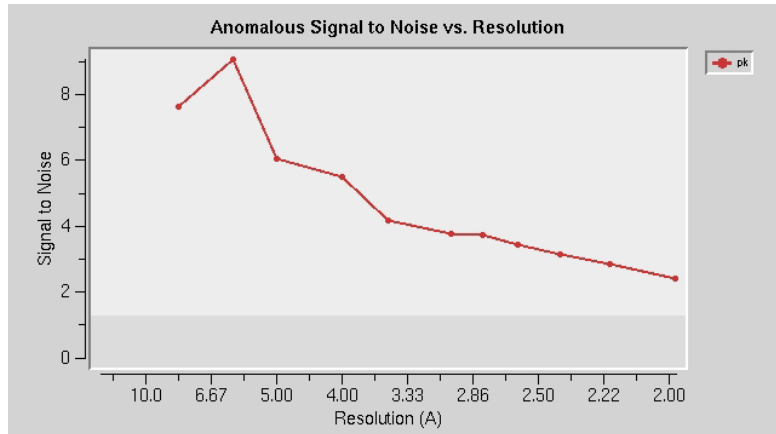




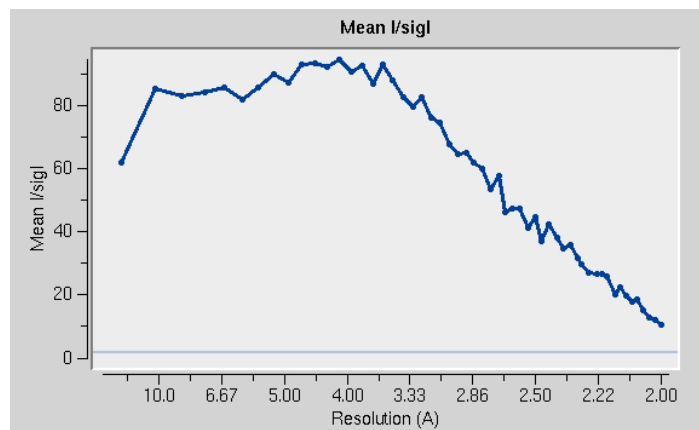
The “Wavelength Pairs” is presented only for MAD results. It shows the correlation coefficient percentage (CC) between the signed anomalous difference ΔF for two different wavelengths as a function of the resolution shell. It is recommended to truncate the data where CC falls below about 25 to 30% (Scheldrick, G. M. & Schneider, T. R. (2001). *Proceedings of the NATO Advanced Study Institutes on Methods in Macromolecular Crystallography and Chemical Perspectives in Crystallography of Molecular Biology*, 25 May – 4 June 2000, Erice, Italy. *Methods in Macromolecular Crystallography*, edited by D. Turk & L. Johnson, pp. 72-81. IOS Press.). The pair of wavelengths compared for each line is indicated using the same abbreviations as listed above. The region of the plot where the CC is below 30% is indicated in gray.



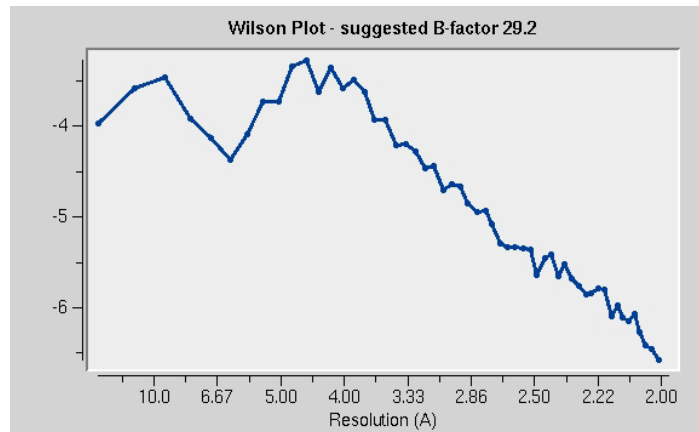
The “Check Twinning” button should be used when you expect twinning problems with your data, which will populate the remaining graphs in the subpage. In addition, note that the three plots described above will be re-rendered with a modified distribution of resolution shells, to match the horizontal axis of the four new plots created (for example, see the re-rendered “Anomalous Signal to Noise vs. Resolution” plot below).



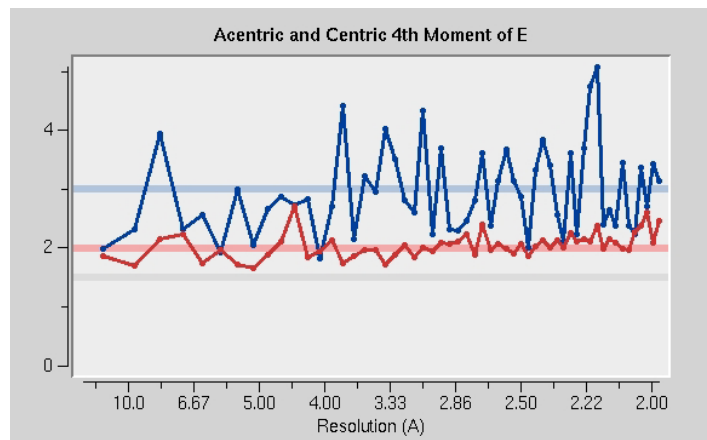
The new plots created are “Mean $I/\sigma I$,” “Wilson Plot,” “Acentric and Centric 4th Moment of E” and “Cumulative Intensity Distribution (Acentric and Centric),” as shown below. The “Mean $I/\sigma I$ ” plots the mean $I/\sigma I$ value for each resolution shell, with a pale blue line at $\langle I/\sigma I \rangle = 2$. This value is in keeping with the convention of only using reflections up to a resolution where the highest-resolution shell has $\langle I/\sigma I \rangle \geq 2$.



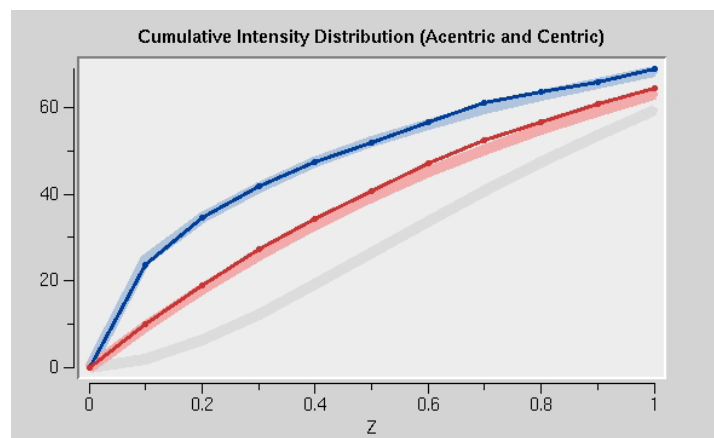
The “Wilson Plot,” as calculated by TRUNCATE, plots the logarithm of the mean I_{obs} value for each resolution shell divided by the sum of atomic scattering factors squared, as a function of $2 \sin(\theta/\lambda)$ (here labeled by resolution). The slope of the high resolution data (< 4.0 Å) is used to estimate the Wilson B-factor for the dataset, which is indicated in the title as the “suggested B-factor”. High values of the Wilson B-factor may indicate a problem with the data, for example crystal decay.



The “Acentric and Centric 4th Moment of E ” plot, which is also calculated by TRUNCATE, is a measure of twinning: the thin, dark blue lines represent the 4th moment of the normalized structure amplitudes E for centric reflections, and the thin, dark red lines the 4th moment of E for acentric reflections. If there is no twinning, the 4th moments of E should be about 2.0 for acentric reflections and 3.0 for centric reflections. These ideal values are represented by thick, pale lines. By contrast, the gray line shows the expected 4th moment of E at 1.5 for acentric reflections (2.0 for centric reflections) of perfectly twinned data. How the experimental curves compare to the expected values can be used as an indicator of the presence of twinning.

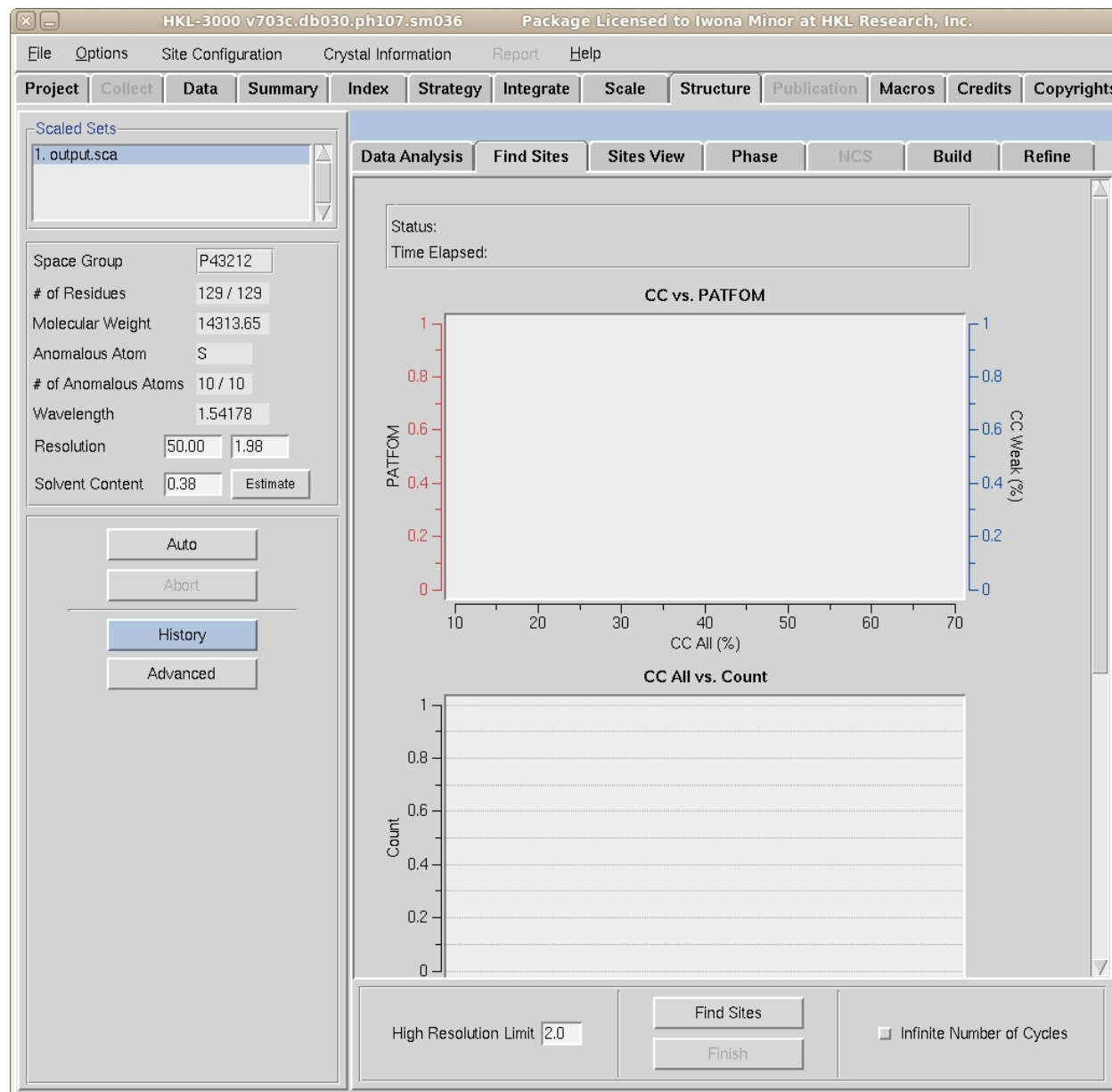


The “Cumulative Intensity Distribution (Acentric and Centric)” plot the cumulative distribution function (CDF) for the intensities in the data set. (The CDF is plotted with the probability (Z) along the horizontal axis and the intensity values along the vertical axis, following the convention used by TRUNCATE.) The blue lines represent the CDF for centric reflections, and the red lines the CDF for acentric reflections. Thick, pale lines represent the theoretical CDF for data with no twinning, while thin, dark lines display the experimentally observed CDF. The pale gray line shows the theoretical (sigmoidal) CDF for acentric reflections for perfectly twinned data. If your crystal is not twinned, then the thin, dark blue and red lines should coincide with the thick, pale blue and red lines. If the experimental acentric CDF overlaps with the gray line the data likely exhibit perfect merohedral twinning.



Substructure determination

The next step after analyzing your data is locating the substructure – the heavy atoms that produce the anomalous signal. Substructures are referred to as “sites” in HKL-3000. Move to the **Find Sites** subpage by clicking the **Find Sites** tab on the **Structure** page. As in the **Data Analysis** subpage, graphs and other data that do not fit within the screen can be accessed by moving the scrollbar to the right. The Find Sites subpage displays three graphs: “CC vs. PATFOM,” “CC All vs. Count,” and “CC All vs. Mean Occupancy.”

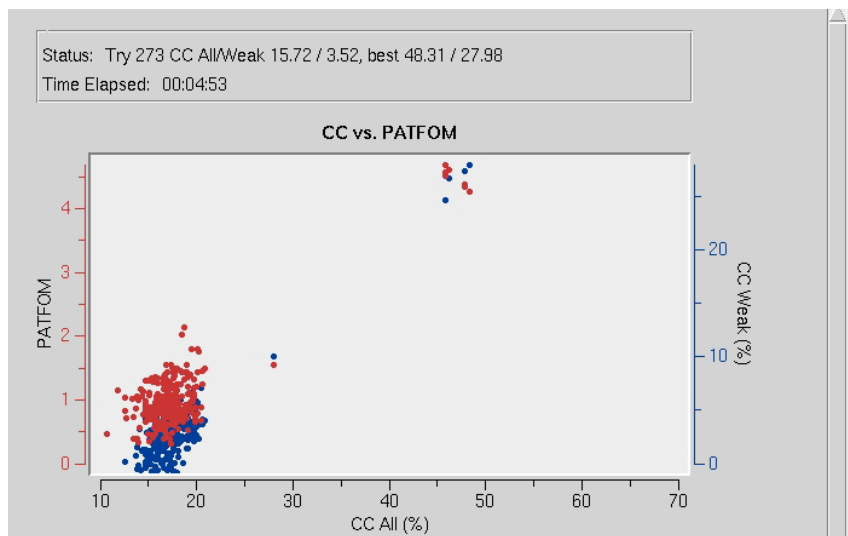


Substructure determination is done with the SHELXD program. At the bottom of the **Find Sites** subpage is a box for setting the “High Resolution Limit,” which controls the highest resolution data used in the substructure search. Sheldrick’s rule of thumb is to use a high resolution cutoff at least 0.5 Å greater than the highest resolution reflections in the set. For example, if the highest resolution reflections diffract to 2.0 Å, try a substructure limit of 2.5 Å. It is also useful to consult the “Anomalous Signal to Noise vs.

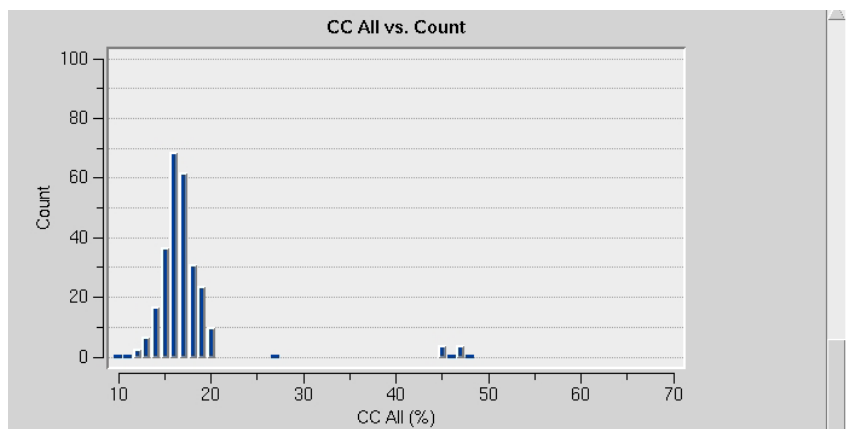
Resolution” plot on the **Data Analysis** subpage and determine to what resolution there is a significant anomalous signal. Including high resolution data without a significant anomalous signal will only add noise to the substructure search process.

After choosing the high resolution limit, click the “Find Sites” button, and the program will search for substructure solutions until it finds a one with CC All/CC Weak percentages above 30/15, as recommended by Sheldrick. (The first number is the correlation coefficient percentage for “All” reflections, and the second the CC percentage for “Weak” reflections. See the SHELXD documentation for more information.) In some difficult cases, the first solution with All/Weak CC values above 30/15 may not prove to be the best overall substructure solution. In this case, check the “Infinite Number of Cycles” box to instruct SHELXD to search for sites indefinitely. You will be required to press the “Finish” button when you are satisfied you have found an adequate substructure.

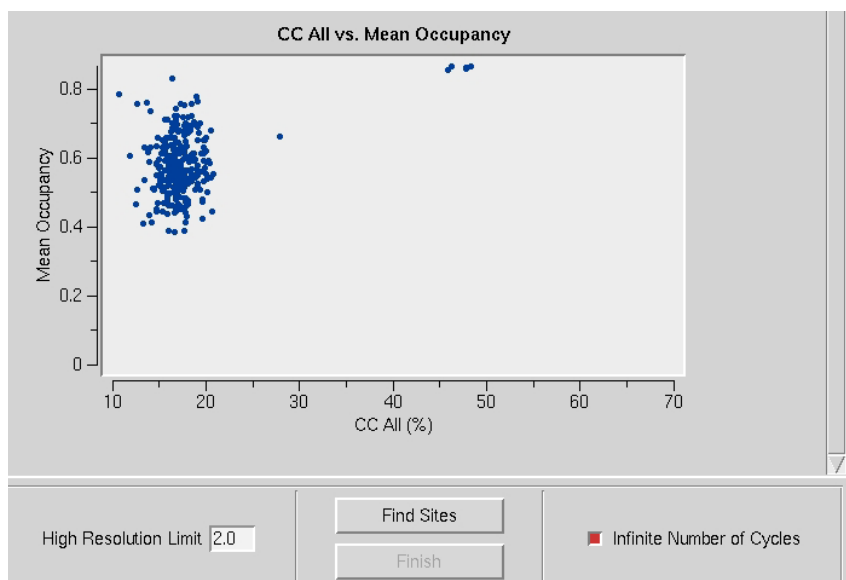
Three plots summarize the independent substructure solutions. The most important is the first: “CC vs PATFOM,” which plots both the *CC Weak* (blue dots) and the *PATFOM* (Patterson figure of merit, red dots) statistics as a function of the *CC All* statistic. Two things are important when searching for good substructure solutions. First, all three values (*CC All*, *CC Weak*, and *PATFOM*) for the best solutions should be significantly greater than the corresponding values for the other solutions. Second, there should be clear separation between the best solutions and the large population of non-solutions (which may be more clearly seen when “Infinite Number of Cycles” is selected). Note in the “CC vs. PATFOM” plot shown below, the solutions group into two distinct clusters.



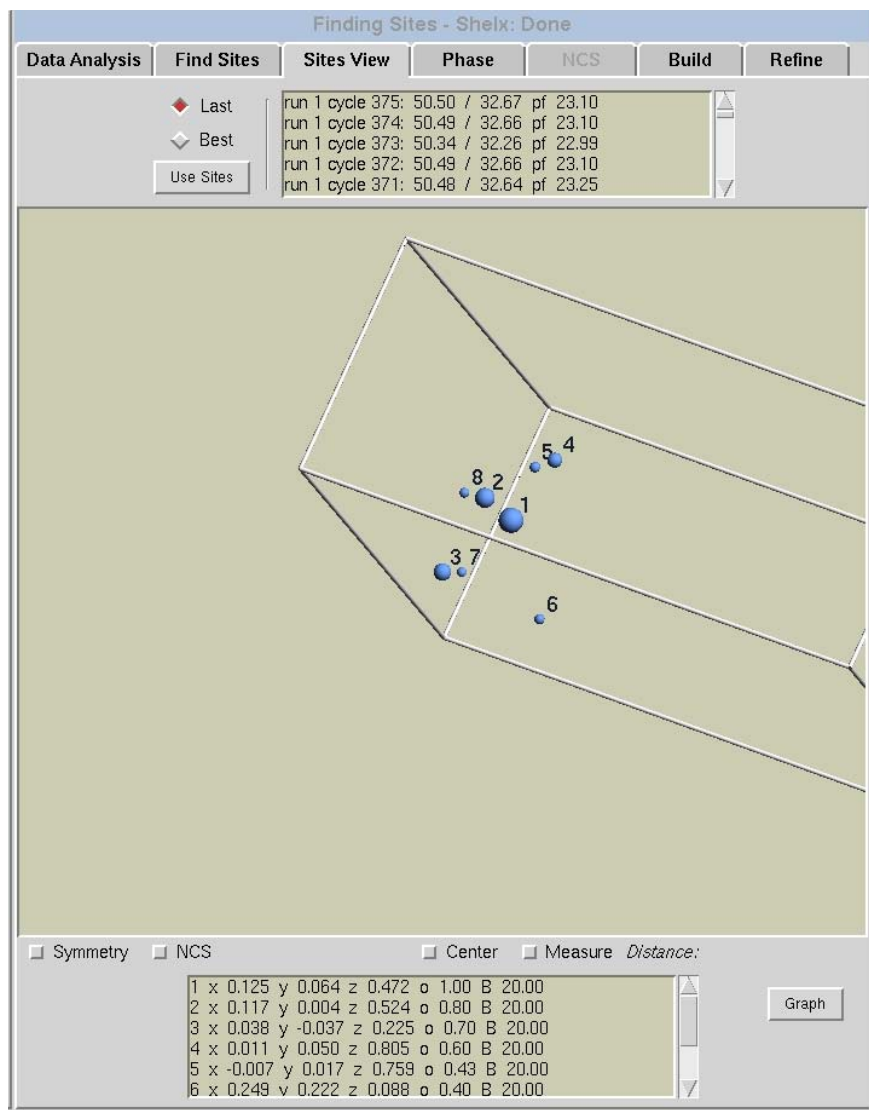
The second plot “CC All vs. Count” shows a histogram of the CC All percentages for all solutions calculated, which may be used to determine the relative size of each cluster of solutions. From that plot you know how many times a particular solution occurs.



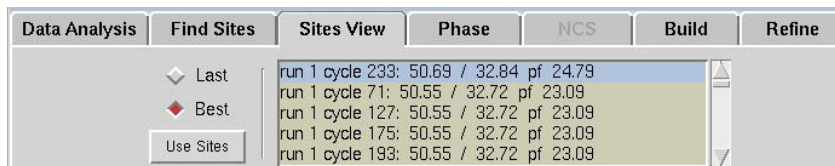
The third plot “CC All vs. Mean Occupancy” shows the mean occupancy for all sites in for each substructure solution as a function of *CC All*. SHELXD uses relative occupancies, where the strongest site (or atom) in the substructure is assigned occupancy 1.0, and the other sites are scaled relative to the first.



The results of substructure determination, as calculated on the **Find Sites** subpage, are displayed on the **Sites View** subpage (shown below). The sites of the solution are displayed in a three-dimensional representation of the unit cell. The 3-D display of the sites may be rotated by left-clicking and dragging on the window, and zoomed in or out by right-clicking and dragging.



The currently selected substructure solution (by default, the solution with the highest *CC All*, *CC Weak* and *PATFOM* values in the most recent **Find Sites** run) is shown as a set of blue spheres, numbered in order of occupancy (with 1 the site with highest occupancy, and so on). The radii of the spheres are proportional to the occupancy of the sites (higher occupancy sites have larger spheres).

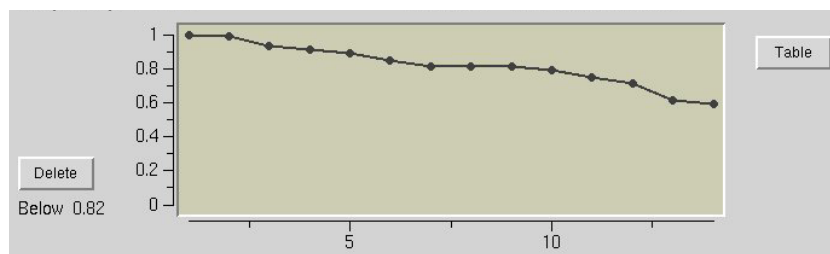


All of the substructure solutions are shown in the list (above) at the top of the **Sites View** subpage, which can be sorted chronologically (select "Last") or in descending order of correlation coefficient (select "Best"). Note that if "Find Sites" on the **Find Sites** subpage is run multiple times (e.g., with different parameter settings), all solutions from all runs are listed. Clicking on another solution in the list of solutions at the top of the subpage displays the sites in the solution as a set of red cones; click "Use Sites" to select that solution instead. When a new solution is chosen, the red cones change into blue spheres and the previous set of blue spheres are removed.

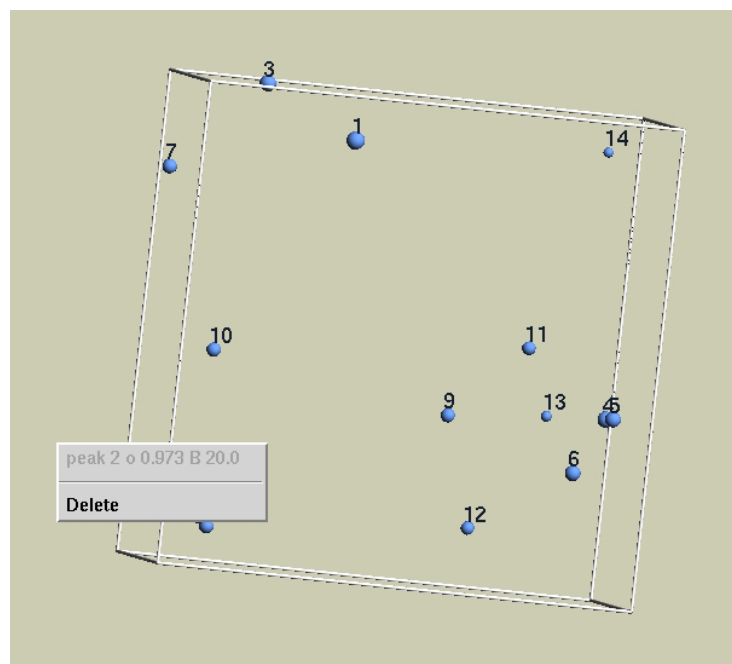
In addition to rotating and zooming in the on the sites as described above, there are other tools for managing the currently selected substructure (below). Select “Symmetry” to display symmetry-related sites (light blue spheres). Select “Center” and then click on a site with the middle mouse button to center on that site. Select “Measure” and select two sites with the middle mouse button to display the distance between them. (The NCS checkbox is used for viewing the results of non-crystallographic symmetry detection, as described in the *NCS* section below).

<input type="checkbox"/> Symmetry	<input type="checkbox"/> NCS	<input type="checkbox"/> Center	<input checked="" type="checkbox"/> Measure	Distance: 5 11 12.06 Å
1	x 0.606	y 0.900	z 0.042	o 1.00 B 20.00
2	x 0.841	y 0.907	z -0.207	o 0.99 B 20.00
3	x 0.796	y 1.007	z -0.211	o 0.93 B 20.00
4	x 0.623	y 1.006	z 0.128	o 0.91 B 20.00
5	x 0.741	y 1.078	z 0.123	o 0.89 B 20.00
6	x 0.626	y 1.020	z 0.181	o 0.85 B 20.00

The table at the the bottom of the Site View window (above) lists all sites in the currently selected substructure, including their coordinates (x, y, z), occupancies (o) and B-factors (B). Click the “Graph” button to toggle the display from a table to a graph (below) showing occupancies of the sites.



You may manually remove low occupancy sites from the selected substructure before phasing two different ways. One way is to left-click the site occupancy graph (above) at a desired occupancy, which appears under the “Delete” button to the left. Click “Delete” to remove all sites below the specified occupancy; this is especially useful if there is a significant difference in occupancy between groups of sites. Individual sites may also be removed by middle-clicking a sphere in the 3-D window and selecting “Delete” in the popup menu (below). Note that both “Center” and “Measure” must be unchecked first.

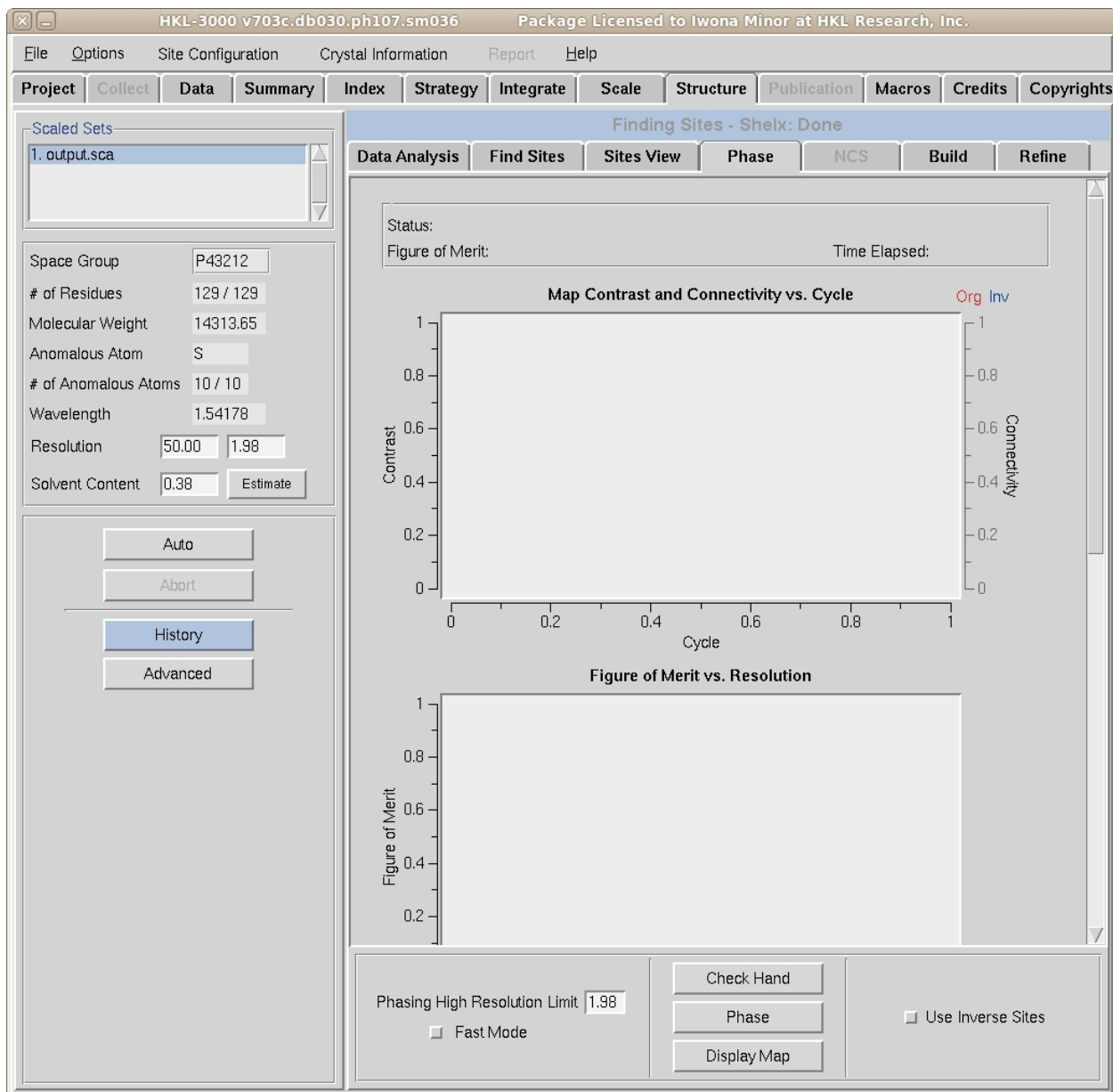


Other, more rarely used features relating to substructure solution, such as setting the minimum site-to-site distance, identifying the number of disulfides bridges, allowing sites in special positions, etc. may also be accessed through the “Advanced” button in the left sidebar. The “Advanced Mode” dialog also permits saving or loading substructure atom positions to a file.

Phasing

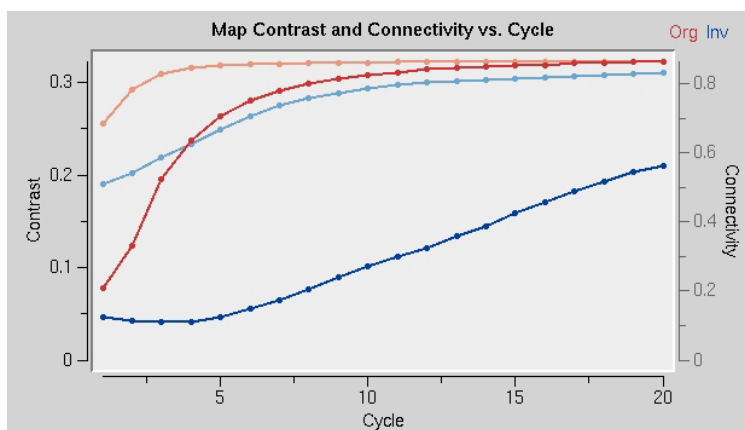
NOTE: while the examples below show phasing using SAD data, the tools described can also be used to solve structures by MAD.

Once you have selected (and possibly edited) a substructure solution, you may move on to the **Phase** subpage via its lab. The subpage is shown below. As before, some of the contents of the subpage may not be visible without using the scrollbar to the right.



Click the "Check Hand" button first to determine the handedness of the substructure—some substructures may be inverted from their positions in the phased structure, which cannot be detected by SHELXD. To determine the proper hand, HKL-3000 will run SHELXE twice to calculate and refine two sets of electron density maps, using phases from either the original substructure or the same substructure with positions inverted (i.e., the opposite hand) respectively.

The first plot (“Map Contrast and Connectivity vs. Cycle”) interactively displays the results from both SHELXE runs. Statistics for the map calculated from the original sites are in red, and for the map from the inverted sites in blue. Two different parameters shown—map connectivity (lighter lines), and map contrast (darker lines). Typically one of the two substructure hands will result in a much higher quality electron density map, and this is presumed to be the correct hand. For example, in the figure below, the original substructure produces a much better map and thus it does not need to be inverted.



In fact, when the inverted substructure produces higher contrast values, HKL-3000 will automatically select the “Use Inverse Sites” option and will change the space group if appropriate (i.e. for enantiomorphic space groups). If your data is in an enantiomorphic space group, and the space group is changed after selecting inverse sites, you should go back to the **Scale** tab and rescale the data in the new space group to keep consistency in your data files.

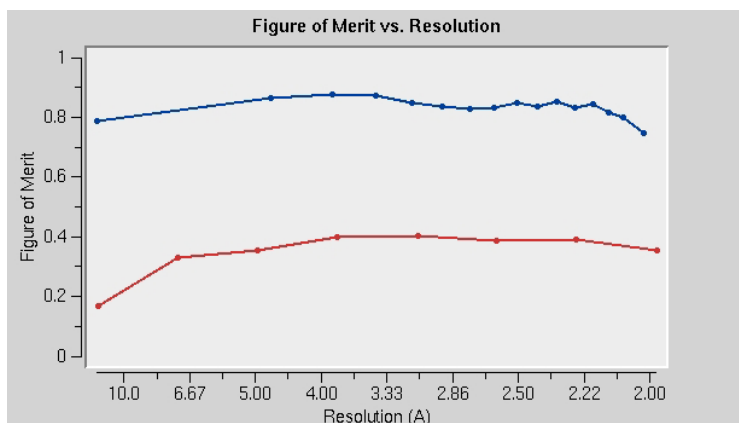
NOTE: Currently site inversion does not work correctly for space groups *I4₁*, *I4₁22* and *F4₁32*.

Phasing High Resolution Limit <input type="text" value="2.00"/>	<input type="button" value="Check Hand"/>	<input checked="" type="checkbox"/> Use Inverse Sites
<input type="checkbox"/> Fast Mode	<input type="button" value="Phase"/>	
	<input type="button" value="Display Map"/>	

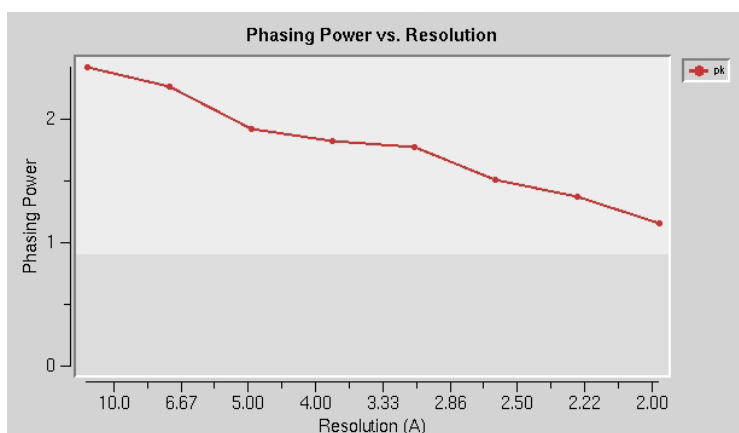
After determining the proper handedness for the substructure, click the “Phase” button (above) to begin calculating experimental phases. The “Phasing High Resolution Limit” controls which reflections are processed by the MLPHARE program, which may be a lower resolution than the high resolution limit for the whole data set (or sets). However, when the MLPHARE phases are refined by density modification, they will be extended to all reflections (see below).

After clicking “Phase,” the SHELXE run will be repeated, and the substructure will be improved with MLPHARE, which refines the coordinates, occupancies and B-factors for the substructure atoms and uses the refined positions to generate phase information. HKL-3000 will show statistics from phasing in two plots: “Figure of Merit vs. Resolution” and “Phasing Power vs. Resolution”, as well as a table with the current values of the substructure coordinates, occupancies and B-factors.

The red line on the “Figure of Merit vs Resolution” plot (below) shows the mean figures of merit (FOM) values for the MLPHARE-calculated phases by resolution shell. The mean FOM by resolution shell for the phases further improved by DM (see below) are shown in a blue line. This plot is dynamically redrawn through subsequent runs and cycles of refinement.



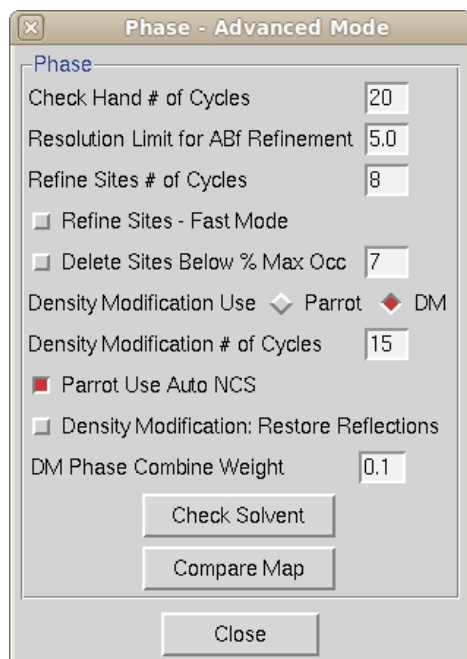
The “Phasing Power vs. Resolution” displays the mean “phasing power” by resolution shell (see the MLPHARE manual for more details). Phasing power values greater than 0.9 are considered to be significant, and accordingly the region of the plot below 0.9 is marked in gray.



The final chart displays the current refined values for the substructure atoms. Like the other plots on the Phase subpage, this chart is dynamically updated through subsequent refinement cycles.

	x	y	z	o	b	d1	d2	d3
1	0.605	0.900	0.041	1.00	22.80	0.02	0.09	0.09
2	0.841	0.908	-0.207	0.81	18.65	0.02	0.08	0.08
3	0.796	1.007	-0.212	0.74	18.29	0.03	0.06	0.06
4	0.623	1.007	0.128	0.82	19.72	0.05	0.06	0.06
5	0.742	1.077	0.123	0.84	20.37	0.03	0.08	0.08
6	0.625	1.019	0.178	0.80	18.96	0.04	0.16	0.18
7	0.825	0.929	-0.204	0.80	21.78	0.02	0.18	0.18
8	0.653	1.065	-0.031	0.66	17.34	0.03	0.07	0.07
9	0.801	1.018	-0.265	0.80	23.97	0.04	0.15	0.19

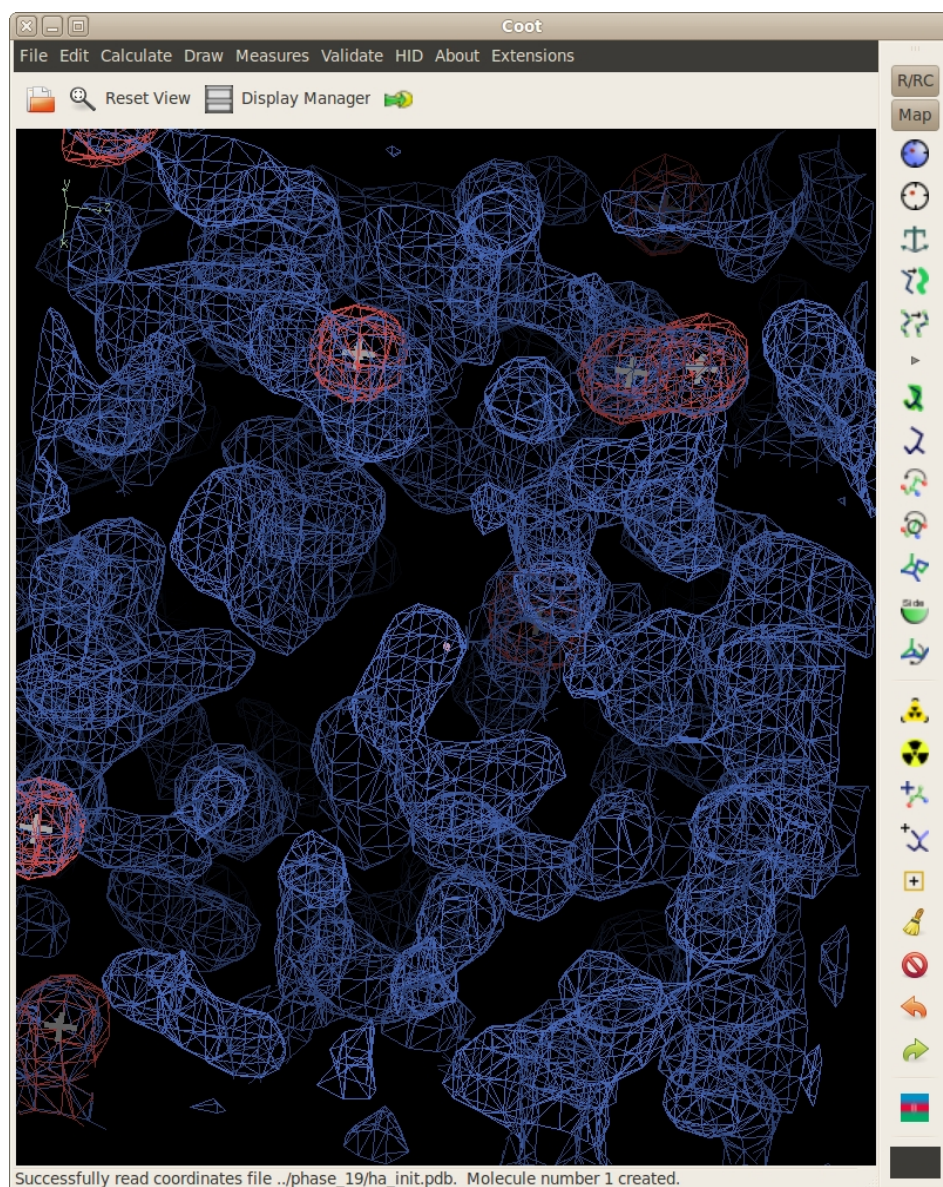
In the initial MLPHARE run, B-factors are refined isotropically. The results from each MLPHARE cycle are analyzed, and some atoms in the substructure may be removed in the subsequent runs on the basis of high B-factor values and/or low occupancy. The default cutoff values prompting the removal of a substructure atom may be altered in the “Advanced Mode” dialog (click the “Advanced” button in the left sidebar.) A number of other default parameters of the phasing process may be altered here as well.



If any site is removed, an additional MLPHARE run will be performed, and in the last run B-factors will be refined using anisotropic approximation. See the Advanced Mode dialog to modify the default resolution limit for anisotropic (ABf) refinement. The table showing information on anomalous sites also shows the distances each atom shifted in position (in Å) after the three main MLPHARE runs: position and occupancy refinement (d1), position and isotropic B-factor refinement (d2), and position and anisotropic B-factor refinement (d3).

After the substructure is refined and the initial phases are calculated with MLPHARE, HKL-3000 refines the phases by density modification (solvent flattening). By default, this is done using the DM program, though Parrot may be used instead (change this default through the “Advanced Mode” dialog). As mentioned above, the mean FOM for the phases calculated by the current cycle of density modification is dynamically updated on the FOM plot (in blue). You may change the number of DM cycles using the “Advanced Mode” menu.

After density modification is completed, HKL-3000 automatically launches COOT to display the resulting electron density maps using the phases calculated. The COOT instance is configured to show three sets of information: the refined anomalous substructure, the anomalous density map red, and the experimental density map using the calculated phases in blue (see below).



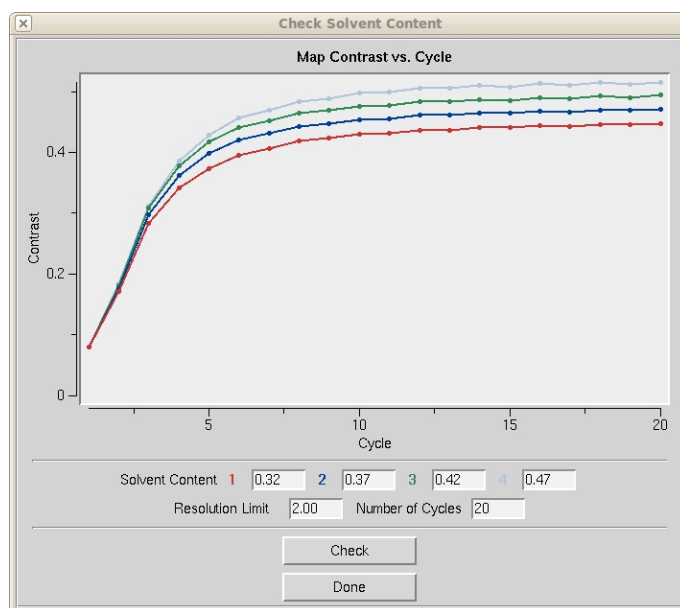
If the substructure was determined properly, the peaks on the anomalous map should correspond to the positions of the substructure atoms. If you are satisfied with the quality of the experimental map, you may proceed to the next steps of structure determination by moving on to the **Build** subpage.

If you are not satisfied with the phasing results, you have several options. First, you may go back to the **Sites View** tab to check the refined substructure, and then use the refined sites as the input for another round of phasing. Note that the refined substructure has been added to the list of solutions on the **Sites View** subpage (below). Click on a refined site, click the “Use Sites” button, and phase again.

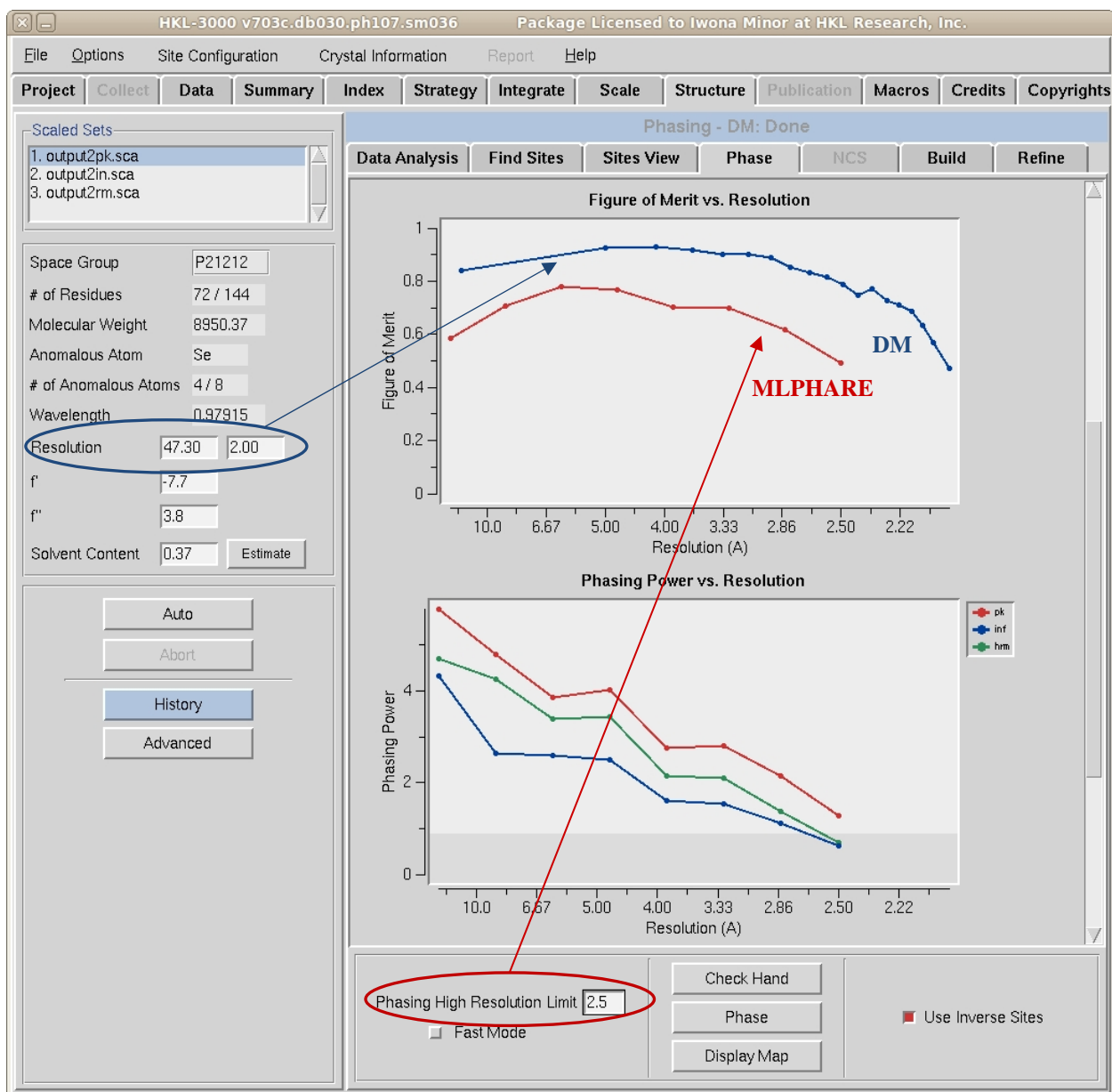
Phasing - DM: Done						
Data Analysis	Find Sites	Sites View	Phase	NCS	Build	Refine
<div><div><div>◊ Last</div><div>◆ Best</div></div><div>Use Sites</div></div>		<div><div>run 1 cycle 233: 50.69 / 32.84 pf 24.79</div><div>run 1 cycle 233: 50.69 / 32.84 pf 24.79 refined_1</div><div>run 1 cycle 233: 50.69 / 32.84 pf 24.79 map_find_1</div><div>run 1 cycle 71: 50.55 / 32.72 pf 23.09</div><div>run 1 cycle 127: 50.55 / 32.72 pf 23.09</div></div>				

Second, you may also use the sites identified after the analysis of the anomalous map, which sometimes reveals additional atoms in the substructure. These substructures are listed as the “map_find” solutions in the table above.

Third, the estimated solvent content may not be accurate, which can significantly affect the results of density modification. To test different values of the solvent content, click “Check Solvent” from the “Advanced Mode” dialog on the **Phase** subpage. The “Check Solvent Content” dialog (below) allows you to manually enter up to four different solvent contents (by default, HKL-3000 will fill in the original estimate in box 2). When the “Check” button is clicked, the program will run SHELXE four times, one for each solvent content value, and display the resulting map contrast by refinement cycle for each run. The solvent content that provides the map with the best contrast is likely to be closest to the true value



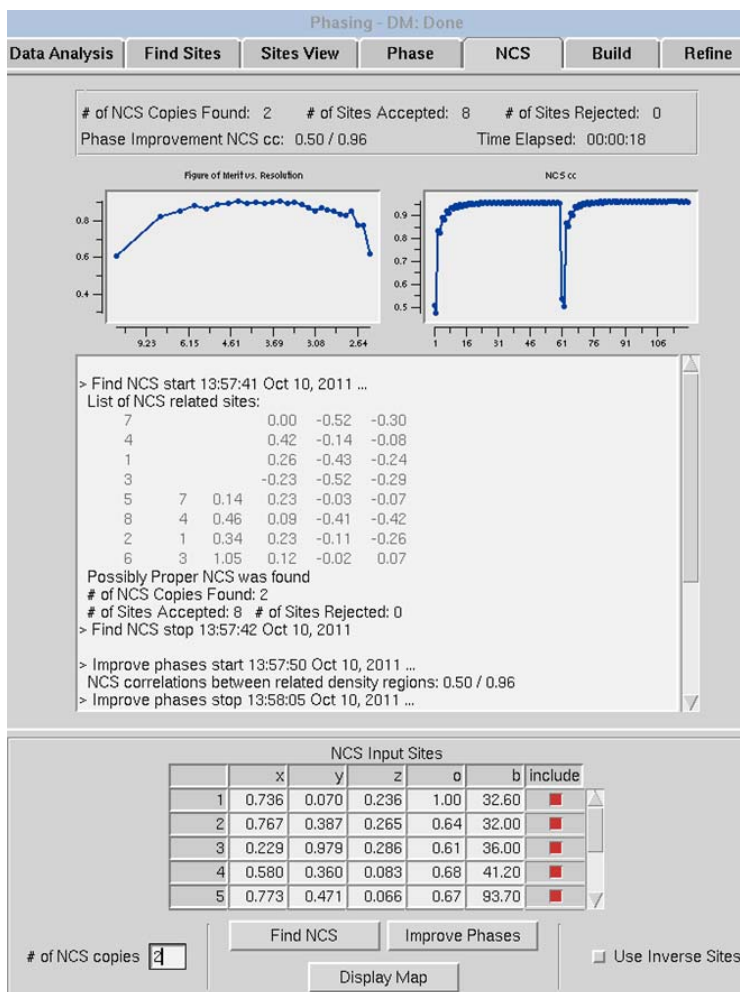
Fourth, you may change the “Phasing High Resolution Limit” value. The default value is suggested by HKL-3000 on the basis of the “Anomalous Signal to Noise vs. Resolution” analysis performed on the **Data Analysis** subpage. As mentioned above, if you change the phasing resolution limit MLPHARE will refine the substructure and will calculate phases only to the currently set limit. However, DM will perform a phase improvement together with a phase extension to the resolution which is set in the left panel of the main HKL-3000 window, regardless of the phasing resolution limit (see below). Using a smaller subset of low resolution reflections for phasing may be of benefit, as the anomalous differences for the lowest reflections are likely to be larger.



If you do not want to refine the substructure atoms, you may check the “Fast Mode” option at the bottom of the window. In this case, only the SHELXE and DM runs will be executed and the resulting density modified map will be displayed in COOT.

NCS

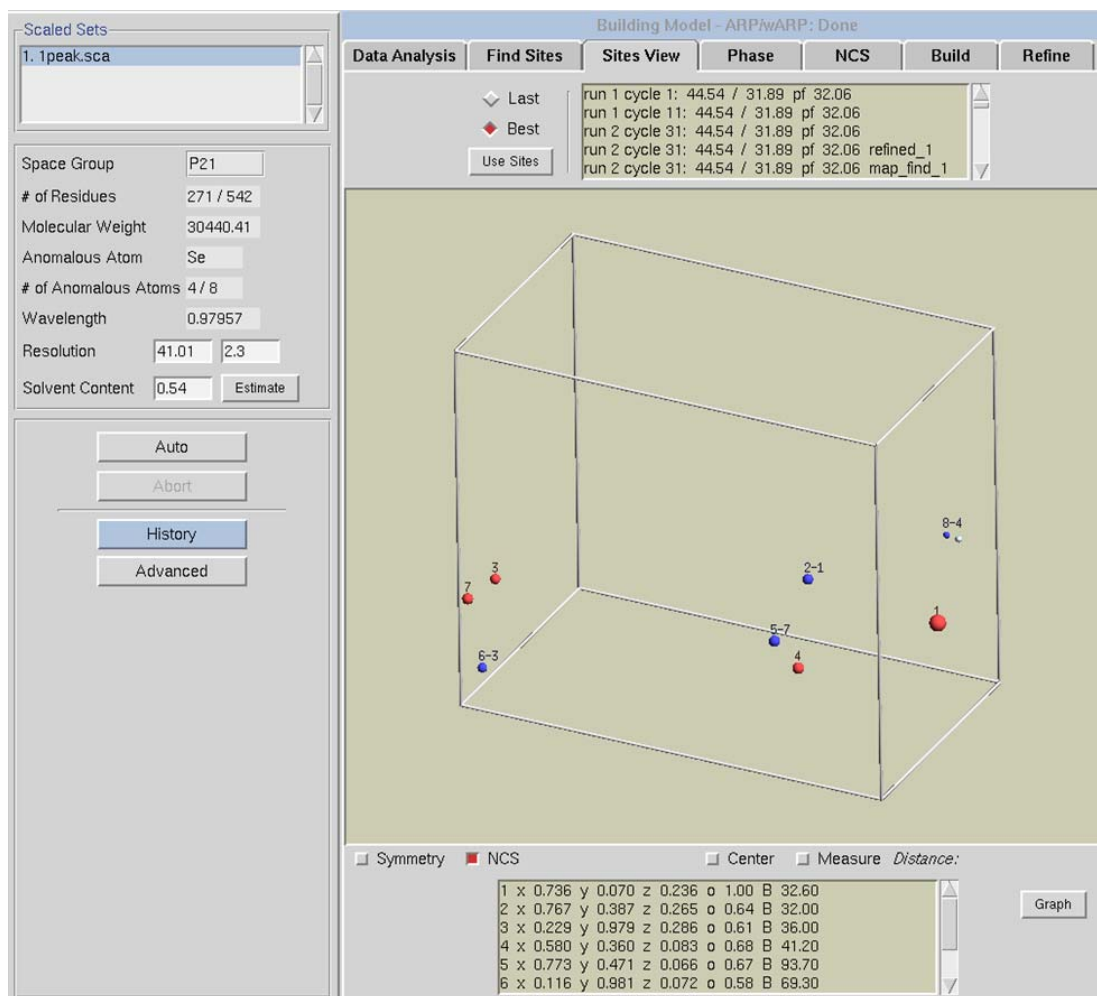
If you have more than one copy of the molecule in the asymmetric unit you may use non-crystallographic symmetry (NCS) for a phase improvement. HKL-3000 includes a subpage for automatically detecting NCS and calculating the appropriate operator(s), accessed by the **NCS** tab (see below). NCS operators are calculated with the PROFESS and RESOLVE programs after an analysis of the symmetry of the substructure sites. The substructure for NCS evaluation is listed as “NCS input sites” at the bottom of the **NCS** subpage, and other substructures may be selected through the **Sites View** subpage.



For the NCS calculations you may use all sites or only selected sites with high occupancy values—only the sites with a checked red square in the “include” column. If it is not clear how many copies of the protein are in the asymmetric unit you should test all possible numbers of NCS copies. The middle part of the NCS subpage shows the results of each NCS search: whether NCS with the specified number of copies was found, a list of the correlating sites, and the number of sites that were used and the number rejected.

NOTE: If the “Check Hand” procedure of the Phase subpage indicates that the sites in the currently selected substructure should be inverted, you should be certain to check “Use Inverse Sites” on the NCS subpage as well to avoid calculating incorrect NCS operators.

If NCS operators were successfully detected, you may go back to the **Sites View** subpage to review the results of the NCS search. The NCS related sites are colored and labeled: the sites in the first copy are shown in red, sites in the other copies are blue, and sites excluded from NCS detection are shown in grey.



Once you successfully detect NCS operators, click the “Improve Phases” button on the **NCS** subpage. HKL-3000 will start the DM program, which will use NCS averaging with the detected operators to further refine the experimental phases. Statistics from the DM runs are shown in two small plots at the top of the **NCS** subpage, with a plot of phase FOM as a function of resolution shell to the left, and a plot of the NCS correlation coefficient for each cycle of refinement. Following improvement of the phases via NCS averaging, HKL-3000 will spawn an instance of COOT that will display three maps: the anomalous density map (red), the electron density map after solvent flattening alone (blue), and the density map after further phase refinement by NCS averaging (light blue).

Parrot may also be used for NCS averaging (via the “Advanced Mode” dialog from the NCS subpage). Parrot is also capable of using a partial polypeptide model instead of a heavy atom substructure to detect NCS. If you know the NCS operator(s) from another source, you may enter them directly in the “NCS Operators” box on the “Advanced Mode” dialog. The rotation operators may be described by either rotation matrices or Euler angles, specified in these formats:

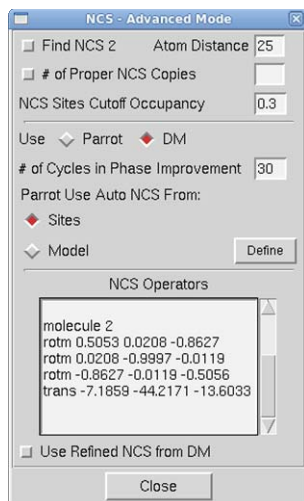
Rotation matrix:

```
molecule 2
rotm 0.5053 0.0208 -0.8627
rotm 0.0208 -0.9997 -0.0119
rotm -0.8627 -0.0119 -0.5056
trans -7.1859 -44.2171 -13.6033
```

Euler:

```
molecule 2
euler 61.48 177.41 -93.83
trans 34.26 156.80 2.64
```

Finally, when DM is used for NCS averaging, it refines the input NCS operators during the averaging process. If you want to use the improved NCS operators in subsequent refinement steps, you should check “Use Refined NCS from DM” in the “Advanced Mode” dialog (below).

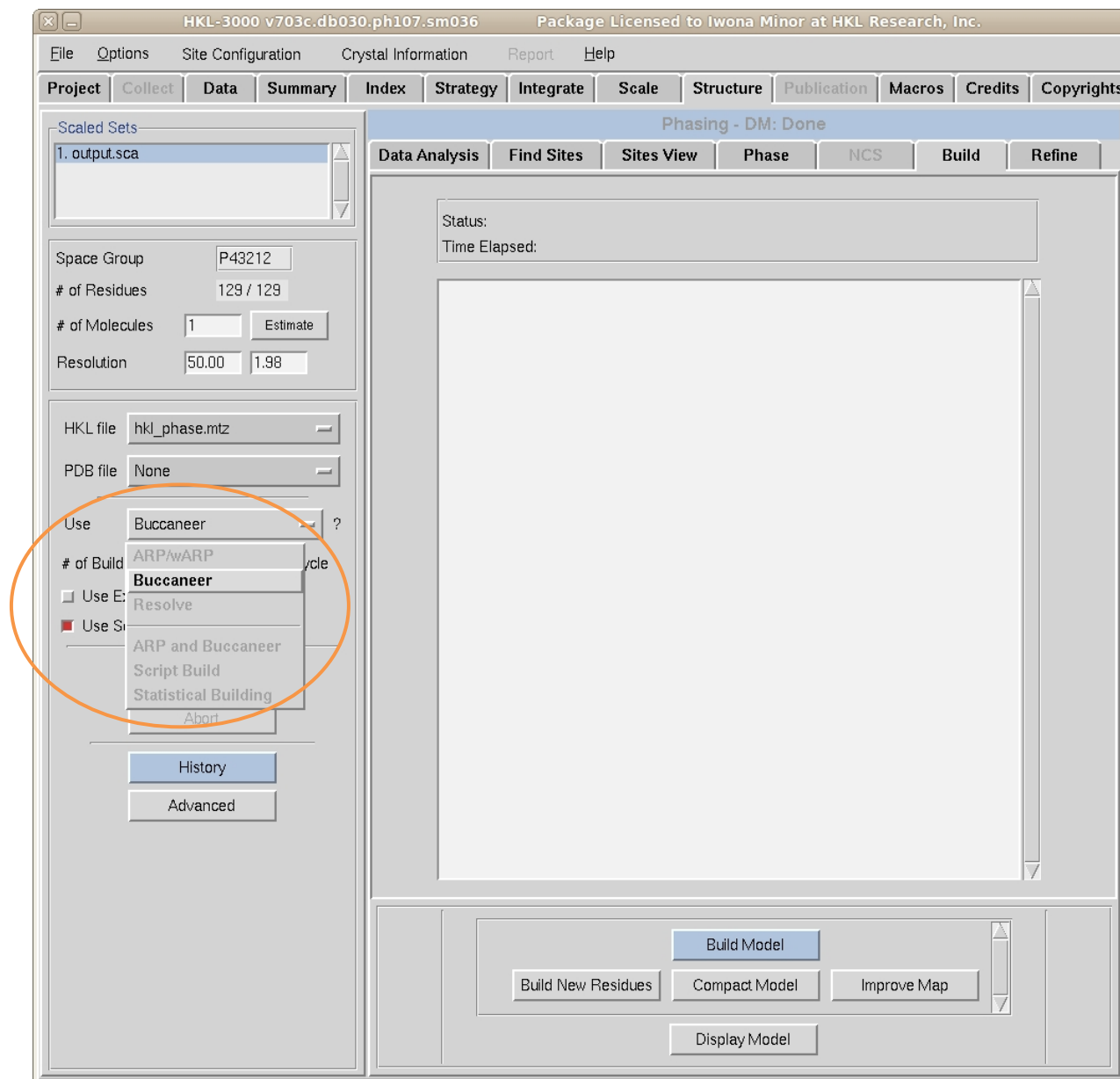


Model building

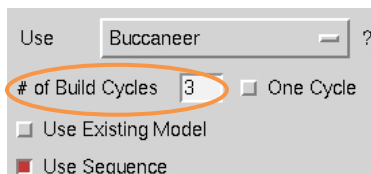
Once phasing is done and you have produced a reasonable electron density map, click the **Build** tab to start model building. Currently HKL-3000 includes options to use several programs (ARP/wARP, Buccaneer and Resolve) for model building, as well as some specialized program settings:

- “ARP and Buccaneer” uses alternating cycles of ARP/wARP and Buccaneer.
- “Script Build” uses Resolve and REFMAC in an iterative model-building procedure implemented as a CSH shell script (`resolve_build.csh`) by Tom Terwilliger. This procedure can take a few hours.
- “Statistical Building” runs ARP/wARP numerous times and uses statistical methods to generate a consensus combined model. This uses a script written by Zbyszek Otwinowski.

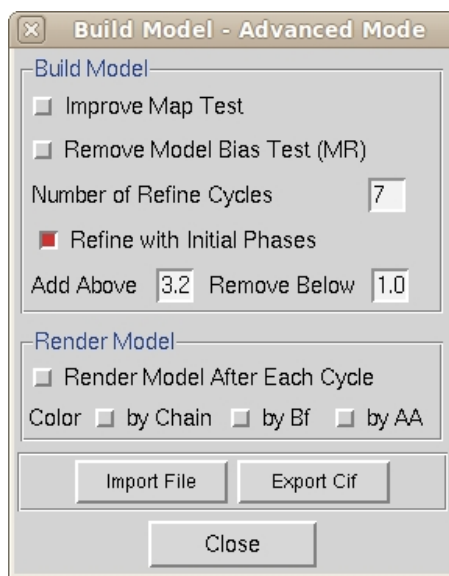
Dependent on resolution or your preferences you may choose any of these programs/procedures using the “Use” pulldown in the middle of the left sidebar as shown below.



During the first run of model building you will need a set of structure factors with phases (required) and a sequence (optional). Following successful phasing, there will be either 1 or 2 sets of structure factors—those with experimental MLPHARE phases refined by density modification (`hkl_phase.mtz`), and if applicable, those with experimental phases further improved by NCS (`hkl_phase_ncs.mtz`). The set of experimental amplitudes and phases is selected in the “HKL file” pulldown (the NCS phases if present, and the density-modified phases otherwise). As no coordinates exist prior to the first step of model building, the pulldown for “PDB file” will initially be empty. You may change the number of model building cycles in the main window (the “One Cycle” checkbox builds a preliminary model quickly to verify if the structure solution is correct).



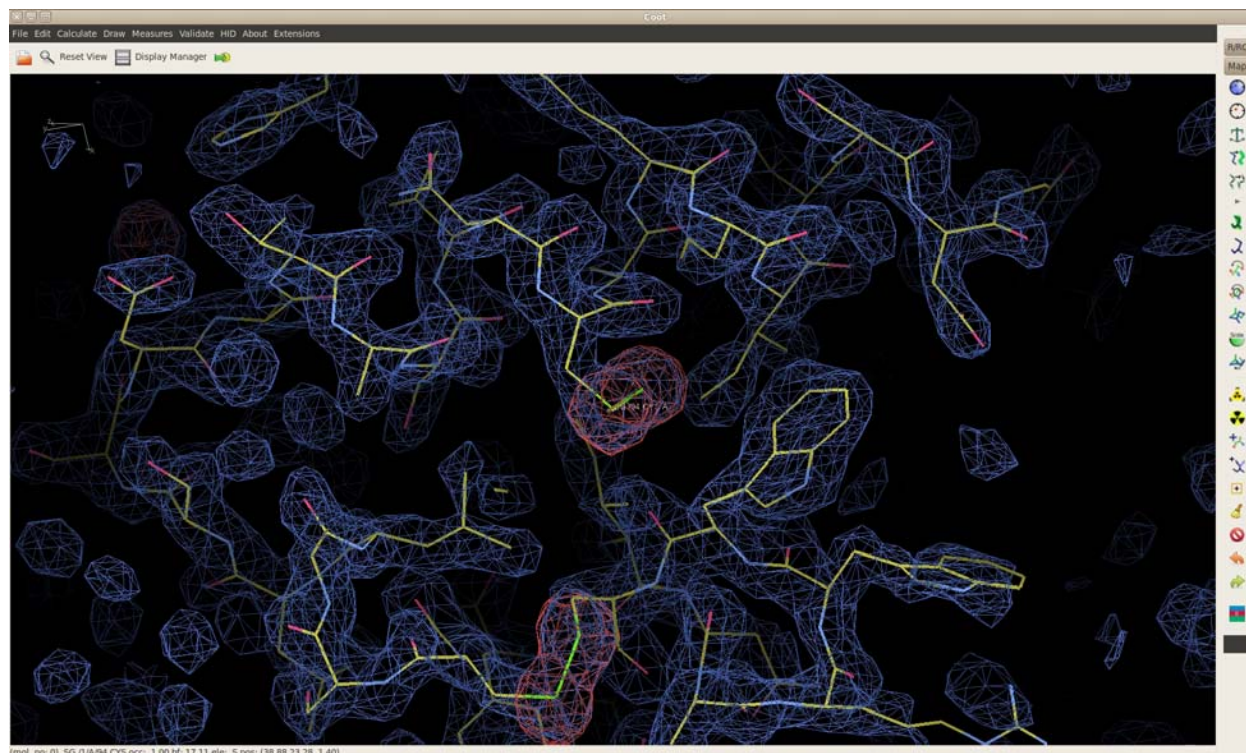
As with other subpages in the **Structure** page, you may modify the default parameters of the computations in the “Advanced Mode” dialog (below). For example, you may define the number of Refmac refinement cycles between building cycles (“Number of Refine Cycles”).



The “Improve Map Test” is only used when building a model with Resolve; this instructs the program to also refine the experimental phases during the model building process. The “Remove Model Bias Test (MR)” primarily applies to molecular replacement and is described in the section on the Molecular Replacement Pathway below. If you are choosing multiple cycles of model building you may check “Render Model After Each Cycle” to have an overview of the current model visualized in the PyMOL program.

After model building is completed, HKL-3000 starts Coot to display the current model, the substructure atoms, the $2F_o - F_c$, and experimental phase electron density maps and the $F_o - F_c$ difference map. In basic cases where the initial electron density map is of high quality, you may build an almost complete model. However, when you have low resolution data or there is some problem with phasing, the initial model will

be incomplete and have to be extended. You may view the current model using COOT, by selecting the `hkl_build_last.pdb` model from in the “PDB file” pulldown and clicking the “Display Model” button.



You may monitor the progress of the model building in the central window of the **Build** subpage. During the building cycles, it will show the number of residues and chains built, along with the number of residues docked into sequence. After the building cycles, the secondary structure elements that have been built and docked in the sequence will be shown graphically. During the refine cycles, the current values of R-factors and the mean figure of merit (FOM) at each step (below).

Building Model - Buccaneer

analysis Find Sites Sites View Phase NCS Build Re

Status:
Time Elapsed: 00:04:30

> Build model start 14:15:02 Sep 30, 2011 ...
Directory: build_model_37
Map in hkl_phase.mtz

Build main cycle: 1

Build cycle 1: 65 aa (50%) in 3 chains, 27 aa (21%/m) in the longest chain
48 aa have been docked (37%)
Build cycle 2: 96 aa (74%) in 6 chains, 27 aa (21%/m) in the longest chain
58 aa have been docked (45%)
Build cycle 3: 96 aa (74%) in 6 chains, 27 aa (21%/m) in the longest chain
58 aa have been docked (45%)
Refine cycle 0: R = 0.536 Rfree = 0.554 Fom = 0.610

> Secondary structure statistics

Chain A

KVFGRCLEAAAMKRHGLDNYRGYSLGNMCAAKFESINFNTQATNRNTDGS TDYGI LGIINS

RWWCNDGRTPGSRNLNIPCSALLSSDI TASVNCACKK I VSDGNGNNAWAWNRCKGTDV

QAWIRGCRL

Chain B

KVFGRCLEAAAMKRHGLDNYRGYSLGNMCAAKFESINFNTQATNRNTDGS TDYGI LGIINS

RWWCNDGRTPGSRNLNIPCSALLSSDI TASVNCACKK I VSDGNGNNAWAWNRCKGTDV

Refine cycle 1: R = 0.505 Rfree = 0.554 Fom = 0.629
Refine cycle 2: R = 0.494 Rfree = 0.543 Fom = 0.643
Refine cycle 3: R = 0.489 Rfree = 0.538 Fom = 0.646
Refine cycle 4: R = 0.487 Rfree = 0.543 Fom = 0.647

Building Model - Buccaneer: Done

analysis Find Sites Sites View Phase NCS Build Re

Status:
Time Elapsed: 00:22:27

Refine cycle 6: R = 0.309 Rfree = 0.373 Fom = 0.783
Refine cycle 7: R = 0.308 Rfree = 0.373 Fom = 0.782
Bond distances rmsd: 0.021, bond angles rmsd: 1.896

Build main cycle: 20

Build cycle 1: 131 aa (102%) in 1 chains, 131 aa (102%/m) in the longest chain
128 aa have been docked (99%)
Build cycle 2: 131 aa (102%) in 1 chains, 131 aa (102%/m) in the longest chain
128 aa have been docked (99%)
Refine cycle 0: R = 0.375 Rfree = 0.433 Fom = 0.736

> Secondary structure statistics

Chain A

KVFGRCLEAAAMKRHGLDNYRGYSLGNMCAAKFESINFNTQATNRNTDGS TDYGI LGIINS

RWWCNDGRTPGSRNLNIPCSALLSSDI TASVNCACKK I VSDGNGNNAWAWNRCKGTDV

QAWIRGCRL

Refine cycle 1: R = 0.344 Rfree = 0.409 Fom = 0.757
Refine cycle 2: R = 0.324 Rfree = 0.391 Fom = 0.771
Refine cycle 3: R = 0.317 Rfree = 0.384 Fom = 0.778
Refine cycle 4: R = 0.313 Rfree = 0.380 Fom = 0.781
Refine cycle 5: R = 0.311 Rfree = 0.377 Fom = 0.782
Refine cycle 6: R = 0.309 Rfree = 0.374 Fom = 0.783
Refine cycle 7: R = 0.308 Rfree = 0.374 Fom = 0.782
Bond distances rmsd: 0.021, bond angles rmsd: 1.896

Map out hkl_build_4.mtz, model out hkl_build_4.pdb
> Build model end 13:15:08 Sep 21, 2011

You may use the models from previous building runs as the starting point for subsequent building runs by checking “Use Existing Model” (shown below), and starting another build run by clicking “Build Model.” To use a file from a previous run, select the desired file in the “PDB file” pulldown. By default this will be the most recent model built. In this way, the output from one model building program (e.g. Resolve) can easily be passed as input to another (e.g. ARP/wARP).

Use ?

of Build Cycles ☐ One Cycle

☒ Use Existing Model

☐ Use Sequence

NOTE: the model building program will not use a starting model unless “Use Existing Model” is checked, even if there is a file selected in “PDB file”!

An alternative approach is manual correction of the model in COOT. After any build step, select the structure factor file and PDB file desired in “HKL file” and “PDB file,” respectively. (By default, these will be the files most recently generated.) Clicking “Display Model” at the bottom of the subpage will start an instance of Coot displaying the model and electron density maps ($2F_o - F_c$ and $F_o - F_c$) in the selected files selected. The anomalous density map and the heavy atom structure will also be displayed. After manually correcting and saving the model in Coot in the same directory, you will see the modified PDB file in the “PDB file” pulldown. By checking “Use Existing Model,” the manually corrected model can be used as an input for the next run of model building.

Other useful functions are accessed through other buttons:

Abort. Pressing “Abort” will cause HKL-3000 to immediately stop building. If at least one model was built, the most recent model built along with the most recent electron density maps will be displayed in Coot.

History. All files (*.pdb, *.mtz, etc.) created by the model building process are stored in a special session directory named `build_model_n` (where *n* is an integer), which is automatically created after starting a new building session for a given set of experimental phases. (See *Appendix A* for more details about how HKL-3000 manages files and directories.) By default, the list of files that appear in the “HKL file” and “PDB file” pulldowns are in the most recently created `build_model_n` session directory. However, if you close and restart the program, or if you go back and calculate new experimental phases on the **Phase** subpage, the next time you click “Build Model” on the **Build** subpage, a new `build_model_n` session directory will be created and the old models and structure factors from prior sessions will no longer be accessible in the file pulldowns. The “History” button allows you to return to a previous session. For example, if the current session is `build_model_17` but better models were built in the prior session `build_model_12`, clicking “History” and selecting the earlier session changes the contents of the pulldowns to list the files from the prior session directory.

Build new residues, Compact residues, and Improve map. These buttons invoke special functionality of Resolve to extend existing models, attempt to group all chains into the same asymmetric unit, and refine experimental phases while building, respectively. Please refer to the Resolve documentation for more details.

Refinement

After building an initial model, you may use the **Refine** subpage to do subsequent iterative refinement of your structure. REFMAC is used for maximum likelihood least squares model refinement. The fundamental procedure on the **Refine** subpage is to run REFMAC on the currently selected “HKL file” and “PDB file,” by clicking the blue “Refine Model” button. For all refinement steps executed, the “R, Fom vs. Cycle” plot at the top of the window will render a plot of R_{free} factor (orange), R factor (red) and mean figure of merit (blue) by cycle. Results from multiple runs are cumulatively added to the same plot, so the overall trend in statistics improvement may be seen over multiple sessions.

The left sidebar provides a number of options for controlling the functionality of REFMAC. By default, refinement is done in the most recent `build_model_n` session directory, rather than creating a new directory. If additional files are needed for the options listed below, they should be copied into this session directory. These options are “# of Refine Cycles” controls the number of cycles performed per “Refine Model” run.

If ligands or non-standard residues are present in the model that require custom restraint libraries, check “Use Ligand Library” and click “Define” to either explicitly choose a ligand restraint library (in *.cif format), or to generate new restraints from a ligand structure using LIBCHECK.

The “Use NCS” option controls whether NCS restraints are used during refinement. In “Auto” mode, the REFMAC will attempt to determine which residues are related by NCS automatically. In “Manual” mode, the ranges of residues related by NCS are manually specified with the “Define” button.

The “Use TLS” option controls whether anisotropic Translation-Libration-Screw motion refinement of rigid domains (or groups) of the model should be performed. To use TLS refinement, the rigid groups must be assigned using the “Define” button. The groups may either be specified manually, or a file containing TLS operators (*.tls) can be loaded.

Other parameters that control the operation of REFMAC may be adjusted through the “Advanced Mode” dialog, which are described in more detail in the REFMAC manual.

After cycles of maximum likelihood refinement, the model can be manually inspected and edited through Coot. The “Manual Model Build” button launches Coot with several models (the model in “PDB file” and the phasing substructure) and maps (the $2F_o-F_c$ and F_o-F_c maps in “HKL file”, the anomalous different map, and the initial experimental density-modified map). After manually rebuilding and saving the edited structure in Coot, the new file name is automatically selected in “PDB file.”

In addition to the REFMAC refinement performed by “Refine Model” and the manual rebuilding accessed by “Manual Model Build”, the **Refine** subpage provides a number of other tools to aid in structure refinement listed as a set of buttons. These tools include:

- *Sec Str Statistics*. This renders a table listing regions of the structure with identifiable secondary structure.
- *Add Waters*. This invokes a procedure in ARP/wARP to refine and extend the set of ordered waters.
- *Rotamer Refine*. This invokes a Coot procedure to identify residue side chains with improbable configurations and replace them with more probable rotamers.

- *Render Model.* This launches PyMOL to render the current model.
- *Check Waters.* This displays a dialog that shows a histogram of the distribution of water B-factors, which may be used to automatically remove waters with unrealistic B-factor values.

Molecular replacement (MR) pathway

Defining MR project parameters

To solve the structure of a protein by molecular replacement (MR), a project needs to be created by clicking “New Project” on the **Project** page, as described above for setting up a SAD/MAD project. The procedure for setting up a project for MR is identical save that “Molecular Replacement,” rather than “SAD/MAD,” should be selected as the “Phasing Method” (see below).

Each project represents a single macromolecule.
Each should contain only crystals of that macromolecule, but can include crystals with mutants, alternate tags, and ligands.

Edit Project lyso

Name: lyso

Description:

Tags: Edit Tags

Component: Protein Small Molecule Virus

Phasing Method: SAD/MAD **Molecular Replacement** Ligand Screen

Sequence: KVFGRCELAAMKRGHLDNYRGYSLGNWVCAAKFESNFNTQATNRNTDGS
TDYGILQINSRWWCNDGRTPGSRNLCNIPCSALLSSDITASVNCACKIVSDG
NGMINAWAWRNRCGTDVQAWIRGCR

Obtain Sequence from: NCBI id Download

Project Finished

Done Cancel

After setting up the project parameters by clicking “Done,” and selecting a Scalepack file on the **Project** page, you should go to the **Structure** page. Note that in MR phasing mode, the tabs shown on the **Structure** page are different. Specifically, these subpages are **Analyze Data**, **Prepare Model**, **Run MR**, **Rebuild Model**, and **Refine Model**.

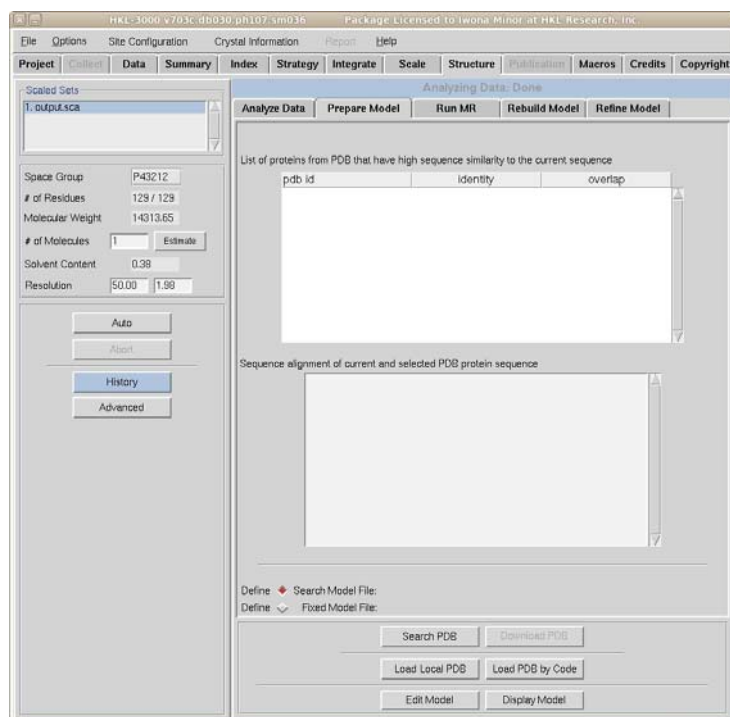
Data analysis

The **Analyze Data** subpage in MR phasing mode is very similar to the **Data Analysis** subpage in SAD/MAD phasing mode. The plots generated in the SAD/MAD case are generated here as well (please refer to the *SAD/MAD pathway* section of the manual for more details about these plots. There are two small differences. First, the plots that are only relevant to SAD/MAD data are omitted (“Anomalous Signal to Noise vs. Resolution” and “Wavelength Pairs”). Second, the completeness statistics do not need to be calculated separately from the twinning statistics; clicking the “Analyze Data” button will populate all five charts.

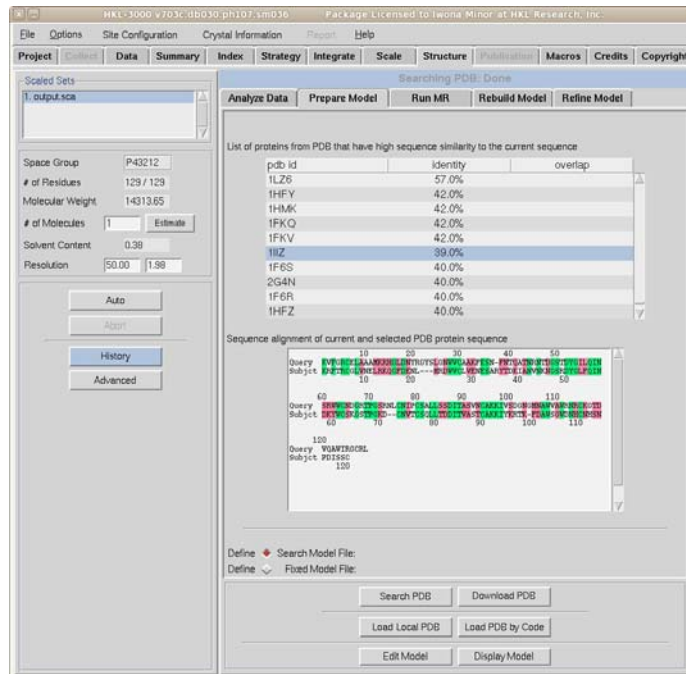
Similarly, as in the SAD/MAD case the “Estimate” button on the left sidebar is used to calculate the most probable solvent content. The solvent content estimation dialog works identically in either MR or SAD/MAD phasing mode.

Preparing the search model

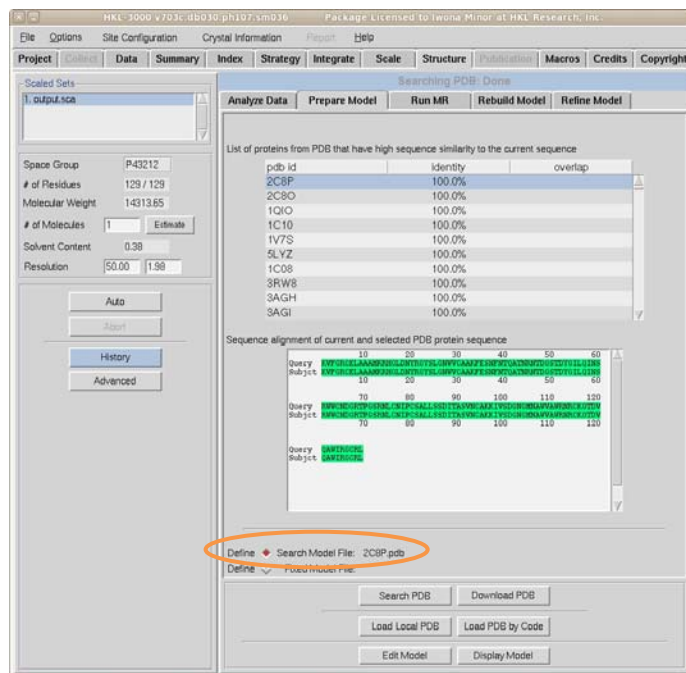
The next step is preparing a proper search model in the **Prepare Model** subpage.



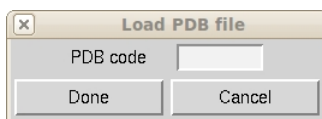
You may find search models in multiple ways. First, HKL-3000 can search the PDB for an appropriate structure, by clicking the “Search PDB” button. HKL-3000 will look for structures in the PDB with 30% or greater sequence identity to the sequence of your protein. If matching PDB structures are found, the top part of the subpage will display a list of PDB structures and the percentage of sequence identity to your sequence for each one. Underneath the list of PDB hits, the second window displays a sequence alignment of your protein to the currently selected PDB protein (highlighted in blue in the list). Identical residues are marked in green, and similar residues in red (see below).



The model chosen should have a high degree of identity. To automatically obtain the model for a particular structure in the list, select the structure and click the “Download PDB” button. A PDB file will appear at the bottom of the subpage (see below).

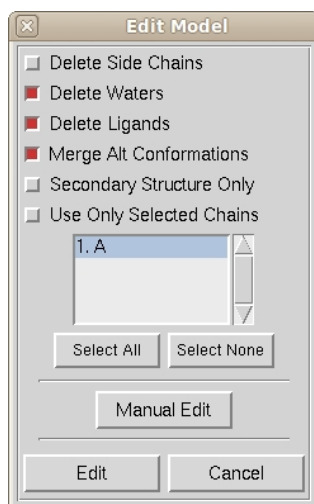


You may also obtain a specific PDB structure by clicking “Load PDB by Code” and entering its four letter accession code.





Alternatively, if you already have the PDB file for your search model, you can load in the structure file by clicking the “Load Local PDB” button and selecting the file.

Refined structures, such as those from the PDB, may not serve as good search models without some preparation. For example, refined structures usually contain water molecules and ligands, which might be missing, modified and/or located in different places in your structure. The “Edit Model” button opens a dialog that presents a variety of options for transforming the search model. Choose the changes you wish to make by selecting the appropriate checkboxes and click “Edit” to modify the model.



Alternately, you may click the “Manual Edit” button, which will launch a text editor to edit the PDB file data directly. If an edited search model is produced with either the “Manual Edit” or the “Edit” buttons, the coordinates will be copied to a new file with a name automatically generated by HKL-3000 (see below)

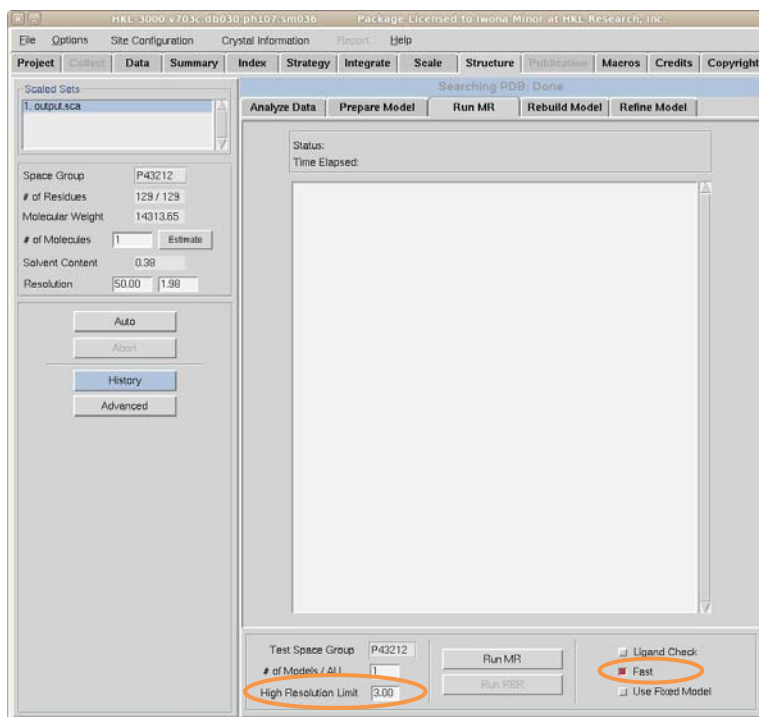
```
Define  Search Model File: hkl_mr_edit_1.pdb
Define  Fixed Model File:
```

Clicking the “Display Model” button launches the Coot program with the current model loaded, if you prefer to edit the search model using graphical tools. After editing the model, save the modified model coordinates with a different file name. Note that HKL-3000 is not able to automatically detect the name of the new file you created in Coot; you will have to manually upload it to HKL-3000 using the “Load Local PDB” button.

Note that one of two different types of models may be defined on the Prepare Model page, either a “Search Model File,” or a “Fixed Model File.” By default, “Search Model File” is chosen, which indicates that a full translation and rotation search for the best orientation of the search model will be performed. In contrast, marking the model as a “Fixed Model File” will omit the translation and rotation search, and the model will remain in the same orientation as in the initial file. This is useful in the process of fitting two or more different search models into the same asymmetric unit, described in more detail below

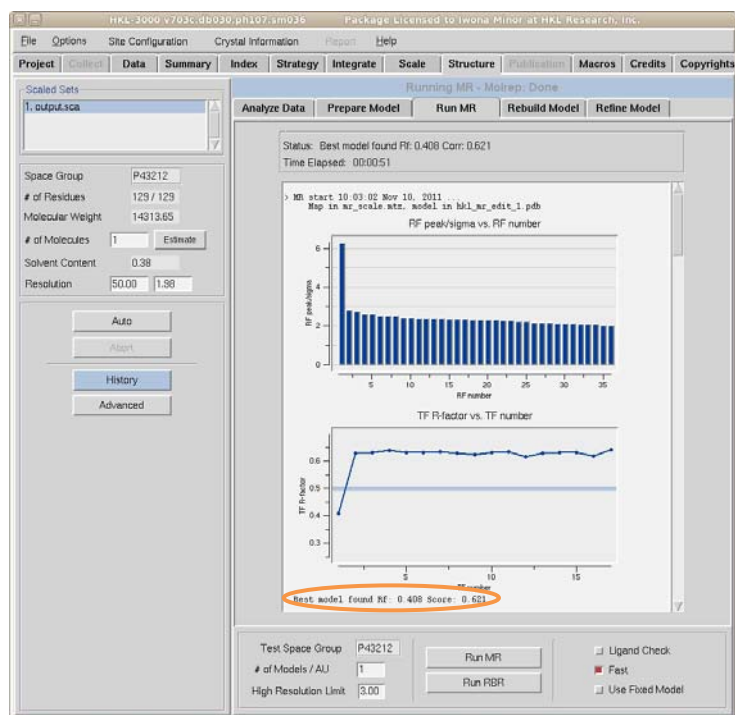
Running MR

After preparing the search model, the next step is accessed via the **Run MR** tab. By default, HKL-3000 uses the MOLREP program for molecular replacement, with the “Fast” option selected. (Selecting “Fast” indicates that MOLREP should be run in “fast mode”, and conversely, deselecting the option will run “slow mode”). The value in the “High Resolution Limit” box determines the high resolution limit of the set of reflections used to find a molecular replacement solution. When the “Run MR” button is clicked, a full rotation and translation search for the input model is performed. When the “Run RBR” (rigid body refinement) button is clicked, the rotation and translation search is omitted. This mode is for situations where the search model is largely isomorphous to the data being phased (for example, using the *apo*-protein structure to solve the structure of the same protein soaked in or co-crystallized with ligand).

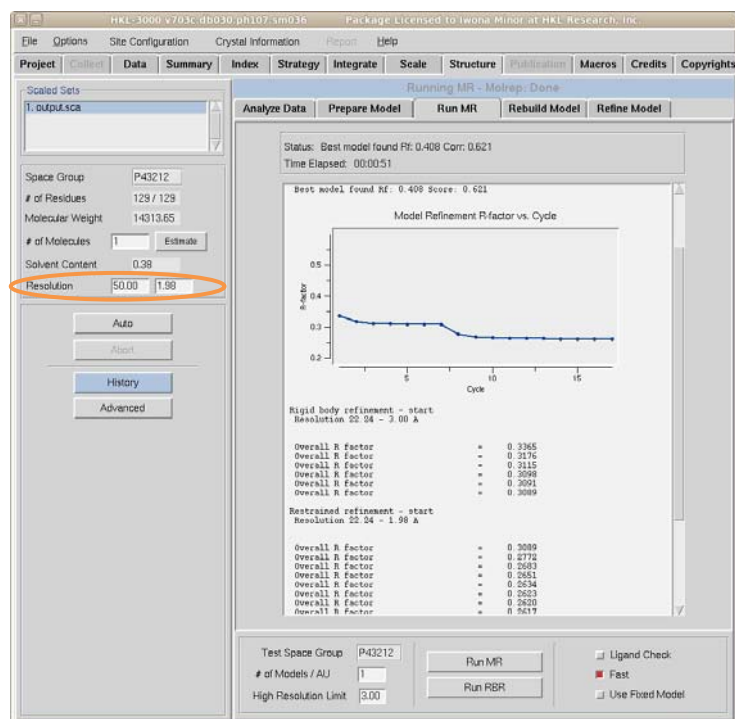


After MOLREP is run by clicking the “Run MR” button, two plots showing the results of the cross rotational function (RF) and translational function (TF) calculations are shown. The “RF peak/sigma vs. RF number” plot shows the ratio of R-factor to σ for the largest peaks in the cross RF (greater is better). The “TF R-factor vs. TF number” plot shows the R-factors for the best TF solutions (lesser is better). The lower plot also contains a pale blue line at an R-factor of 0.5 to indicate that TF solutions lesser than this value are likely to be significant. Generally, good MR solutions should be significantly better than the other solutions.

When the MR procedure succeeds, MOLREP will also print statistics for the best overall solution: the overall R-factor (Rf) and a score (consult the MOLREP documentation for more details).



Following identification of the correct rotation and translation, the search model is refined with REFMAC against the experimental structure factor amplitudes. Rigid body refinement is used first using data up to the MR resolution limit. Next, the model is subjected to restrained refinement using all available data (i.e. all the way to the high resolution limit shown in the left sidebar; see below).



A plot of the R-factors as a function of REFMAC refinement cycle is displayed (shown above). After refinement, Coot is automatically launched, and the modified model, along with the calculated $2F_o - F_c$ electron density and $F_o - F_c$ difference maps.

Rebuilding and refining MR models

After successful MR phasing, if large sections of the electron density need to be built, the same programs and tools for building initial models after SAD/MAD phasing (ARP/wARP, Resolve and Buccaneer) may also be used to rebuild search models in MR solutions. To this end, the functionality of the **Rebuild Model** subpage in MR mode is identical to that of the **Build** subpage in SAD/MAD mode. Please consult the *Model building* section above for more details.

Similarly, once the initial search model has been rebuilt to match the new structure factor data, the **Refine Model** subpage may be used to refine the structure using REFMAC. As in the case of model building, the functionality of the **Refine Model** subpage in MR mode is identical to that of the **Refine** subpage in SAD/MAD mode.

MR of heteromeric complexes: using multiple search models

HKL-3000 may also be used for more difficult MR problems, such as solving the structure of heteromeric complexes. For large heteromers, it may be necessary to use more than one search model. When working with a heteromeric complex, special care must be set the appropriate sequence set in the **Project** page. Currently, HKL-3000 cannot separately track multiple polypeptide sequences for the same project. Until this feature is implemented, as a workaround the sequences for the different polypeptides can be concatenated together and added to the sequence box as one large combined sequence. This concatenation should take stoichiometry into account. For example, consider a crystal of the hypothetical complex A_2B , which comprises two subunits of polypeptide A and one subunit of polypeptide B. The combined sequence of A_2B should contain two copies of the sequence of A and one copy of the sequence of B (see below).

Project | Collect | Data | Summary | Index | Strategy | Integrate | Scale | Structure | Publication | Macros | Credits | Copyrights

Project Name: project_heterotrimer
Crystal: crystal1
Experimenter: anna
Date: Dec 16, 2011

Buttons: Load, Save, Edit Project, Edit Crystal, New Project, New Crystal, Evaluate Model

Principal Component: Protein

Protein Data

Name:
Number of Residues: 669
Organism:
Description:
Molecular Weight: 75542.98
NCBI accession code:
Swiss-Prot accession code:
No. of Homologues:
Last check in PDB:

Phasing Method: Molecular Replacement

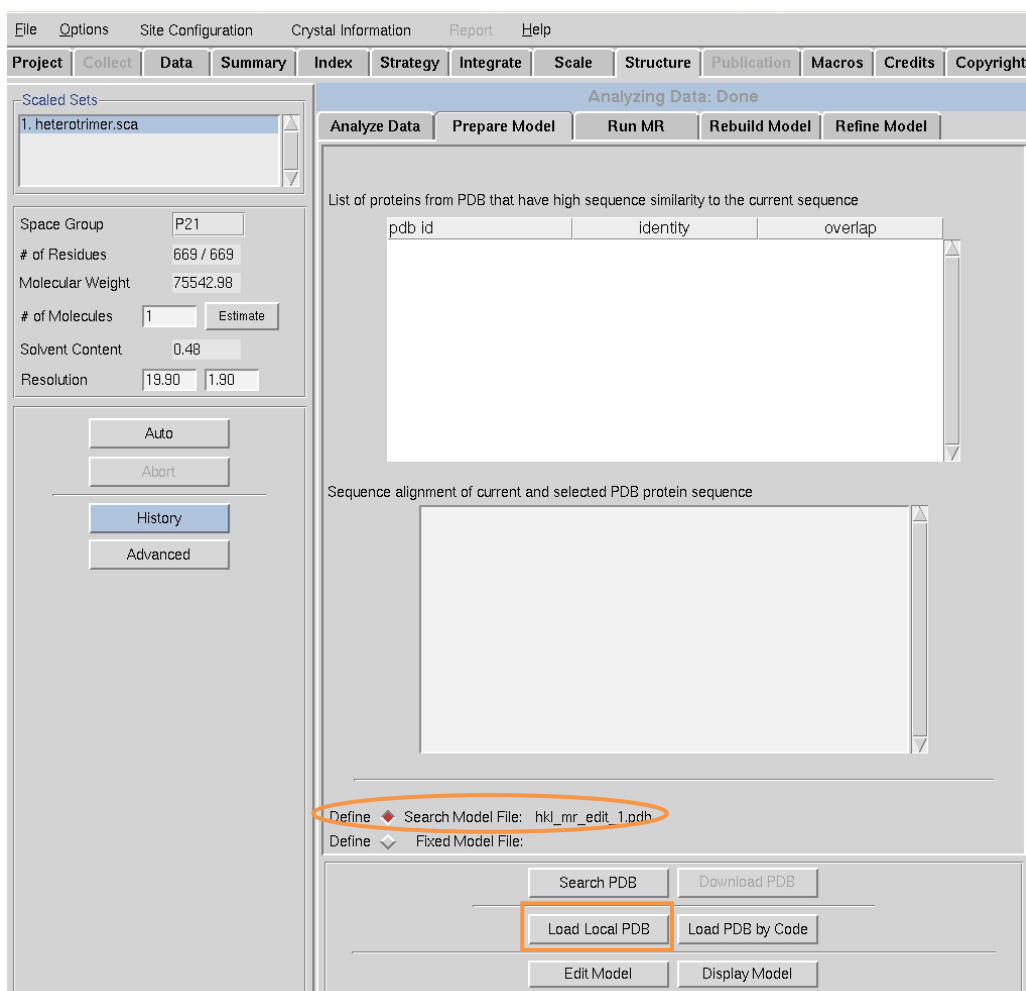
Ala (A): 42 Gly (G): 57 Met (M): 12 Ser (S): 48
Cys (C): 21 His (H): 21 Asn (N): 36 Thr (T): 39
Asp (D): 36 Ile (I): 54 Pro (P): 27 Val (V): 45
Glu (E): 30 Lys (K): 6 Gln (Q): 48 Trp (W): 12
Phe (F): 9 Leu (L): 30 Arg (R): 45 Tyr (Y): 51

Scaled Data

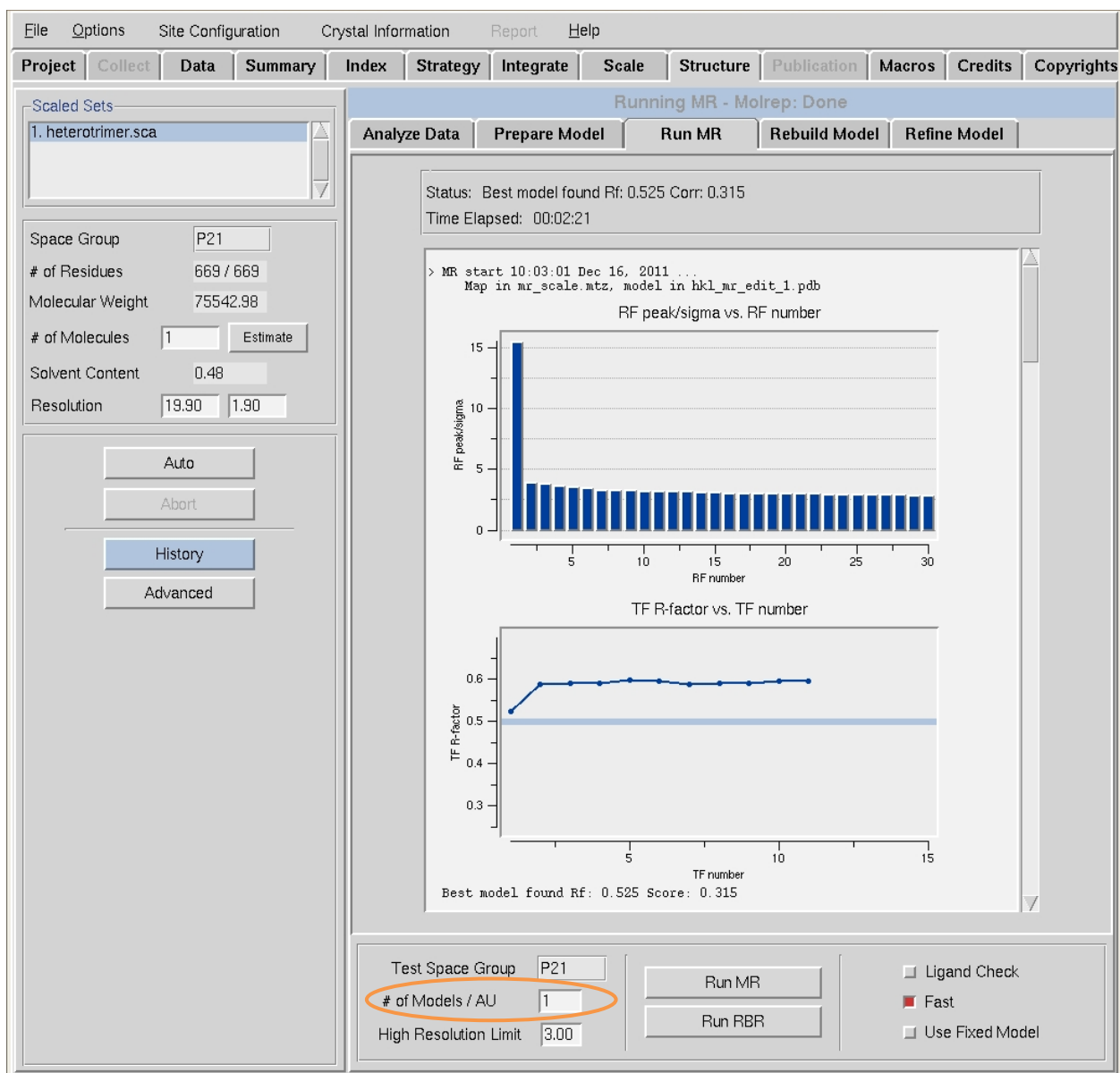
Structure File Name: /home/anna/data/MR_heterotrimer/heterotrimer.sca
Cell: 65.403 79.428 72.199 90.000 105.201 90.000
Space Group: P21 Wavelength: 0.97980 Mode: Peak Order: 1

Once the combined sequence is defined, you may begin structure determination on the **Structure** page. As always, the first step is analysis of the diffraction data and the estimation of number of macromolecules in the asymmetric unit. Note that HKL-3000 considers the sequence listed in the Project page to be a single macromolecule, and the value of “# of Molecules” should reflect that fact. Continuing the example from above, if there are 2 copies of the A₂B complex in the asymmetric unit, even though there are technically 6 macromolecules (4 subunits of A and 2 subunits of B) present, the value of “# of Molecules” should be set to 2.

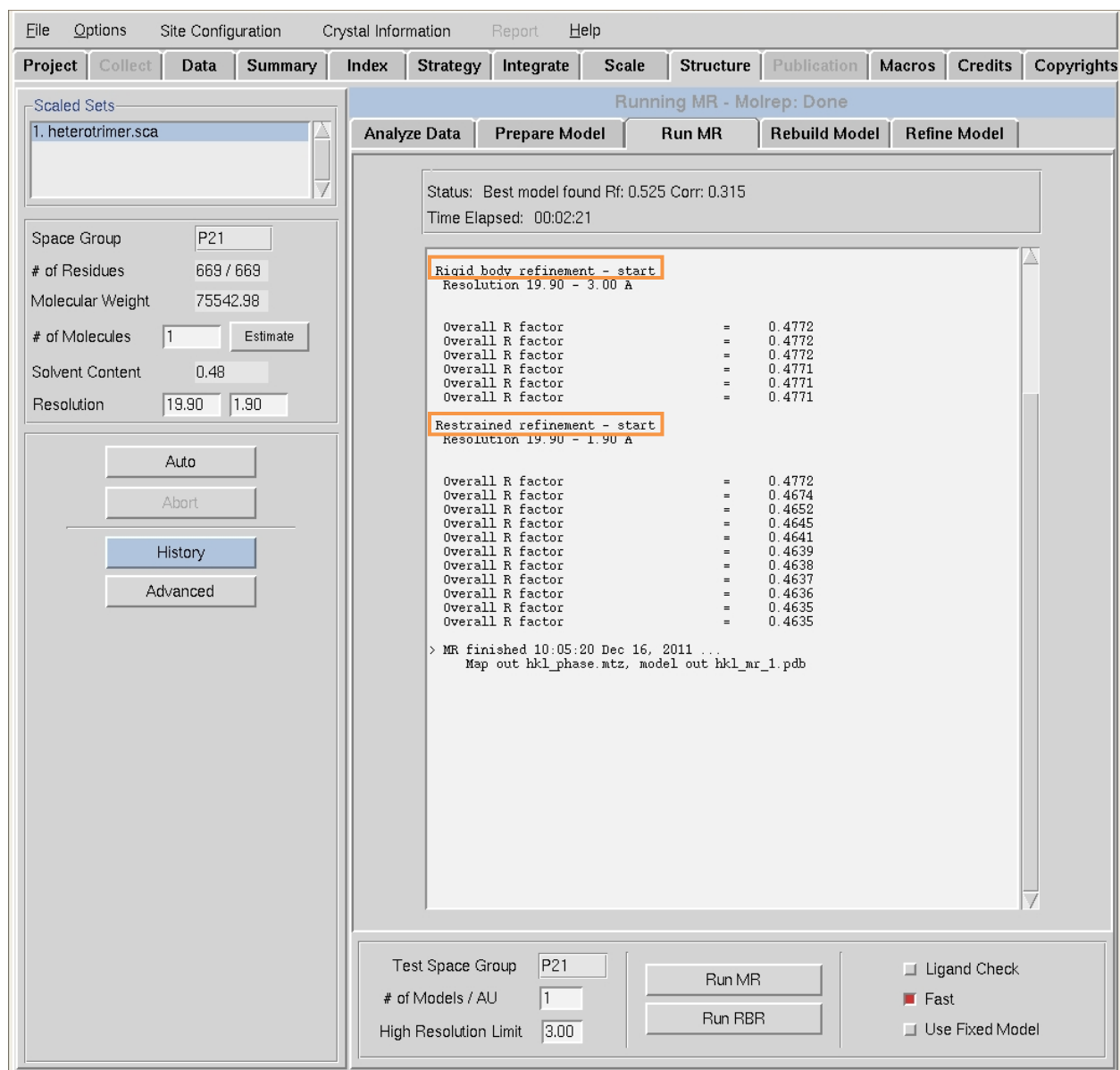
While working on a heteromeric complex you will have to obtain at least two different search models, which will be loaded one at a time. In the example shown below, the first search model is loaded from a local PDB file and edited.



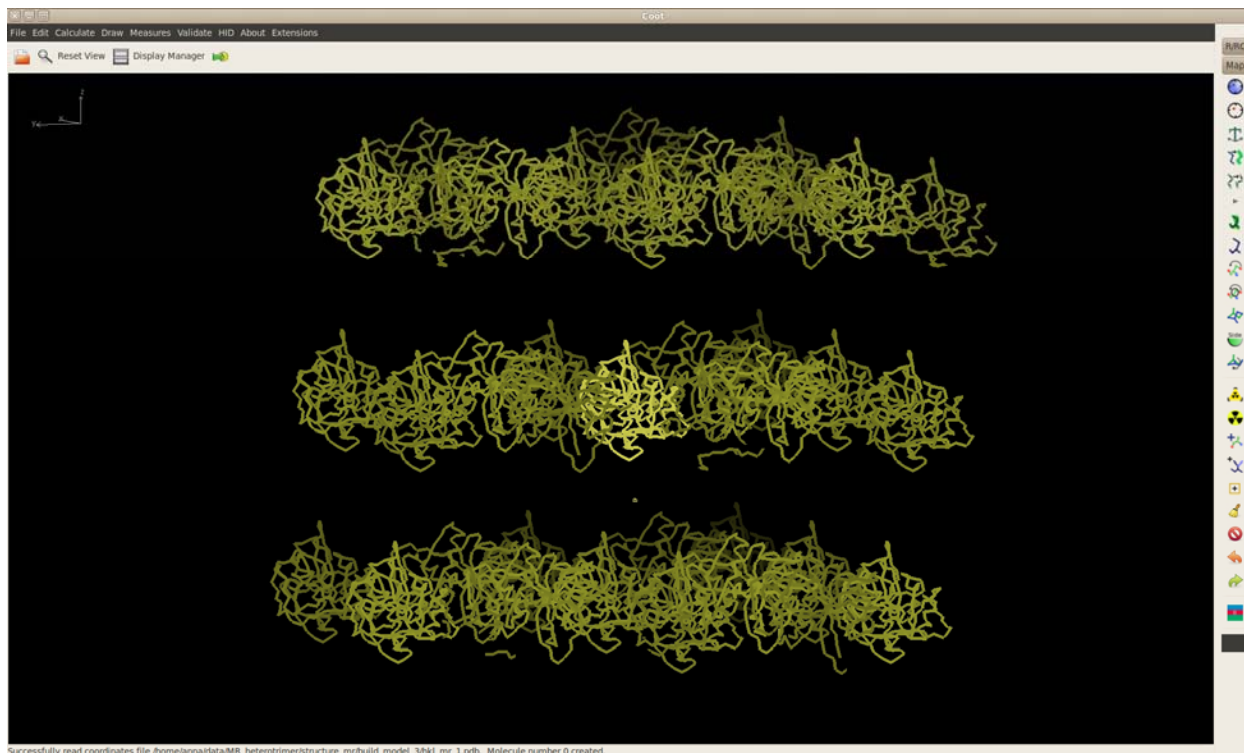
When the first search model is prepared, go to the **Run MR** tab to search for MR solutions for this model.



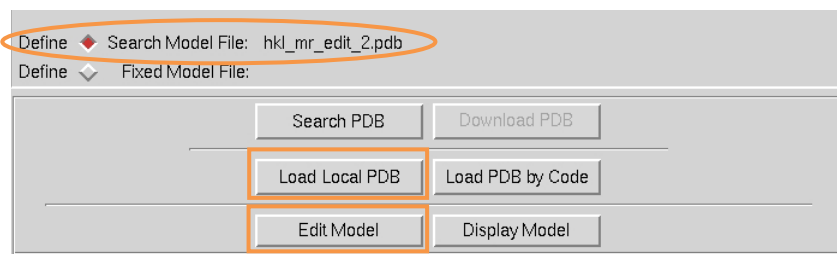
As before, after the MR search procedure, HKL-3000 will subsequently do rigid body and restrained refinement on the search model.



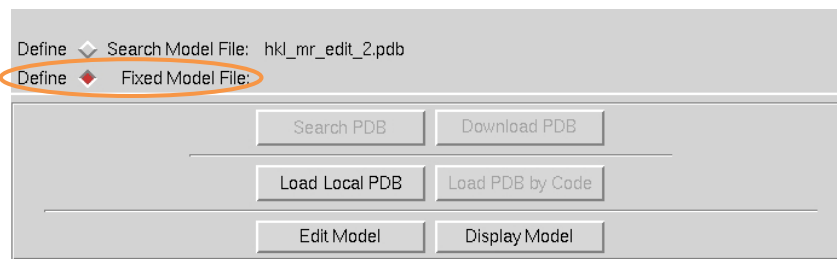
With luck, the MR procedure for the first search model will be at least partially successful, and a partial model of the complex will be displayed.



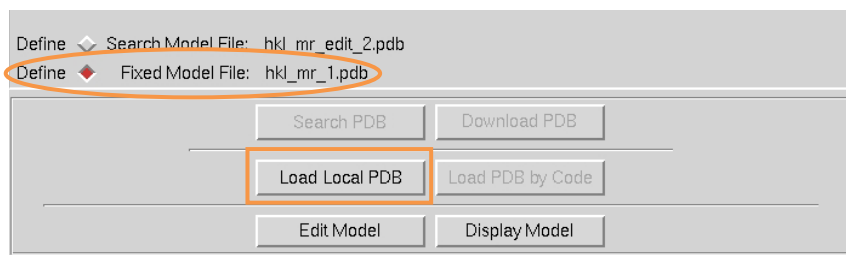
If the solution for the first search model appears to be correct, you may continue the MR process by searching for a second model. The second search model may be loaded in through the **Prepare Model** subpage as before. In the example below, the second search model was loaded from a local file and then edited.



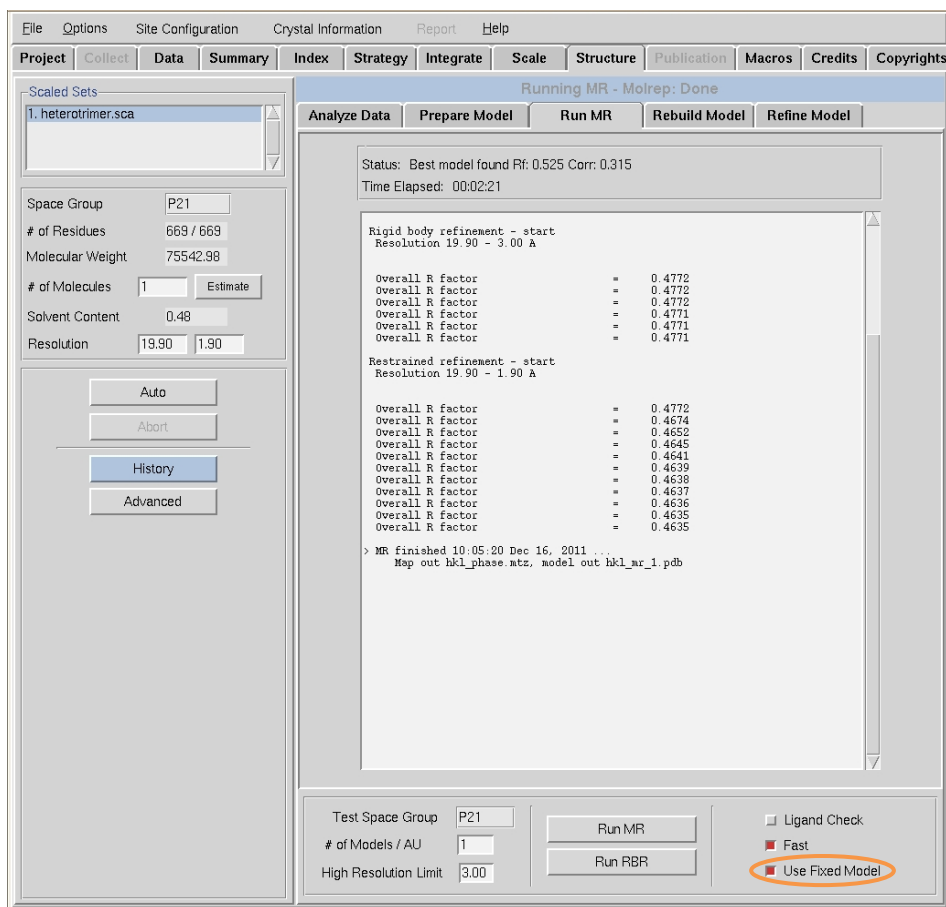
However, the difference in preparing the second search model is that the previous search model(s) must also be included. As the optimal translation and rotation for the first search model was already been determined, its position can be fixed. To add the previous, refined search model and keep it fixed in orientation, change the toggle from “Search Model File” to “Fixed Model File”.



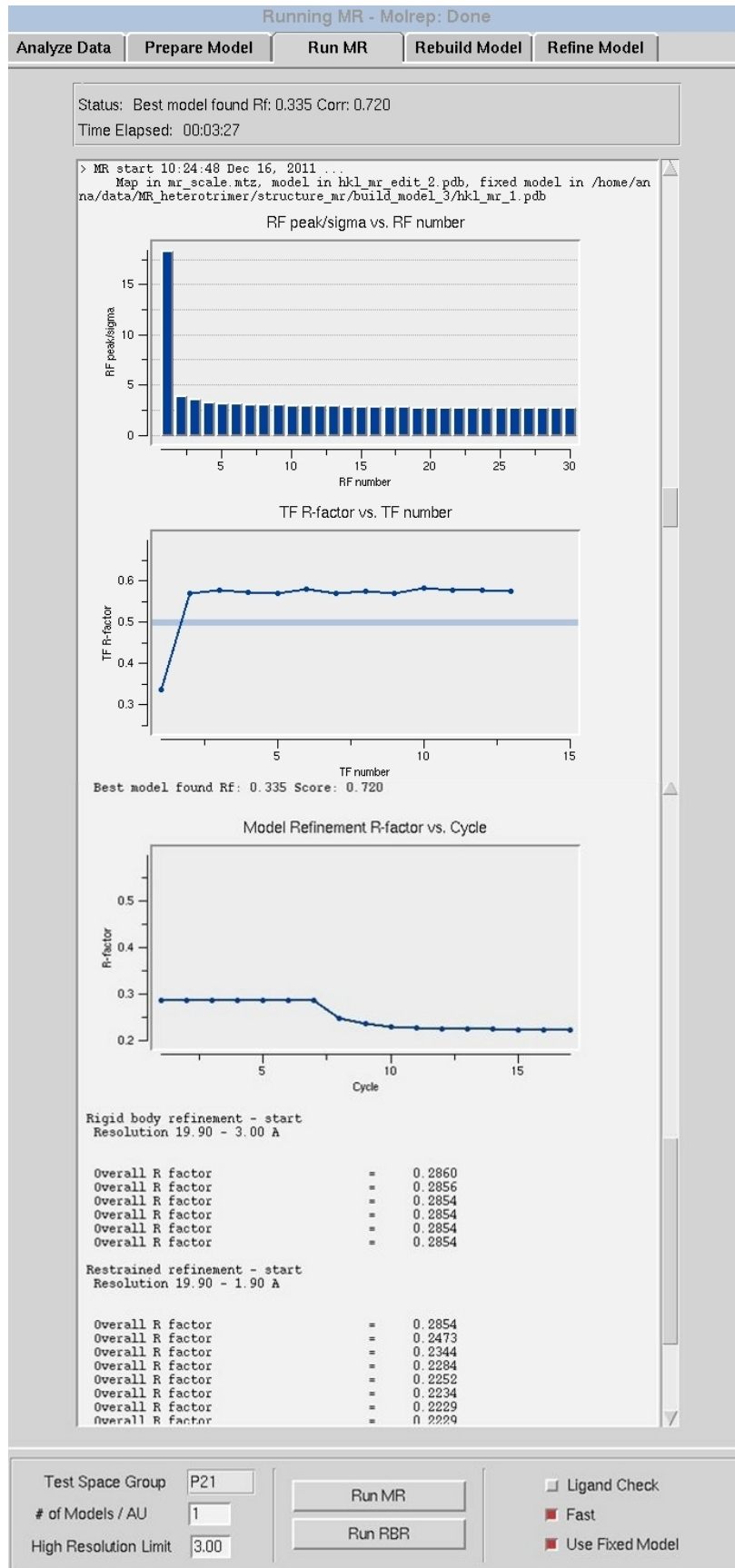
Once this box is checked, you can load the file with the previous solution into HKL-3000 via the “Load Local PDB” button. Essentially, there are “slots” for two models, one used in the rotation/translation search, and a second kept in its original, fixed orientation. The “Load Local PDB” (and “Download PDB” and “Load PDB by Code”) will load files into whichever slot is selected.



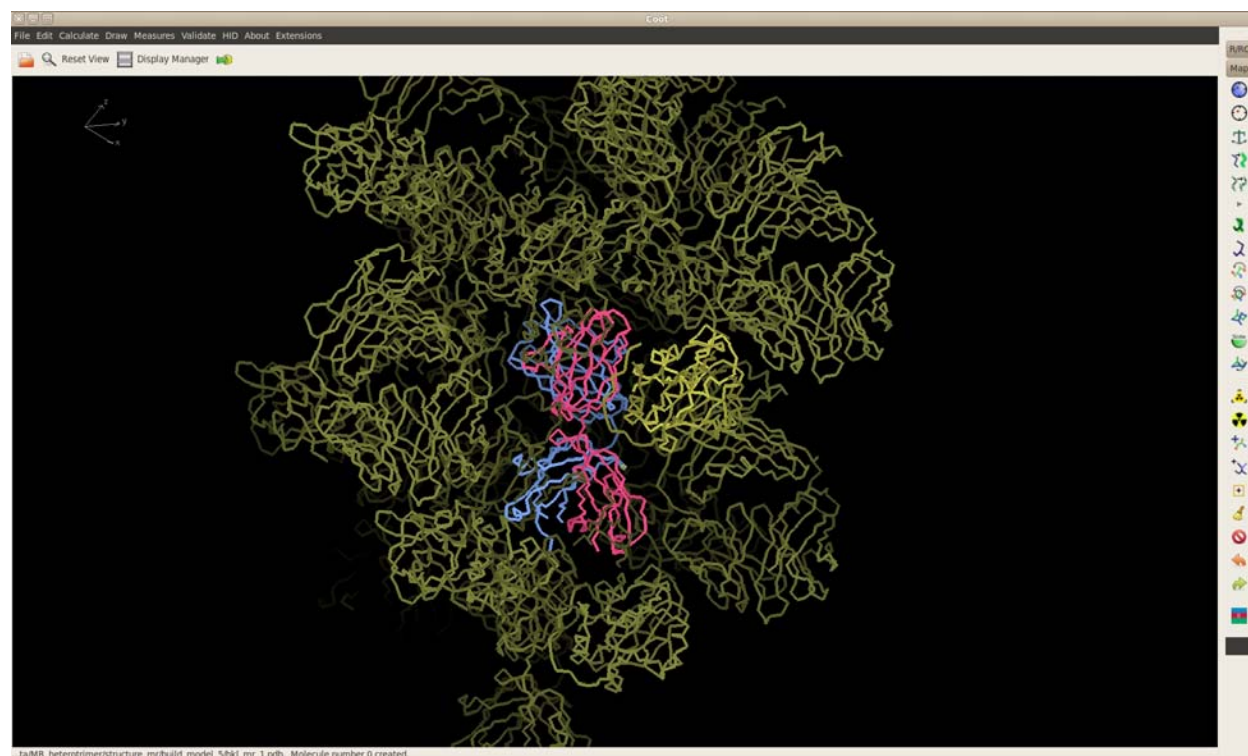
When both the new search model and the prior fixed solution are prepared, you may proceed to the **Run MR** tab. Note: the fixed model specified on the Prepare Model subpage will not be used unless the “Use Fixed Model” option is also selected (see below).



With this option set, click “Run MR” again.



Once the MR procedure is finished, HKL-3000 will launch Coot, displaying the best solutions for both models, as well as the calculated electron density. At this point, you may investigate if the packing of the heteromeric complex and density maps are reasonable.



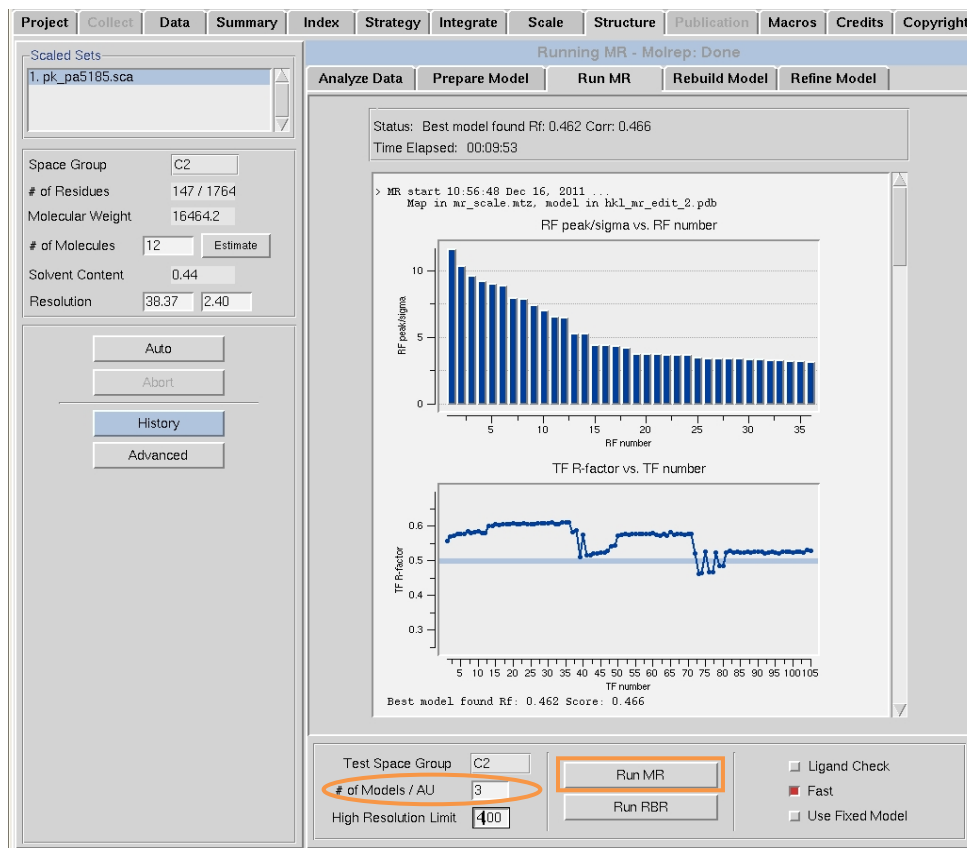
MR of homomeric complexes

For homo-oligomeric complexes, the sequence of only one polypeptide chain needs to be specified (see below).

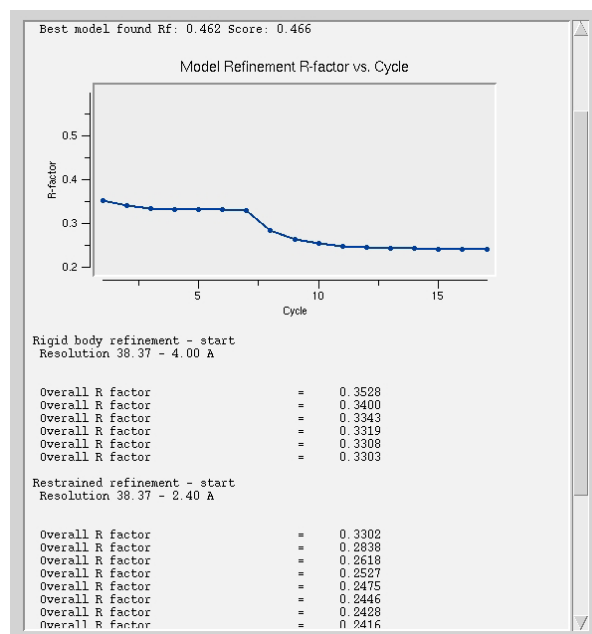
Project	Collect	Data	Summary	Index	Strategy	Integrate	Scale	Structure	Publication	Macros	Credits	Copyrights																				
<div> <div> Project Name: manycopies Crystal: crystal1 Experimenter: anna Date: Dec 16, 2011 <div> Load Save </div> <div> Edit Project Edit Crystal </div> <div> New Project New Crystal </div> <div> Evaluate Model </div> </div> <div> Principal Component: Protein Protein Data Name: Number of Residues: 147 Organism: Description: Molecular Weight: 16464.2 NCBI accession code: Swiss-Prot accession code: No. of Homologues: Last check in PDB: </div> </div>																																
<div> <div> MATAPRPLREGQYLHFQPISTRWHDNDVGHVNNVTYYAFDFA VNTYLIERGGLDIOGGGEVIGLVSSSCDVFAPVAFQRIEVLRL YARLGNSSVGYELALFLLEGQREACAAGRIVHVFERRSSRPVA IPQELRDALAALOSSAQ </div> <div> Phasing Method: Molecular Replacement <table> <tr> <td>Ala (A): 16</td> <td>Gly (G): 10</td> <td>Met (M): 2</td> <td>Ser (S): 10</td> </tr> <tr> <td>Cys (C): 2</td> <td>His (H): 4</td> <td>Asn (N): 5</td> <td>Thr (T): 5</td> </tr> <tr> <td>Asp (D): 6</td> <td>Ile (I): 7</td> <td>Pro (P): 7</td> <td>Val (V): 13</td> </tr> <tr> <td>Glu (E): 9</td> <td>Lys (K): 0</td> <td>Gln (Q): 9</td> <td>Trp (W): 1</td> </tr> <tr> <td>Phe (F): 8</td> <td>Leu (L): 13</td> <td>Arg (R): 13</td> <td>Tyr (Y): 7</td> </tr> </table> </div> </div>													Ala (A): 16	Gly (G): 10	Met (M): 2	Ser (S): 10	Cys (C): 2	His (H): 4	Asn (N): 5	Thr (T): 5	Asp (D): 6	Ile (I): 7	Pro (P): 7	Val (V): 13	Glu (E): 9	Lys (K): 0	Gln (Q): 9	Trp (W): 1	Phe (F): 8	Leu (L): 13	Arg (R): 13	Tyr (Y): 7
Ala (A): 16	Gly (G): 10	Met (M): 2	Ser (S): 10																													
Cys (C): 2	His (H): 4	Asn (N): 5	Thr (T): 5																													
Asp (D): 6	Ile (I): 7	Pro (P): 7	Val (V): 13																													
Glu (E): 9	Lys (K): 0	Gln (Q): 9	Trp (W): 1																													
Phe (F): 8	Leu (L): 13	Arg (R): 13	Tyr (Y): 7																													

After data analysis, prepare a search model for MR as before

However, before running the MR procedure proper, some parameters of the MR process need to be adjusted, using information about the oligomerization state of the macromolecule. For example, in the figure below, the protein crystallized forms homotetramers, and by creative adjustment of the “# Models / AU” parameter, the tetramer may be used as the search model. The solvent content analysis suggests that the asymmetric unit probably contains 12 protein molecules, which corresponds to 3 tetrameric assemblies. To search for 3 assemblies, you have to update value of the “# Models / AU” box and then press “Run MR.”



As before, after molecular replacement HKL-3000 will automatically start REFMAC for rigid body and restrained refinement of the search model.



If the statistics of the resulting solution are unsatisfactory, try using a smaller fragment of the assumed oligomeric assembly as a search model, such as a homodimer instead of a homotetramer. In the example above, the value of “# Models / AU” would be 6 (representing 6 copies of the dimer in the AU).

When the analyzed molecule is monomeric, and there are many copies of the molecule in the asymmetric unit, MR search may become more complicated. It is possible that after searching for the first model, only some of the molecules expected to be present in the asymmetric unit will be localized. As in the heteromeric complex case, the first partial solution can be “fixed” and used in conjunction with additional rounds of MR.

Appendix A: a note about file locations and management

During the structure solution process, a large number of files (substructure atom positions, structure factor files, model coordinates, output logs, etc.) will be produced “under the hood.” In general, HKL-3000 will handle all input and output for you and you should not have to manage these files directly. In particular, the program will automatically name all directories and files for you according to a regular system, and you will not need to specify the names of output files. However, it may sometimes be useful to view particular files, and it is helpful to understand where HKL-3000 stores the files it produces.

As all of the files produced are associated with one or more Scalepack (*.sca) files, all of this information will be written into a directory named `structure` (or `structure_mr` for molecular replacement) in the same directory as the Scalepack file. For example, if the reflection file `/home/ania/lysozyme/proc/lysozyme.sca` is chosen for determination on the **Project** page, all SAD/MAD structure solution files will be written into `/home/ania/lysozyme/proc/structure/`. Molecular replacement files will be written into `/home/ania/lysozyme/proc/structure_mr/`. If multiple Scalepack files are selected on the **Project** page, the directory containing the first file is used. (For this reason, it is strongly recommended when working with multiple *.sca files to ensure they are in the same directory before starting structure solution.)

Within the `structure` (or `structure_mr`) directory, each step of the process will create a new subdirectory to store the files relevant to that step, with the number of the step in sequence appended. For example this might be a typical list of subdirectories in the `structure` directory after a structure solution run:

```
find_sites_1
phase_2
find_sites_3
phase_4
phase_5
build_model_6
```

Important: you should not rename directories in this list, or add new directories or files to the `structure` (or `structure_mr`) directory, as this can corrupt HKL-3000's internal state!

Whenever HKL-3000 asks you to select the name of a file from a list, it will use the most recently created directory (e.g. with the highest step number). In the rare situations where you need to select the name of a file from an earlier step, (for example, on the **Build** subpage under the **Structure** tab) you may click the blue “History” button in the left sidebar to explicitly select a prior directory to search for files.