

Chapter 4

Computational aspects of high-throughput crystallographic macromolecular structure determination

Paul D. Adams^{1,‡}, Ralf W. Grosse-Kunstleve¹ and Axel T. Brunger²

¹Lawrence Berkeley Laboratory, 1 Cyclotron Road, BLDG 64R0121, Berkeley CA 94720, USA. ²The Howard Hughes Medical Institute and Departments of Molecular and Cellular Physiology, Neurology and Neurological Sciences, Structural Biology, and Stanford Synchrotron Radiation Laboratory, Stanford University, J. H. Clark Center E300-C, 318 Campus Drive, Stanford, CA 94305, USA.

‡Email: PDAdams@lbl.gov

Introduction

The desire to understand biological processes at a molecular level has led to the routine application of X-ray crystallography. However, significant time and effort are usually required to solve and complete a macromolecular crystal structure. Much of this effort is in the form of manual interpretation of complex numerical data using a diverse array of software packages, and the repeated use of interactive 3-dimensional graphics. The need for extensive manual intervention leads to two major problems: significant bottlenecks that impede rapid structure solution (Burley et al., 1999), and the introduction of errors

due to subjective interpretation of the data (Mowbray et al., 1999). These problems present a major impediment to the success of structural genomics efforts (Burley et al., 1999; Montelione & Anderson, 1999) that require the whole process of structure solution to be as streamlined as possible. See Chapter [40](#) for a detailed description of structural genomics. The automation of structure solution is thus necessary as it has the opportunity to produce minimally biased models in a short time. Recent technical advances are fundamental to achieving this automation and make high-throughput structure determination an obtainable goal.

High-throughput structure determination

Automation in macromolecular X-ray crystallography has been a goal for many researchers. The field of small-molecule crystallography, where atomic resolution data are routinely collected, is already highly automated. As a result, the current growth rate of the Cambridge Structural Database (CCSD) (Allen et al., 1983) is more than 15000 new structures per year. This is approximately ten times the growth rate of the Protein Data Bank (PDB) (Berman et al., 2000). See Chapters 11, 12, and 13 for further details of structural databases. Automation of macromolecular crystallography can significantly improve the rate at which new structures are determined. The goal of automation moved to a position of prime importance with the development of the concept of structural genomics (Burley et al., 1999; Montelione & Anderson, 1999) and the routine application of high-resolution macromolecular crystallography to study protein/ligand complexes for drug discovery (Nienaber et al., 2000). In order to exploit the information present in the

rapidly expanding sequence databases it has been proposed that the structural database must also grow. Increased knowledge about the relationship between sequence, structure and function will allow sequence information to be used to its full extent. The success of structural genomics requires macromolecular structures to be solved at a rate significantly faster than at present. This high-throughput structure determination depends on automation to reduce the bottlenecks related to human intervention throughout the whole crystallographic process. Automation of structure solution from the experimental data relies on: the development of algorithms that minimize or eliminate subjective input, the development of algorithms which automate procedures that were traditionally performed by manual intervention, and finally the development of software packages which allow a tight integration between these algorithms. Truly automated structure solution requires the computer to make decisions about how best to proceed in the light of the available data.

The automation of macromolecular structure solution applies to all of the procedures involved beginning with data collection to structure refinement. There have been many technological advances that make macromolecular X-ray crystallography easier. In particular, cryo-protection to extend crystal life (Garman, 1999), the availability of tunable synchrotron sources (Walsh et al., 1999a), high-speed CCD data collection devices (Walsh et al., 1999b) and the ability to incorporate anomalously scattering selenium atoms into proteins have all made structure solution much more efficient (Walsh et al., 1999b). The desire to make structure solution more efficient has lead to

investigations into the optimal data collection strategies for multi-wavelength anomalous diffraction (Gonzalez et al., 1999; Gonzalez, 2007) and phasing using single anomalous diffraction with sulfur or ions (Dauter et al., 1999; Dauter & Dauter, 1999). It has been shown that MAD phasing using only 2 wavelengths can be successful (Gonzalez et al., 1999). The optimum wavelengths for such an experiment are those that give a large contrast in the real part of the anomalous scattering factor (e.g. the inflection point and high-energy remote). However, it has also been shown that, in general, a single wavelength collected at the anomalous peak is sufficient to solve a macromolecular structure (Rice et al., 2000). Such an approach minimizes the amount of data to be collected and increases the efficiency of synchrotron beamlines, and is becoming a more widely used technique.

Data analysis

The first step of structure solution, once the raw images have been processed, is assessment of data quality. The intrinsic quality of the data must be quantified and the appropriate signal extracted. Observations that are in error must be rejected as outliers. Some observations will be rejected at the data-processing stage, where multiple observations are available. However, if redundancy is low then probabilistic methods can be used (Read, 1999). The prior expectation, given either by a Wilson distribution of intensities or model-based structure-factor probability distributions, is used to detect outliers. This method is able to reject strong observations that are in error, which tend to

dominate the features of electron-density and Patterson maps. This method could also be extended to the rejection of outliers during the model refinement process.

When using isomorphous substitution or anomalous diffraction methods for experimental phasing the relevant information lies in the differences between the multiple observations. In the case of anomalous diffraction, these differences are often very small, being of the same order as the noise in the data. In general the anomalous differences at the peak wavelength are sufficient to locate the heavy atoms, provided that a large enough anomalous signal is observed (Grosse-Kunstleve & Brunger, 1999; Weeks et al., 2003). However, in less routine cases it can be very important to extract the maximum information from the data. One approach used in MAD phasing is to analyze the data sets to calculate F_A structure factors, which correspond to the anomalously scattering substructure (Terwilliger, 1994). Several programs are available to estimate the F_A structure factors: XPREP (Bruker, 2001), MADSYS (Hendrickson, 1991) and SOLVE (Terwilliger & Berendzen, 1999a). In another approach a specialized procedure for the normalization of structure factor differences arising from either isomorphous or anomalous differences has been developed in order to facilitate the use of direct methods for heavy atom location (Blessing & Smith, 1999).

Merohedral twinning of the diffraction data can make structure solution difficult and in some cases impossible. The twinning occurs when a crystal contains multiple diffracting domains that are related by a simple transformation such as a two-fold rotation about a

crystallographic axis, a phenomenon that occurs in certain space groups or under certain combinations of cell dimensions and space group symmetry (Parsons, 2003). As a result, the observed diffraction intensities are the sum of the intensities from the two distinctly oriented domains. Fortunately, the presence of twinning can be detected at an early stage by the statistical analysis of structure factor distributions (Yeates, 1997). If the twinning is only partial, it is possible to detwin the data. Perfect twinning typically makes structure solution using experimental phasing methods difficult, but the molecular replacement method and refinement (see below) still can be successfully used.

Heavy atom location and computation of experimental phases

The location of heavy atoms in isomorphous replacement or the location of anomalous scatterers was traditionally performed by manual inspection of Patterson maps. However, in recent years labeling techniques such as seleno-methionyl incorporation have become widely used. This leads to an increase in the number of atoms to be located, rendering manual interpretation of Patterson maps extremely difficult. As a result automated heavy atom location methods have proliferated. The programs SOLVE (Terwilliger & Berendzen, 1999a) and CNS (Brunger et al., 1998; Grosse-Kunstleve & Brunger, 1999) use Patterson based techniques to find a starting heavy atom configuration that is then completed using difference Fourier analyses. Shake-and-Bake (SnB) (Weeks & Miller, 1999), SHELX-D (Sheldrick & Gould, 1995) and HySS (Grosse-Kunstleve & Adams, 2003) use direct methods reciprocal-space phase refinement combined with modifications in real-space. SnB refines phases derived from randomly positioned atoms, while

SHELX-D derives starting phases by automatic inspection of the Patterson map. All methods have been used with great success to solve substructures with more than 60 selenium sites. SHELX-D and SnB have been used to find up to 150 and 160 selenium sites respectively. The HySS program from PHENIX provides a high degree of automation, terminating the search once a successful solution has been found.

After the heavy atom or anomalously scattering substructure has been located, experimental phases can be calculated and the parameters of the substructure refined. A number of modern maximum-likelihood based methods for heavy atom refinement and phasing are readily available: MLPHARE (Otwinowski, 1991), CNS (Brunger et al., 1998), SHARP (de La Fortelle & Bricogne, 1997), SOLVE (Terwilliger & Berendzen, 1999a) and Phaser (McCoy et al., 2004). Programs such as PHENIX (Adams et al., 2002) have the advantage of fully integrating heavy atom location (using HySS), site refinement/phasing (using SOLVE or Phaser), and automated choice of heavy atom hand.

Density modification

Often the raw phases obtained from the experiment are not of sufficient quality to proceed with structure determination. However, there are many real space constraints, such as solvent flatness, that can be applied to electron density maps in an iterative fashion to improve initial phase estimates. This process of density modification is now routinely used to improve experimental phases prior to map interpretation and model

building. However, due to the cyclic nature of the density modification process, where the original phases are combined with new phase estimates, introduction of bias is a serious problem. The γ correction was developed to reduce the bias inherent in the process, and has been applied successfully in the method of solvent-flipping (Abrahams, 1997). The γ correction has been generalized to the γ perturbation method in the DM program, part of the CCP4 suite (Collaborative Computational Project 4, 1994), and can be applied to any arbitrary density modification procedure, including non-crystallographic symmetry averaging and histogram matching (Cowtan, 1999). After bias removal, histogram matching is significantly more powerful than solvent flattening for comparable volumes of protein and solvent (Cowtan, 1999). More recently a reciprocal-space maximum-likelihood formulation of the density modification process has been devised and implemented in the program RESOLVE (Terwilliger, 2000; Terwilliger 2002a; Terwilliger 2003a). This method has the advantage that a likelihood function can be directly optimized with respect to the available parameters (phases and amplitudes), rather than indirectly through a weighted combination of starting parameters with those derived from flattened maps. In this way the problem of choices of weights for phase combination is avoided. The concept of statistical density modification has been developed further in the program PIRATE (Cowtan, 2004), where many different probability distributions are used to classify the density.

Molecular Replacement

The method of molecular replacement is commonly used to solve structures for which a homologous structure is already known. As the database of known structures expands as a result of structural genomics efforts this technique will become more and more important. The method attempts to locate a molecule or fragments of a molecule, whose structure is known, in the unit cell of an unknown structure for which experimental data are available. In order to make the problem tractable it has traditionally been broken down into two consecutive three-dimensional search problems: a search to determine the rotational orientation of the model followed by a search to determine the translational orientation for the rotated model (Rossmann & Blow, 1962). The method of Patterson Correlation (PC) refinement is often used to optimize the rotational orientation prior to the translation search, thus increasing the likelihood of finding the correct solution (Brunger, 1997). With currently available programs, structure solution by molecular replacement usually involves significant manual input. Recently however, methods have been developed to automate molecular replacement. One approach has used the exhaustive application of traditional rotation and translation methods to perform a complete 6-dimensional search (Sheriff et al., 1999). More recently, less time consuming methods have been developed. The EPMR program implements an evolutionary algorithm to perform a very efficient 6-dimensional search (Kissinger et al., 1999). A Monte-Carlo simulated annealing scheme is used in the program Queen of Spades to locate the positions of molecules in the asymmetric unit (Glykos & Kokkinidis, 2000). To improve the sensitivity of any molecular replacement search algorithm maximum likelihood methods have been developed in the Phaser program (Read, 2001; Storoni et

al., 2004; McCoy et al., 2005). The traditional scoring function of the search is replaced by a function that takes into account the errors in the model and the uncertainties at each stage. This approach is seen to greatly improve the chances of finding a correct solution using the traditional approach of rotation (Storoni et al., 2004) and translation searches (McCoy et al., 2005). In addition the method performs anisotropic correction of the experimental data and a statistically correct treatment of simultaneous information from multiple search models using multivariate statistical analysis (Read, 2001). This allows information from different structures to be used in highly automated procedures while minimizing the risk of introducing bias. In the future, molecular replacement algorithms may permit experimental data to be exhaustively tested against all known structures to determine whether a homologous structure is already present in a database, which could then be used as an aid in structure determination.

Map interpretation

The interpretation of the initial electron density map, calculated using either experimental phasing or molecular replacement methods, is often performed in multiple stages (described below) with the final goal being the construction of an atomic model. If the interpretation cannot proceed to an atomic model, it is often an indication that the diffraction data collection must be repeated with improved crystals. Alternatively, repeating previous computational steps in data analysis or phasing may generate revised hypotheses about the crystal, such as a different space group symmetry or estimate of unit cell contents. Clearly, completely automating the process of structure solution will

require that these eventualities are taken into consideration and dealt with in a rigorous manner.

The first stage of electron density map interpretation is an overall assessment of the information contained in a given map. The standard deviation of the local root-mean-square electron density can be calculated from the map. This variation is high when the electron-density map has well defined protein and solvent regions and is low for maps calculated with random phases (Terwilliger & Berendzen, 1999b; Terwilliger, 1999). A similar, more discriminating, analysis can be performed by calculation of the skewness of the histogram of electron density values in the unit cell (Podjarny, 1976). It has also been shown that the correlation of the local root-mean-square density in adjacent regions in the unit cell can be used as a measure of the presence of distinct, contiguous solvent and macromolecular regions in an electron density map (Terwilliger & Berendzen, 1999c).

Currently the process of analyzing an experimental electron density map to build the atomic model is a time consuming, subjective process and almost entirely graphics based. Sophisticated programs such as COOT (Emsley & Cowtan, 2004), O (Jones et al., 1991), XtalView (McRee, 1999), QUANTA (Oldfield, 2000), TurboFrodo (Jones, 1978) and MAIN (Turk, 2000) are commonly used for manual rebuilding. These greatly reduce the effort required to rebuild models by providing: libraries of sidechain rotamers and peptide fragments (Kleywegt & Jones, 1998), map interpretation tools and real space refinement of rebuilt fragments (Jones et al., 1991). However, it has been shown that there are substantial differences in the models built manually by different people when presented

with the same experimental data (Mowbray et al., 1999). The majority of time spent in completing a crystal structure is in the use of interactive graphics to manually modify the model. This manual modification is required either to correct parts of the model that are incorrectly placed or to add parts of the model that are currently missing. This process is prone to human error because of the large number of degrees of freedom of the model and the possible poor quality of regions of the electron density map.

Although interactive graphics systems for manual model building have made the process dramatically simpler, there have also been significant advances in making the process of map interpretation and model building truly automated. One route to automated analysis of the electron density map is the recognition of larger structural elements, such as α -helices and β -strands. Location of these features can often be achieved even in electron density maps of low quality using exhaustive searches in either real space (Kleywegt & Jones, 1997) or reciprocal space (Cowtan, 1998; Cowtan, 2001), the latter having a significant advantage in speed because the translation search for each orientation can be calculated using a Fast Fourier Transform. The automatic location of secondary structure elements from skeletonized electron density maps can be combined with sequence information and databases of known structures to build an initial atomic model with little or no manual intervention from the user (Oldfield, 2000). This method has been seen to work even at relatively low resolution ($d_{\min} \sim 3.0 \text{ \AA}$). However, the implementation is still graphics based and requires user input. A related approach in the program MAID also uses a skeleton generated from the electron density map as the start point for locating secondary structure elements (Levitt, 2001). Trial points are extended in space by

searching for connected electron density at C_{α} distance (approximately 3.7Å) with standard α -helical or β -strand geometry. Real space refinement of the fragments generated is used to improve the model. Both of these methods suffer from the limitation that they do not combine the model building process with the generation of improved electron density maps derived from the starting phases and the partial models.

In order to completely automate the model building process, methods have been developed that combine automated identification of potential atomic sites in the map with model refinement. In the ARP/warp system an iterative procedure is used that describes the electron density map as a set of unconnected atoms from which protein-like patterns, primarily the mainchain trace from peptide units, are extracted. From this information and knowledge of the protein sequence a model can be automatically constructed (Perrakis et al., 1999). This powerful procedure, known as warpNtrace in ARP/wARP, can gradually build a more complete model from the initial electron density map and in many cases is capable of building the majority of the protein structure in a completely automated way. Unfortunately this method currently has the limitation of a need for relatively high-resolution data ($d_{\min} < 2.3\text{\AA}$). Data which extend to this resolution are available for less than 60% of the ~16500 X-ray structures in the PDB. Therefore other approaches have been developed to automatically interpret maps at lower resolution (Holton et al., 2000; Terwilliger 2002b, 2003b, 2003c, 2003d). In the PHENIX system (Adams et al., 2002), the combination of secondary structure fragment location and fragment extension by RESOLVE (Terwilliger 2003d) with iterated structure refinement by *phenix.refine* (Afonine et al., 2005) for map improvement provides an automated

model building method that is relatively insensitive to resolution and is capable of typically building 70% or more of a structure even at 3.0Å. With this technology, it is now possible to investigate the variability of models by building many models against the same data (DePristo et al., 2004; DePristo et al., 2005; Terwilliger et al., 2007).

Methods have recently been developed for the automated location and fitting of small molecules into difference electron density maps, a process critical to the crystallographic screening of potential therapeutic compounds bound to their target molecules. These methods have used reduction of the difference electron density to a simpler representation (Zwart et al., 2004; Aishima et al., 2005), or systematic searching against the density map with rigid fragments of the small molecule (Terwilliger et al., 2006). This is still an active area of research, where problems of small molecule disorder and partial occupancy present significant challenges to robust automation.

Refinement

In general the atomic model obtained by automatic or manual methods contains some errors and must be optimized to best fit the experimental diffraction data and prior chemical information. In addition, the initial model is often incomplete and refinement is carried out to generate improved phases that can then be used to compute a more accurate electron density map. However, the refinement of macromolecular structures is often difficult for several reasons. First, the data to parameter ratio is low, creating the danger

of overfitting the diffraction data. This results in a good agreement of the model to the experimental data even when it contains significant errors. Therefore, the apparent ratio of data to parameters is often increased by incorporation of chemical information, i.e. bond length and bond angle restraints obtained from ideal values seen in high resolution structures (Hendrickson, 1985). Second, the initial model often has significant errors often due to the limited quality of the experimental data or a low level of homology between the search model and the true structure in molecular replacement. Third, local (false) minima exist in the target function. The more local minima and the deeper they are will more likely lead to a failed refinement. Fourth, model bias in the electron density maps complicates the process of manual rebuilding between cycles of automated refinement.

Methods have been devised to address these difficulties. Cross validation, in the form of the free *R*-value, can be used to detect overfitting (Brunger, 1992). The radius of convergence of refinement can be increased by the use of stochastic optimization methods such as molecular dynamics-based simulated annealing (Brunger et al., 1987). Most recently, improved targets for refinement of incomplete, error-containing models have been obtained using the more general maximum likelihood formulation (Murshudov et al., 1997; Pannu et al., 1998). The resulting maximum likelihood refinement targets have been successfully combined with the powerful optimization method of simulated annealing to provide a very robust and efficient refinement scheme (Adams et al., 1999). For many structures, some initial experimental phase information is available from either isomorphous heavy atom replacement or anomalous diffraction methods. These phases

represent additional observations that can be incorporated in the refinement target. Tests have shown that the addition of experimental phase information greatly improves the results of refinement (Pannu et al., 1998; Adams et al., 1999). It is anticipated that the maximum likelihood refinement method will be extended further to incorporate multivariate statistical analysis, thus allowing multiple models to be refined simultaneously against the experimental data without introducing bias (Read, 2001).

The refinement methods used in macromolecular structure determination work almost exclusively in reciprocal-space. However, there has been renewed interest in the use of real-space refinement algorithms that can take advantage of high quality experimental phases from anomalous diffraction experiments or non-crystallographic symmetry averaging. Tests have shown that the method can be successfully combined with the technique of simulated annealing (Chen et al., 1999).

The parameterization of the atomic model in refinement is of great importance. When the resolution of the experimental data is limited then it is appropriate to use chemical constraints on bond lengths and angles. This torsion angle representation is seen to decrease overfitting and improve the radius of convergence of refinement (Rice & Brunger, 1994). If data are available to a high enough resolution, additional atomic displacement parameters can be used. Macromolecular structures often show anisotropic motion, which can be resolved at a broad spectrum of levels ranging from whole domains down to individual atoms. The use of the Fast Fourier Transform to refine atomic anisotropic displacement parameters in the program REFMAC has greatly improved the

speed with which such models can be generated and tested (Murshudov et al., 1999). The method has been shown to improve the crystallographic R-value and free R-value as well as the fit to geometric targets for data with resolution higher than 2Å. New programs, such as *phenix.refine* (Afonine et al., 2005), are being developed with the explicit goal of increasing the automation of structure refinement, which still remains a significant bottleneck in structure completion.

Validation

Validation of macromolecular models and their experimental data (Vaguine et al, 1999) is an essential part of structure determination (Kleywegt, 2000). This is important both during the structure solution process and at the time of coordinate and data deposition at the Protein Data Bank, where extensive validation criteria are also applied (Berman et al., 2000). More recently the MolProbity structure validation suite has been developed (Lovell et al., 2003; Davis et al., 2004). This applies numerous geometric validation criteria to assess both global and local correctness of the model. This information can be readily used to correct errors in the model. See Chapters 14 and 15 for more descriptions of validation methods based on stereochemistry and atomic packing. In the future, the repeated application of validation criteria in automated structure solution will help avoid errors that can still occur as a result of subjective manual interpretation of data and models.

Challenges to automation

Non-crystallographic symmetry

It is not uncommon for macromolecules to crystallize with more than one copy in the asymmetric unit. This leads to relationships between atoms in real space and diffraction intensities in reciprocal space. These relationships can be exploited in the structure solution process. However, the identification of non-crystallographic symmetry (NCS) is generally a manual process. A method for automatic location of proper NCS (i.e. a rotation axis) has been shown to be successful even at low-resolution (Vonnrhein & Schulz, 1999). A more general approach to finding NCS relationships uses skeletonization of electron density maps (Spraggon, 1999). A monomer envelope is calculated from the solvent mask generated by solvent flattening. The NCS relationships between monomer envelopes can then be determined using standard molecular replacement methods. When a model is being built automatically it has been shown that the NCS relationships can be extracted from the local features of the electron density map (Pai et al., 2006).

These methods could be used in the future to automate the location of NCS operators and determination of molecular masks. In the case of experimental phasing using heavy atoms or anomalous scatterers, it is possible to locate the NCS from the sites (Lu, 1999; Terwilliger 2002c). The RESOLVE program automates this process such that NCS

averaging can be automatically performed as part of the phase improvement procedure. Non-crystallographic symmetry information can also be used in structure refinement (Kleywegt, 1996) to either decrease the number of refined parameters (NCS constraints), or increase the number of restraints (NCS restraints). The CNS program implements both of these methods, and most other refinement programs implement NCS restraints. It should be noted that NCS sometimes could be very close to crystallographic symmetry. For example, a translational relationship between the molecules in the asymmetric unit can lead to very weak reflections that may be interpreted as systematic absences resulting from crystallographic centering or rotational symmetry in a molecular complex. As a result, this may lead to an assignment of a higher-symmetry space group than is the case of the true crystal symmetry. These possible complications should always be considered during structure solution as they can lead to stalled R-factors in structure refinement.

Disorder

Except in the rare case of very well ordered crystals of extremely rigid molecules, disorder of one form or another is a component of macromolecular structures. This disorder may take the form of discrete conformational substates for sidechains (Wilson & Brunger, 2000), surface loops, or small changes in the orientation of entire molecules throughout the crystal. The degree to which this disorder can be identified and interpreted typically depends on the quality of the diffraction data. With low to medium resolution data dual side chain conformations are occasionally observed. With high resolution data (1.5Å or better) multiple side chain and mainchain conformations are often seen. The

challenge for automated structure solution is the identification of the disorder and its incorporation into the atomic model without the introduction of errors as a result of misinterpreting the data. Disorder of whole molecules within the crystal, as a result of small differences in packing between neighboring unit cells, cannot be visualized in electron density maps. However, the effect on refinement statistics such as the R and free-R value can be significant because no single atomic model can fit the observed diffraction data well. One approach to the problem is to simultaneously refine multiple models against the data (Burling & Brunger, 1994). An alternative approach is the refinement of Translation-Libration-Screw (TLS) parameters for whole molecules or subdomains of molecules (Winn et al., 2001). This introduces only a few additional parameters to be refined while still accounting for the majority of the disorder. However, it still remains a challenge to automatically identify subdomains. The use of normal modes as an alternative parameterization for the molecular flexibility has the potential for refinement of structures at much lower resolution (Delarue & Dumas, 2004; Poon et al., 2007), while also avoiding the need to identify subdomains.

Conclusions

Over the last decade there have been many significant advances towards automated structure determination. Programs such as PHENIX (Adams et al., 2002) and AutoSHARP (Vonnrhein et al., 2006) combine large functional blocks in an automated fashion. The program CNS (Brunger et al., 1998) provides a framework in which

different algorithms can be combined and tested using a powerful scripting language. The CCP4 suite (CCP4, 1994) provides a large number of separate programs that can be easily run from a graphic user interface.

Some progress towards full automation has been made by linking together existing programs which is typically achieved using scripting languages and/or the World Wide Web. However, long-term robust solutions, such as the PHENIX system (Adams et al., 2002), are fully integrating the latest crystallographic algorithms within a modern computer software environment. Eventually, complete automation will need structure solution to be intimately associated with data collection and processing. When automated software permits the heavy atom location and phasing steps of structure solution to be performed in a few minutes, it will enable real time assessment of diffraction data as it is collected at synchrotron beamlines. Map interpretation will need to be significantly faster than the present situation, with initial analysis of the electron density taking minutes rather than the hours or days required currently.

Cited Literature

Abrahams JP (1997): Bias reduction in phase refinement by modified interference functions: Introducing the gamma correction. *Acta Cryst.* D53:371-376.

Adams PD, Pannu NS, Read RJ, Brunger AT (1999): Extending the limits of molecular replacement through combined simulated annealing and maximum likelihood refinement.

Acta Cryst. D55:181-190.

Adams PD, Grosse-Kunstleve RW, Hung L-W, Ioerger TR, McCoy AJ, Moriarty NW, Read RJ, Sacchettini JC, Sauter NK, Terwilliger TC (2002): PHENIX: building new software for automated crystallographic structure determination *Acta Cryst.* D58:1948-1954.

Afonine PV, Grosse-Kunstleve RW, Adams PD (2005): A robust bulk-solvent correction and anisotropic scaling procedure. *Acta Cryst.* D61:850-855.

Aishima J, Russel DS, Guibas LJ, Adams PD, Brunger AT (2005): Automated crystallographic ligand building using the medial axis transform of an electron-density isosurface *Acta Cryst.* D61:1354-1363.

Allen FH, Kennard O, Taylor R (1983): Systematic Analysis of Structural Data as a Research Technique in Organic Chemistry. *Acc. Chem. Res.* 16:146-153.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000): The Protein Data Bank. *Nucleic Acids Res* 28:235-42.

Blessing RH, Smith GD (1999): Difference structure-factor normalization for heavy-atom or anomalous-scattering substructure determinations. *J. Appl. Cryst.* 32:664-670.

Bruker Analytical X-ray Solutions, Madison, WI (2001).

Brunger AT, Kuriyan J, Karplus M (1987): Crystallographic R factor refinement by molecular dynamics. *Science* 235:458-460.

Brunger AT (1992): The Free R value: a Novel Statistical Quantity for Assessing the Accuracy of Crystal Structures. *Nature* 355:472-474.

Brunger AT (1997): Patterson correlation searches and refinement. *Methods Enzymol.* 276:558–580.

Brunger AT, Adams PD, Clore GM, Gros P, Grosse-Kunstleve RW, Jiang J-S, Kuszewski J, Nilges M, Pannu NS, Read RJ, Rice LM, Simonson T, Warren GL: Crystallography & NMR system (CNS) (1998): A new software system for macromolecular structure determination. *Acta Cryst.* D54:905-921.

- The design and implementation of the widely used CNS program is described. The use of a scripting language to develop, test, and implement new features is a powerful feature of the software.

Burley SK, Almo SC, Bonanno JB, Capel M, Chance MR, Gaasterland T, Lin D, Sali A, Studier FW, Swaminathan S (1999): Structural genomics: beyond the human genome project. *Nat. Genet.* 23:151-157.

Burling FT, Brunger AT (1994): Thermal motion and conformational disorder in protein crystal structures: Comparison of multi-conformer and time-averaging models. *Israel Journal of Chemistry* 34:165-175.

Chen Z, Blanc E, Chapman MS (1999): Real-space molecular-dynamics structure refinement. *Acta Cryst.* D55:464-468.

Collaborative Computational Project, Number 4 (1994): The CCP4 Suite: Programs for Protein Crystallography. *Acta Cryst.* D50:760-763.

Cowtan K (1998): Modified phased translation functions and their application to molecular-fragment location. *Acta Cryst.* D54:750-756.

Cowtan K (1999): Error estimation and bias correction in phase-improvement calculations. *Acta Cryst.* D55:1555-1567.

Cowtan K (2001): Fast Fourier feature recognition. *Acta Cryst.* D57:1435-1444.

Cowtan K (2004): Statistical phase improvement without a solvent boundary. *Acta Cryst.* A60, S14.

Davis IW, Murray LW, Richardson JS, Richardson DC (2004): MolProbity: structure validation and all-atom contact analysis for nucleic acids and their complexes. *Nucleic Acids Research* 32:W615-619.

Dauter Z, Dauter M (1999): Anomalous signal of solvent bromides used for phasing of lysozyme. *J. Mol. Biol.* 289:93-101.

Dauter Z, Dauter M, de La Fortelle E, Bricogne G, Sheldrick GM (1999): Can anomalous signal of sulfur become a tool for solving protein crystal structures? *J. Mol. Biol.* 289:83-92.

Deacon AM, Ealick SE (1999): Selenium-based MAD phasing: setting the sites on larger structures. *Structure* 7:161-6.

Delarue M, Dumas P (2004): On the use of low-frequency normal modes to enforce collective movements in refining macromolecular structural models. *Proc. Natl. Acad. Sci. (USA)* 101:6957-6962.

DePristo MA, de Bakker PIW, Blundell TL (2004): Heterogeneity and inaccuracy in protein structures solved by X-ray crystallography. *Structure* 12:831-838.

DePristo MA, de Bakker PIW, Johnson RJK, Blundell TL (2005): Crystallographic refinement by knowledge-based exploration of complex energy landscapes. *Structure* 13:1311-1319.

Emsley P, Cowtan K (2004): Coot: model-building tools for molecular graphics" *Acta Cryst* D60:2126-2132.

Garman E (1999): Cool data: quantity AND quality. *Acta Cryst.* D55:1641-1653.

Glykos NM, Kokkinidis M (2000): A stochastic approach to molecular replacement. *Acta Cryst.* D56:169-174.

Gonzalez A, Pedelacq J-D, Sola M, Gomis-Rueth FX, Coll M, Samama J-P, Benini S (1999): Two-wavelength MAD phasing: In search of the optimal choice of wavelengths. *Acta Cryst.* D55:1449-1458.

Gonzalez A (2007): A comparison of SAD and two-wavelength MAD phasing for radiation-damaged Se-MET crystals. *J. Synchrotron Radiation* 14:43-50.

Grosse-Kunstleve RW, Brunger AT (1999): A highly automated heavy-atom search procedure for macromolecular structures. *Acta Cryst.* D55:1568-1577.

Grosse-Kunstleve RW, Adams PD (2003): Substructure search procedures for macromolecular structures. *Acta Cryst.* D59:1966-1973.

Hendrickson WA (1985): Stereochemically restrained refinement of macromolecular structures. *Meth. Enzymol.* 115:252-270.

Hendrickson WA (1991): Determination of Macromolecular Structures from Anomalous Diffraction of Synchrotron Radiation. *Science* 254:51-58.

Holton T, Ioerger TR, Christopher JA, Sacchettini JC (2000): Determining protein structure from electron-density maps using pattern matching. *Acta Cryst.* D56:722-734.

Howell PL, Blessing RH, Smith GD, Weeks CM (2000): Optimizing DREAR and SnB parameters for determining Se-atom Substructures. *Acta Cryst.* D56:604-617.

Jones TA (1978): A graphics model building and refinement system for macromolecules. *J. Appl. Cryst.* 11:268-272.

Jones TA, Zou J-Y, Cowan SW, Kjeldgaard M (1991): Improved methods for the building of protein models in electron density maps and the location of errors in these models. *Acta Cryst.* A47:110-119.

Kissinger CR, Gehlhaar DK, Fogel DB (1999): Rapid automated molecular replacement by evolutionary search. *Acta Cryst.* D55:484-491.

Kleywegt GJ (1996): Use of non-crystallographic symmetry in protein structure refinement. *Acta Cryst.* D52:842-57.

Kleywegt GJ, Jones TA (1997): Template convolution to enhance or detect structural features in macromolecular electron-density maps. *Acta Cryst.* D53:179-185.

Kleywegt GJ, Jones TA (1998): Databases in Protein Crystallography. *Acta Cryst.* D54:1119-1131.

Kleywegt GJ (2000): Validation of protein crystal structures. *Acta Cryst.* D56:249-265.

de La Fortelle E, Bricogne G (1997): Maximum-Likelihood Heavy-Atom Parameter Refinement in the MIR and MAD Methods. *Methods Enzymol.* 276:472-494.

Levitt DG (2001): A new software routine that automates the fitting of protein X-ray crystallographic electron-density maps. *Acta Cryst.* D57:1013-1019.

Lovell SC, Davis IW, Arendall III WB, de Bakker PIW, Word JM, Prisant MG, Richardson JS, Richardson DC (2003): Structure Validation by C α Geometry: ϕ, ψ and C β Deviation. *Proteins: Structure, Function and Genetics* 50:437-450.

Lu G (1999): FINDNCS: A program to detect non-crystallographic symmetries in protein crystals from heavy atoms sites. *J. Appl. Cryst.* 32:365-368.

McCoy AJ, Storoni LC, Read RJ (2004): Simple algorithm for a maximum-likelihood SAD function. *Acta Cryst.* D60:1220-1228.

McCoy AJ, Grosse-Kunstleve RW, Storoni LC, Read RJ (2005): Likelihood-enhanced fast translation functions. *Acta Cryst.* D61:458-464.

McRee DE (1999): XtalView/Xfit - A Versatile Program for Manipulating Atomic Coordinates and Electron Density. *J. Structural Biology* 125:156-165.

Montelione GT, Anderson S (1999): Structural genomics: keystone for a Human Proteome Project. *Nature Structural Biology* 6:11-12.

Mowbray SL, Helgstrand C, Sigrell JA, Cameron AD, Jones TA (1999): Errors and reproducibility in electron-density map interpretation. *Acta Cryst.* D55:1309-1319.

Murshudov GN, Vagin AA, Dodson EJ (1997): Refinement of macromolecular structures by the maximum-likelihood method. *Acta Cryst.* D53:240-255.

Murshudov GN, Vagin AA, Lebedev A, Wilson KS, Dodson EJ (1999): Efficient anisotropic refinement of macromolecular structures using FFT. *Acta Cryst.* D55:247-255.

Neinaber VL, Richardson PL, Klighofer V, Bouska JJ, Giranda VL, Greer J (2000): Discovering novel ligands for macromolecules using X-ray crystallographic screening. *Nature Biotechnology* 18:1005-1108.

Oldfield T (2000): A semi-automated map fitting procedure. In *Crystallographic Computing 7: Macromolecular Crystallographic Data (Crystallographic Computing)*.

Editors: Bourne PE, Watenpaugh K: Oxford University Press.

Otwinowski Z (1991): Maximum likelihood refinement of heavy atom parameters. In *Isomorphous Replacement and Anomalous Scattering, Proc. Daresbury Study Weekend*.

Warrington: SERC Daresbury Laboratory, Editors Wolf W, Evans PR and Leslie AGW; 80-85.

Pai R, Sacchettini J, Ioerger T (2006): Identifying non-crystallographic symmetry in protein electron-density maps: a feature-based approach. *Acta Cryst.* D62:1012-1021.

Pannu NS, Murshudov GM, Dodson EJ, Read RJ (1998): Incorporation of prior phase information strengthens maximum-likelihood structure refinement. *Acta Cryst.* D54:1285-1294.

Parsons S (2003): Introduction to Twinning. *Acta Cryst.* D59:1995-2003.

Perrakis A, Sixma TK, Wilson KS, Lamzin VS (1997): wARP: Improvement and extension of crystallographic phases by weighted averaging of multiple-refined dummy atomic models. *Acta Cryst.* D53:448-455.

Perrakis A, Morris R, Lamzin VS (1999): Automated protein model building combined with iterative structure refinement. *Nature Structural Biology* 6:458-463.

- An automated method for building and refining a protein model is described. An iterative procedure is used that describes the electron density map as a set of

unconnected atoms from which protein-like patterns are extracted. This method is currently used by crystallographers to automate model building when high resolution data are available (approximately 2.3Å or better).

Podjarny, AD (1976): Thesis, Weizmann Institute of Science, Rehovot.

Poon BK, Chen X, Lu M, Vyas NK, Quijcho FA, Wang Q, Ma J (2007): Normal mode refinement of anisotropic thermal parameters for a supramolecular complex at 3.42-Å crystallographic resolution. *Proc. Natl. Acad. Sci. USA* 104:7869-7874.

Read RJ (1999): Detecting outliers in non-redundant diffraction data. *Acta Cryst.* D55:1759-1764.

Read RJ (2001): Pushing the boundaries of molecular replacement with maximum likelihood. *Acta Cryst.* D57:1373-1382.

- This paper describes the basis for the application of maximum likelihood scoring methods to the problem of molecular replacement. Subsequent implementation of these methods in the program Phaser has made a major contribution to the success of difficult molecular replacement problems.

Rice LM, Brunger AT (1994): Torsion angle dynamics: reduced variable conformational sampling enhances crystallographic structure refinement. *Proteins* 19:277-290.

Rice LM, Earnest TN, Brunger AT (2000): Single Wavelength Anomalous diffraction phasing revisited: a general phasing method? *Acta Cryst.* D56:1413-1420.

Rossmann MG, and Blow DM (1962): The detection of sub-units within the crystallographic asymmetric unit. *Acta Cryst.* 15:24-31.

Sheldrick GM, Gould RO (1995): Structure solution by iterative peaklist optimization and tangent expansion in space group P1. *Acta Cryst.* B51:423-431.

Sheriff S, Klei HE, Davis ME (1999): Implementation of a six-dimensional search using the AMoRe translation function for difficult molecular-replacement problems. *J. Appl. Cryst.* 32:98-101.

Spraggon G (1999): Envelope skeletonization as a means to determine monomer masks and non-crystallographic symmetry relationships: application in the solution of the structure of fibrinogen fragment D. *Acta Cryst.* D55:458-463.

Storoni LC, McCoy AJ, Read RJ (2004): Likelihood-enhanced fast rotation functions. *Acta Cryst.* D60:432-438.

Terwilliger TC (1994): MAD phasing: Bayesian estimates of F_A . *Acta Cryst.* D50:11-16.

Terwilliger TC (1999): σ^2_R , a reciprocal-space measure of the quality of macromolecular electron-density maps. *Acta Cryst.* D55:1174-1178.

Terwilliger TC (2000): Maximum-likelihood density modification. *Acta Cryst.* D56:965-972.

- A procedure is described for reciprocal-space maximization of a likelihood function based on experimental phases and characteristics of the electron-density

map. This powerful approach to phase improvement is able to generate minimally biased phase estimates and will be a valuable tool in the future for all aspects of phase improvement and phase combination.

Terwilliger TC (2002a): Statistical density modification with non-crystallographic symmetry. *Acta Cryst.* D58:2082-2086.

Terwilliger TC (2002b): Automated structure solution, density modification, and model-building. *Acta Cryst.* D58:1937-1940.

Terwilliger TC (2002c): Rapid Automatic NCS identification Using Heavy-Atom Substructures. *Acta Cryst.* D58:2213-2215.

Terwilliger TC (2003a): Statistical density modification using local pattern matching. *Acta Cryst.* D59:1688-1701.

Terwilliger TC (2003b): Automated main-chain model-building by template-matching and iterative fragment extension. *Acta Cryst.* D59:38-44.

- This paper describes a method for automated model building using secondary structure elements and small fragments. This method is readily applied to medium to low resolution data (3.2Å and better).

Terwilliger TC (2003c): Automated side-chain model-building and sequence assignment by template-matching. *Acta Cryst.* D59:45-49.

Terwilliger TC (2003d): Improving macromolecular atomic models at moderate resolution by automated iterative model building, statistical density modification and refinement. *Acta Cryst.* D59:1174-1182.

Terwilliger TC, Klei H, Adams PD, Moriarty NW, Cohn JD (2006): Automated ligand fitting by core-fragment fitting and extension into density. *Acta Cryst.* D62:915-922.

Terwilliger TC, Grosse-Kunstleve RW, Afonine PV, Adams PD, Moriarty NW, Zwart P, Read RJ, Turk D, Hung LW (2007): Interpretation of ensembles created by multiple iterative rebuilding of macromolecular models. *Acta Cryst.* D63:597-610.

- This paper describes the automated building of multiple molecular models against synthetic and real diffraction data sets to better understand what information can be extracted from the multiple models. The results indicate that the variation between the models is a reflection of the uncertainty in the data, rather than different physical conformations of the molecules in the crystal.

Terwilliger TC Berendzen J (1999a): Automated MAD and MIR structure solution. *Acta Cryst.* D55:849-861.

Terwilliger TC, Berendzen J (1999b): Discrimination of solvent from protein regions in native Fouriers as a means of evaluating heavy-atom solutions in the MIR and MAD methods. *Acta Cryst.* D55:501-505.

Terwilliger TC, Berendzen J (1999c): Evaluation of macromolecular electron-density map quality using the correlation of local r.m.s. density. *Acta Cryst.* D55:1872-1877.

Turk D (2000): MAIN 96: An interactive software for density modifications, model building, structure refinement and analysis. In *Crystallographic Computing 7: Macromolecular Crystallographic Data (Crystallographic Computing)*. Editors: Bourne PE, Watenpaugh K: Oxford University Press.

Vaguine AA, Richelle J, Wodak SJ (1999): SFCHECK: A unified set of procedures for evaluating the quality of macromolecular structure-factor data and their agreement with the atomic model. *Acta Cryst.* D55:191-205.

Vonrhein C, Schulz GE (1999): Locating proper non-crystallographic symmetry in low-resolution electron-density maps with the program GETAX. *Acta Cryst.* D55:225-229.

Vonrhein C, Blanc E, Roversi P, Bricogne G (2006): Automated Structure Solution With autoSHARP. *Methods Mol Biol.* 364:215-30.

Walsh MA, Evans G, Sanishvili R, Dementieva I, Joachimiak A (1999a): MAD data collection - current trends. *Acta Cryst.* D55:1726-1732.

Walsh MA, Dementieva I, Evans G, Sanishvili R, Joachimiak A (1999b): Taking MAD to the extreme: Ultrafast protein structure determination. *Acta Cryst.* D55:1168-1173.

Weeks CM, Miller R (1999): The design and implementation of SnB v2.0. *J. Appl. Cryst.* 32:120-124.

Weeks CM, Adams PD, Berendzen J, Brunger AT, Dodson EJ, Grosse-Kunstleve RW, Schneider TR, Sheldrick GM, Terwilliger TC, Turkenburg MG, Uson I (2003): Automatic solution of heavy-atom substructures. *Methods Enzymol* 374:37-83.

Wilson MA, Brunger AT (2000): The 1.0 Å crystal structure of Ca²⁺ bound calmodulin: an analysis of disorder and implications for functionally relevant plasticity. *J. Mol. Biol.* 301:1237-1256.

Winn MD, Isupov MN, Murshudov GN (2001): Use of TLS parameters to model anisotropic displacements in macromolecular refinement. *Acta Cryst.* D57:122-133.

Yeates TO (1997): Detecting and Overcoming Crystal Twinning. *Meth. Enzy.* 276:344-358.

Zwart PH, Langer GG, Lamzin VS (2004): Modelling Bound ligands in protein crystal structures. *Acta Cryst.* D60, 2230-2239