Chapter 14

Phasing Electron Diffraction Data by Molecular Replacement: Strategy for Structure Determination and Refinement

Goragot Wisedchaisri and Tamir Gonen

Abstract

Electron crystallography is arguably the only electron cryomicroscopy (cryo EM) technique able to deliver atomic resolution data (better then 3 Å) for membrane proteins embedded in a membrane. The progress in hardware improvements and sample preparation for diffraction analysis resulted in a number of recent examples where increasingly higher resolutions were achieved. Other chapters in this book detail the improvements in hardware and delve into the intricate art of sample preparation for microscopy and electron diffraction data collection and processing. In this chapter, we describe in detail the protocols for molecular replacement for electron diffraction studies. The use of a search model for phasing electron diffraction data essentially eliminates the need of acquiring image data rendering it immune to aberrations from drift and charging effects that effectively lower the attainable resolution.

Key words: Electron cryomicroscopy (Cryo-EM), Electron crystallography, Electron diffraction, Molecular replacement, Structure refinement

1. Introduction

Electron crystallography became a bona fide method for determining the structures of membrane proteins following the pioneering work by Henderson and Unwin in the mid 1970s (1). Since then, the field of electron crystallography has been steadily evolving with recent advancements in methodology and technology that led to a number of success stories of membrane protein structures that have been determined to resolutions that rival those achieved by X-ray crystallography. Some of these developments are described in detail elsewhere in this book, including advances in grid preparation and

Ingeborg Schmidt-Krey and Yifan Cheng (eds.), *Electron Crystallography of Soluble and Membrane Proteins: Methods and Protocols*, Methods in Molecular Biology, vol. 955, DOI 10.1007/978-1-62703-176-9_14,

sample embedding, improved hardware for data collection, the development of the field emission electron source as well as the development of a highly sophisticated helium-cooled stage (2).

In electron crystallography the membrane protein of interest is crystallized in two-dimensions within a lipid-bilayer where its structure and function can be assayed directly. The environment experienced by the protein closely mimics the native state of the protein in the cell. Lipids and membrane proteins coevolve to form biological membranes and lipids intimately influence the structure and function of membrane proteins. Extensive lipid–protein interactions occur within the lipid bilayer as lipids fit into crevices and irregularities on the protein surface to maintain an electrochemical seal across the membrane. One of the strengths of electron crystallography is that both protein and lipid structure can be determined and lipid–protein interactions studied directly if high enough resolution is achieved as illustrated by a number of examples (3–6). The methods used for growing two-dimensional (2D) crystals are discussed elsewhere in this book.

Once 2D crystals are obtained, structure determination follows either from images of the 2D crystals and/or from electron diffraction data. Images of 2D crystals contain both phase and amplitude information, and can be used directly for structure determination if the resolution is sufficient. The protocols used for image acquisition in electron crystallography are discussed in another chapter in this book (Chapter 8). Once images are collected, they are scanned and Fourier filtered in order to generate a 3D density map using the MRC suite of programs (7), 2DX (8), and/or IPLT (9). In practice it is very difficult to collect high-resolution images of 2D crystals, particularly if these are tilted. The main reasons are instability of the cryo-stage (mechanical and thermal drift) and various charging effects that “smear out” the high-resolution features in the images perpendicular to the tilt axis. The result is that high-resolution structure determination by this technique can take years, depending also on factors such as cryo-EM access, experience, and time that can be dedicated to the image collection.

In contrast, electron diffraction can deliver atomic-resolution information because it is not affected by drift and charging effects and is independent of the contrast transfer function (CTF). A number of recent studies used electron diffraction to determine the structures of membrane proteins to resolutions where water molecules become visible and the density for the protein approaches atomicity (3, 4, 6), with the highest resolution recorded at 1.7 Å anisotropically and 1.9 Å isotropically (3). We outline detailed protocols and strategies for the collection of high-resolution electron diffraction data elsewhere in this book. One drawback of electron diffraction is that it provides only intensities (amplitude information) but not phase information. To solve the structure, phases need to be determined by other means. In X-ray crystallography, heavy-atom labeling can be used to determine phases but this is not possible in electron crystallography because the phasing power is too small, and the phases thus obtained are not accurate enough to reliably interpret the map (10). In the case of AQP0, phases were determined by molecular replacement (MR)—a common procedure in X-ray crystallography but one that was implemented only recently for electron crystallography (11). This procedure rapidly generated an atomic-resolution electron density map into which it was possible to build an atomic model using protocols analogous to X-ray crystallography. In this chapter we outline in detail protocols for molecular replacement in electron crystallography as well as strategies for structure determination and refinement.

2. Materials

2.1. Equipment

A computer running any Linux platforms or Macintosh OS-X, or Microsoft Windows, supported by the CCP4 software for macromolecular crystallography.

2.2. Program

Among the most widely used and well-supported MR programs are Amore (12, 13), Molrep (14, 15), and Phaser (16, 17) which are distributed as part of the Collaborative Computational Project No. 4 (CCP4) software for macromolecular crystallography (18) and MR implementation in CNS (19, 20). All programs mentioned here are freely available for academia. However, we recommend inexperienced readers to use the Molrep program because of the ease of use and its several automated MR features. Molrep is installed as part of the CCP4 program suite and can be run from the CCP4i graphical user interface (21), command lines, or shell scripts. A link for software download and instructions for installation can be found at the CCP4 Web site (www.ccp4.ac.uk). For other MR programs, readers are referred to appropriated program manuals for more information.

3. Methods

3.1. Patterson Function

MR is not a homology modeling technique, in which the amino acid sequence of the unknown structure is simply threaded computationally onto the known structure of a homologous protein. MR is a method to estimate initial phases of an unknown crystal structure using a structure of a related known molecule (search model). MR comprises complex calculations for comparing and matching mathematical information derived from the structure of the search model with data derived from the diffraction intensities of the unknown structure in order to arrive at approximated phases. The mathematics involved is known as the Patterson function—the Fourier transform of the squared amplitude ($|F|^2$) with phases set to zero. In simple terms, the Patterson function corresponds to a map of inter-atomic vectors that can be calculated directly from experimental diffraction intensities ($I = |F|^2$) of the unknown structure without phase information. Likewise, the Patterson function can be calculated from amplitude parts of the search model structure factors without using the phase components. When the structure of the search model is similar to the unknown structure, their inter-atomic vectors resemble each other, resulting in similar Patterson functions.

The Patterson function for a protein structure in a crystal lattice is complicated by having multiple sets of inter-atomic vectors within the same protein molecules (self-vectors) from different orientations in the lattice related by crystallographic symmetry, and inter-atomic vectors across to neighboring molecules (cross-vectors). These vectors provide valuable information on how protein molecules are arranged in the crystal lattice. However, if we draw a sphere centered at the origin with a small enough radius, the vectors included in the sphere are mostly self-vectors. The use of the vector maps (Patterson function) is the basis for MR.

3.2. Data Preparation

Prior to starting MR calculations it is important to check the quality and completeness of the experimental diffraction data, as well as any anisotropy, intensity statistics, and possible twinning (the latter not usually being a problem in electron crystallography). These factors may impact the success of finding MR solutions and MR parameters can be adjusted appropriately in some difficult cases to increase the signal.

Even though there are only 17 possible plane groups for 2D crystals (Table 1), compared to 230 space groups for 3D crystals (of which 65 are for biological macromolecules), finding the right space group still requires some effort. For example, indexing and merging programs cannot distinguish space groups within the same Laue group such as P222, P222$_1$, and P2$_1$2$_1$2, without knowledge of systematic absences. Systematic absences may not be reliable when data are not complete because certain wedges containing reflections of systematic absence information are missing or when low-resolution intensities are overloaded and eliminated by a merging program. Also noncrystallographic translation in the asymmetric unit often introduces awkward intensity profiles that appear as if systematic absences are possible. Nevertheless, molecular replacement may be able to pick out the true space group based on the success of finding a correct solution with high correlation. With additional information on the native Patterson function (also

Table 1
17 space groups in 2D crystals and their 3D equivalents for biological macromolecules

Cell geometry	Plane group	2D space group	2D space group number	IUCr space group	IUCr space group number	Bravais lattice	Systematic absence (*h*, *k*, *l*)
Oblique	*p1*	P1	1	P1	1	Primitive triclinic	–
Oblique	*p2*	P2	2	P2	3*		–
Rectangular	*pm*	P12	3	P2	3*	Primitive monoclinic	–
Rectangular	*pg*	$\mathrm{P12_1}$	4	$\mathrm{P2_1}$	4*		$(0, 2n+1, 0)$
Rectangular	*cm*	C12	5	C2	5	C centered monoclinic	–
	p2mm	P222	6	P222	16		–
Rectangular	*p2mg*	$\mathrm{P222_1}$	7	$\mathrm{P222_1}$	17	Primitive orthorhombic	$(0, 0, 2n+1)$
	p2gg	$\mathrm{P22_12_1}$	8	$\mathrm{P2_12_12}$	18		$(2n+1, 2n+1, 0)$
Rectangular	*c2mm*	C222	9	C222	21	C centered orthorhombic	–
	p4	P4	10	P4	75*		–
Square	*p4mm*	P422	11	P422	89	Primitive tetragonal	–
	p4gm	$\mathrm{P42_12}$	12	$\mathrm{P42_12}$	90		$(0, 2n+1, 0)$
	p3	P3	13	P3	143*		–
	p3m1	P321	14	P321	150		–
Hexagonal	*p31m*	P312	15	P312	149	Primitive hexagonal	–
	p6	P6	16	P6	168*		–
	p6mm	P622	17	P622	177		–

*Denotes polar space groups

known as self-Patterson), one might be able to confirm noncrystallographic translation in the crystal. An advantage of electron crystallography over X-ray crystallography is its ability to capture images of 2D crystals directly where the crystal lattice is visualized and the correct space group can be deduced based on the packing pattern in the lattice.

Most crystallographic programs for structure determination were developed for X-ray crystallography and crystallographic symmetry is dictated by space group information that complies with the International Union of Crystallography (IUCr) (and preferably using the Hermann–Mauguin notation). In electron crystallography, the unit cell of the 2D crystals may be assigned by indexing programs as *a* and *b* on the plane of the crystal. Therefore, when a plane group is converted to a space group for structure determination, crystallographic axes in some space groups that do not follow the convention may need to be specified or swapped (Table 1). For example, in plane group *p2* (P2) the symmetry axis is perpendicular to the plane of the membrane while in plane group *pm* (P12) or *pg* (P12$_1$), the symmetry axis is parallel to the membrane plane. By convention, the symmetry axis in space groups P2 and P2$_1$ is parallel to the *b* axis where $\beta \neq 90°$. The orthorhombic plane group *pgg* (P222$_1$2$_1$) is equivalent to space group P2$_1$2$_1$2 for which the twofold axis runs parallel to the *c* axis (perpendicular to the membrane plane). However, in the plane group *p2mg* (P222$_1$), the 2$_1$-screw axis is parallel to the *b* axis on the membrane plane and should be swapped to the *c* axis.

3.3. Model Preparation

The success of a molecular replacement search often depends on the quality of the search model. The more similar the model to the unknown structure, the higher the correlation of their Patterson functions. The search model is generally identified by amino acid sequence alignment. Structures with sequence identity higher than 25% often present structural similarity. However, some structures can appear to be similar (same fold) even with less than 25% sequence identity but their pairwise superposition could yield large root-mean-square (r.m.s.) deviations. Since Patterson functions are distance based (inter-atomic vectors), models with lower r.m.s. deviation from the unknown structure are generally more suitable than models with larger r.m.s. deviations. If the choice of model is limited, having a decent model is still better than having no model. If multiple unique structures can be identified form homology search, it does not hurt to try them all for MR. However, the model with higher sequence identity is expected to have smaller r.m.s deviation from the unknown structure and that can make MR easier. Alternatively, multiple homologous structures (of the same part in the unknown structure) can be used for MR by superimposing them into one ensemble and using the ensemble as a search model. The ensemble model may work better than a single model in some cases.

After a known homologous structure is selected, the model should be modified to remove hydrogen atoms, alternative conformations, water molecules, and heteroatoms (such as ligands, and metals). Any parts of the models that may pose major differences to the unknown structure should also be removed. If the model has high sequence identity (>50%) to the unknown structure, the model side chains may be left unchanged. If the sequence alignment can be obtained with high confidence, identical residues may be kept intact and only different residues should be modified to alanine. Alternatively, all side chain C_γ may be kept in addition to C_β (as alanine) since removing too many atoms that are otherwise similar to those in the unknown structure will decrease the signal and correct solutions may become less clear in some difficult cases. The program Chainsaw (22) offers a convenient way to generate models from PDB files with choices of model modification from sequence alignment. If the sequence identity is low (<25%), all side chains are often removed by converting the model to poly-alanine (with glycine intact) as sequence alignment becomes less reliable. Protein termini and flexible loops connecting secondary structure elements should be examined and removed if these regions are highly variable (by length and by structure flexibility). This can be guided by checking atomic *B*-factors. Flexible regions in the structure usually have higher *B*-factors than the average for the entire molecule and they can be removed. After the coordinates are modified, atomic *B*-factors of the model should be set to a low level (15–20 Å^2). Alternatively, the model *B*-factors can be computed based on surface accessibility. In this approach, atoms on the surface will be given higher *B*-factors than buried atoms to smear out electron density at the surface to account for conformational flexibility of surface residues.

If the model and the unknown structure contain multiple domains connected by loops, it is possible that their domain arrangement may be different due to flexibility. It is recommended that multiple search models each containing only one domain should be prepared if no clear solutions can be found using the complete multi-domain model. No clear solution is likely caused by the difference in position of each domain relative to another in the search model compared to the domains in the unknown structure. After each domain is separated into different search models, MR should be performed by first searching for the largest domain with highest homology (accounting for largest structure amplitudes of the asymmetric unit content with most similar Patterson function). The multiple copy search strategy is given in detail in Subheading 3.7.

Before performing the molecular replacement search, the model structure is placed in a large P1 unit cell to simplify the Patterson function. In this way, the inter-atomic vectors of the model are only self-vectors clustered near the origin within a certain spherical radius far separated from cross-vectors. In most cases, the model should also be shifted to the origin. Many MR programs have automated features to prepare the model in a large P1 cell and shift the model to the origin. By default, users only need to provide the model coordinates.

3.4. Self-Rotation Function

The self-rotation function is a useful tool for determining how many molecules are present in the asymmetric unit and how they are related by noncrystallographic rotational symmetry. Essentially, the self-rotation function is a product of rotating and imposing the Patterson function onto itself and does not require a search model. The self-rotation function always produces the largest peak at the origin corresponding to its un-rotated self-imposing. If there is more than one molecule in the asymmetric unit that is not related by translation, rotated self-vectors for each molecule within a certain spherical radius can be matched to un-rotated self-vectors from other molecules which then produce additional peaks at the angles corresponding to the relative rotation between these molecules.

3.5. Cross-Rotation Function

The typical MR method comprises two essential steps: (1) determination of the orientation of the unknown structure by rotational searches of the model, and (2) determination of the position of the unknown structure in the asymmetric unit by translational searches of the oriented model from the first step. The first step involves the calculation of the cross-rotation function (RF) which is essentially a correlation between the model and the crystal Patterson functions within a limited spherical radius. By rotating the model, the Patterson function is also rotated by the same angles. Therefore, small increments of rotation angles of the model Patterson function can be sampled to cover the entire angular possibility, and the match to the un-rotated crystal Patterson function can be calculated by RF. At any rotation angle, a large score (peak) appears when the rotated model Patterson vectors and the crystal Patterson vectors coincide. The higher the scores, the more likely the solutions are correct. In practice, the correct solutions are usually separated from the incorrect ones by a large drop in the score. The number of correct solutions also depends on how many molecules are in the asymmetric unit and how they are oriented relative to each other.

3.6. Translation Function

The second step involves the calculation of the translation function (TF). In space groups with a symmetry higher than P1, the Patterson function also includes cross-vectors generated from atoms that belong to different molecules that are related by crystallographic symmetry. When the search model is translated, the cross vectors change as the positions of symmetry-related molecules calculated from the model for that particular space group also change according to the symmetry operators. The correlation between the Patterson functions of the translated model and the crystal data are calculated by the TF. The translation search is relative to the space group symmetry. In polar space groups such as P2 (Table 14.1 indicated by asterisk), the translational shift is necessary only in the plane perpendicular to the rotational symmetry axis as the Patterson function does not change in the direction parallel to the symmetry axis.

3.7. Molecular Replacement Search Strategies

Biochemical knowledge of the proteins such as their oligomeric states and symmetry can often help with the interpretation of MR results such as noncrystallographic symmetry and relative orientations of the molecules in the crystal. Most membrane proteins have simple oligomeric states (related by rotational symmetry instead of complicated point group symmetry) due to the constraint of the membrane plane both in the native membrane bilayer in cells and in 2D crystals. However, one has to be open minded that crystal packing and crystallographic symmetry may border on the multimeric proteins in such a way that the contents of the asymmetric unit are not necessarily the entire biological ensemble. For example, if a dimeric protein related by a twofold rotational noncrystallographic symmetry crystallizes in the space group containing a crystallographic twofold that coincides with the dimer twofold axis, the asymmetric unit content may only contain one subunit.

MR gives a clearer solution when the search model approximates the totality of the asymmetric unit content. When there are two copies or more of the molecule in the asymmetric unit, multiple approaches can be used to find correct solutions. The first approach is to perform a standard MR search for one copy at a time from a search model containing one copy of the molecule. In this approach, the first copy is searched for and the best rotation function can be selected. Note that in polar space groups in which the origin is not defined along the symmetry axis, the first copy is placed in an arbitrary origin and TF is only calculated in the direction perpendicular to the symmetry axis. By fixing the first copy, the model is then used to search for a second copy. If a solution is found, RF and TF are usually further improved. Additional copies can be investigated by fixing the position of initial units until all copies are found. This approach works well in most cases but becomes problematic when the asymmetric unit contains too many copies and the search model only accounts for less than 20% of the asymmetric unit content.

The second approach is to perform searches with two copies simultaneously. This approach has been implemented in the program Molrep (23) as a dyad search (for two identical copies) to find top orientations from RF for a monomer and construct dimer search models (dyads) based on monomer orientations identified from RF. From the properly oriented dyads, the program calculates a special translation function (STF) that gives the intermolecular vector between properly oriented monomers. This information is then used to calculate standard TF and correlation coefficients from the dyad search model. This approach may be useful when solutions cannot be found by searching with only a single copy.

The third approach is to use the locked cross rotation function (LRF) and locked translation function (LTF) (24–26). This approach takes advantage of point group NCS operators if present in the assembly. The point group NCS is first identified using the self-rotation function. Rotational search is then carried out for the first and second copies. Only orientations consistent with the NCS found by self-rotation functions are selected. The NCS is then expanded for additional copies to define their orientations in the entire assembly and to produce LRF. Next, the translational search is performed for the first copy, followed by the NCS expansion for additional copies in the assembly. The LTF is calculated based on vectors among NCS related molecules to position them correctly relative to the center of the NCS in the assembly before performing standard TF to locate the position of the assembly in the unit cell. This approach is only applicable to proteins that have a point group symmetry and proves particularly beneficial to large protein assemblies.

Alternatively, if the homologous protein has a known functional oligomeric assembly, sometimes it can be assumed that the unknown structure may also form a similar assembly. Therefore, the search model containing a complete assembly (dimer, trimer, etc.) should be tried if searching by a monomer does not find clear solutions. However, the assembly of the model can be slightly different from that of the unknown structure by a combination of a small rotation and translation from the center of the NCS for each subunit, resulting in failure of MR. This problem may be overcome but requires some effort to vary the orientations and positions of each monomer in the model assembly and perform several MR runs with different modified models.

3.8. Protocol for Molecular Replacement Using CCP4i

The advancement in computer technologies, software development, and automation has been tremendously helpful in speeding up or eliminating tedious procedures dealing with data and model preparation for MR. Programs such as Molrep (15) and Phaser (16) can finish automated model preparation and MR in all possible space groups in a short period of time. Program pipelines such as Balbes (27) and MrBUMP (28) also offer a convenient way to solve structures by MR based on only the input data and amino acid sequence of the unknown structure.

Here we will focus on the use of Molrep together with additional CCP4 programs (18) and, as a test case, for phasing the electron crystallographic structure of lens aquaporin-0 (AQP0, protein data bank (PDB) accession code 2B6O) (3) using the X-ray structure of bovine aquaporin-1 (AQP1, protein data bank accession code 1J4N) (29). We will use the CCP4i graphic user interface (21) to run the relevant programs. The program names are given at the beginning for each step for users who are interested in running the programs from command lines or shell scripts.

3.8.1. Downloading Files from Protein Data Bank

1. Create a directory where you store files and perform MR. This directory will be referred to as the working directory.
2. Go to the PDB Web site (www.pdb.org).
3. In the search box for "PDB ID or text," type "2B6O" and click search to retrieve a PDB entry for the electron crystallographic structure of AQP0. On the right side of the PDB code, click on the drop-down menu "Download files," and select by clicking "Structure Factor (text)." You will be asked whether to open or save file "2b6o-sf.cif." Save this file in the working directory. This file contains amplitudes (|*F*|) in mmCIF format from diffraction data of aquaporin-0 crystal which will be treated as an unknown structure. The first part of this file contains header information describing in each line each column in the second part. The second part contains reflection data in column format. Columns 4–6 contain *h k l* indices for each reflection, and columns 5 and 6 contain measured amplitudes (|*F*|) and their signal-to-noise ratio expressed as standard deviation (σ|*F*|), respectively. The last column is the reflection status which tells each reflection whether to be used for refinement or as a test set for cross-validation.
4. In the same PDB entry as above, click on "Sequence" tab. Under "Chain Display," click on "[fasta]" to download file "2B6O_A.fasta.txt" containing the aquaporin-0 sequence in fasta format. Rename this file to "2B6O_A.fas" and save it in the working directory where you want to perform MR. Next, use a text-editing program to remove the first line (see below) and keep only the amino acid sequence.

   >2B6O:A|PDBID|CHAIN|SEQUENCE

   Save the file as "2B6O.seq" in the working directory.
5. In the search box on the PDB Web site, type "1J4N," then click search to retrieve a PDB entry for the crystal structure of the AQP1 water channel. On the right side, click on the drop-down menu "Download files," and select by clicking "PDB File (text)." You will be asked to open or save file "1J4N.pdb." Save this file in the working directory. This file contains coordinates of AQP1 which will be used as a search model for molecular replacement.

3.8.2. Setting up CCP4

1. For Linux or Macintosh OS-X, open a terminal window and change directory into working directory. Launch CCP4i by typing "ccp4i" and hit enter. For Windows, double click on CCP4i icon.

2. If CCP4i is started for the first time, users will be asked to create a "CCP4i project." Fill in a one word alias such as "AQP0" for the project name and the full directory paths for the project (where the working directory is located) and temporary directory (for temporary files created by the program to be stored), then click "Apply & Exit." If the directories do not exist, users will be asked to allow the program to create them by clicking on "create directory."

3.8.3. Converting Data to mtz File Format

The CCP4 software suite uses mtz binary files as a format for reflection data. Users are encouraged to visit the CCP4 Web site for more information on the mtz file format. If the reflection data file contains intensities (*I*) (after indexing and merging) instead of amplitudes (|*F*|), the file should be converted using the "data reduction" module. Appropriate tasks may be chosen depending on the data file format. Here the file "2b6o-sf.cif" already contains |*F*| and we will use the program cif2mtz to convert mmCIF format to mtz.

1. To convert reflection data in mmCIF file format to mtz format, open the "Reflection Data Utilities" module on the left side of the CCPi window and select "Convert to/modify/extend MTZ" task. A task interface window appears. Fields colored orange are required and fields in gray can be left blank for default parameters.
2. From the top in the "Job title" field, type in a job title that can be easily recognized for back-tracking such as "convert from mmCIF to mtz." In the next line "Import reflection file in," select from the drop-down menu "mmCIF." The box in front of "create full unique set of reflections" should be checked and "keep existing FreeR data" is selected from the drop-down menu (default). (Optional) If the data does not have FreeR reflections flagged, "generate FreeR data" may be selected to flag FreeR reflections at this step. To enter input files, click on "Browse" on the right of the "To" field and select the file name "2b6o-sf.cif" from a list in the file window. The file name "2b6o-sf.mtz" automatically appears in the "Out" field. This file will be the output reflection file in the mtz format. Type in the fields "AQP0" for crystal name and "HighRes" for data set name. This is useful if multiple data sets from the same crystal or data from multiple crystals will be stored in the same file (in different columns) as their names can be established from the identifiers. However, these fields may be left blank. In the cell space group name or number, type in "P422" for space group. This should be a space group in which diffraction intensity data are indexed and merged. In the cell dimensions field, type a "65.5," b "65.5," c "160.0," alpha "90," beta "90," gamma "90." The numbers are in Ångstrom unit for unit cell dimensions and degrees for angles. In the line "FreeR column label," type "FREE" (default).

3. Click on "Run" drop-down menu and select "Run now." A new job will appear in the job database in the central window. When the run is finished, the job status in the job database window changes from "Running" to "Finished." Output and Log files can be viewed by clicking on (which will highlight) the line corresponding to the desired job and click on "View files from job" drop-down menu on the right hand side and select appropriate files. The log file is in the upper section of the drop-down menu. File names "2b6o-sf.mtz" should appear on the list of output files in the middle section.

3.8.4. Checking Data Quality

Data quality can be checked using different validation tools. In this protocol, we will focus on program Sfcheck (30). Program Truncate (31) (also available in CCP4i) can be used in addition to Sfcheck. Both programs analyze and report statistics for intensities and amplitudes such as data completeness, anisotropy, Wilson *B*-factor, twinning, and pseudo-translation (only in Sfcheck). The output log file should be examined for possible pathological cases.

1. Select "Validation & Deposition" module on the left side of the CCP4i window and select "Validate model and/or data" task. A task interface appears in a new window.
2. In the task interface, type in a job title such as "Sfcheck data analysis" in the "Job title" field. Uncheck boxes in front of "Run Rampage to calculate structure geometry" and "Run Procheck to calculate structure geometry." Check the box in front of "Run Sfcheck to analyse" and select "experimental data only" from the drop-down menu. Next line "Run Sfcheck against," select "native SF" data from the drop-down menu. Uncheck the box in front of "Generate anisothermally corrected SF amplitude." In the "MTZ in" field click on "Browse" on the right and select the file "2b6o-sf.mtz" from the file window. FP, SIGFP, and FREE should appear in the F, Sigma, and FreeR drop-down menus, respectively. In the "Sfcheck Output PS" field, the file "2b6o-sf_sfcheck1.ps" should appear. This postscript format file summarizes the analysis result. Click on "Run" drop-down menu and select "Run now."
3. Examine the file 2b6o-sf_sfcheck1.ps and/or Log file. In the output file, important values to pay attention to are numbers of strong reflections ($I > 1\sigma$ and $I > 3\sigma$), completeness (53.4% in this case), *B*-factor (34 Å^2 by Patterson, and 32.8 Å^2 by Wilson plot), pseudo-translation (not detected in this case), anisotropic distribution of Structure Factors.

3.8.5. Calculating Cell Content and Matthews' Coefficient

Analysis of cell content gives an idea of how many molecules could be present in the asymmetric unit. The correct number is usually consistent with the self-rotation function and the native Patterson analysis. Knowing the number of molecules can be helpful with the interpretation of rotation and translation functions. Here we will use the program Matthews_coeff to analyze the cell content.

1. In the module menu, select "Molecular replacement" module and click on "Analysis" folder. Select "Cell content analysis" task. A task interface appears in a new window.
2. In the "Job title" field, type "content analysis" as a job title. In the next line "Calculate Matthews coefficient for," select "Protein only" from the drop-down menu. Check the box in front of "Read crystal parameters from mtz file." In the MTZ file field, click on "Browse" on the right and select file "2b6o-sf.mtz" from the file window. Space group "P422" and high resolution limit "1.8" should automatically appear and the box in front of this is checked. In the line "Use molecular weight," select "enter in Daltons" from the drop-down menu. Type "28000" in the field on the right of "Molecular weight of protein or nucleic acid." Alternatively, the molecular weight can be entered by selecting "estimated from number of residues" and typing the number of residues per molecule (the program assumes 112.5 Da per amino acid residue), or by selecting "estimated from sequence file" and enter the sequence file in the "Sequence file" field. Click on "Run" drop-down menu and select "Run now."
3. The result will appear in the same task window (and also in the Log file) as follows:

N_{mol}/asu	Matthews coeff	% solvent	*P*(reso)	*P*(tot)
1	3.06	59.89	0.99	0.99
2	1.53	19.77	0.01	0.01

N_{mol}/asu shows possible number of molecules per asymmetric unit. Matthews coeff and % solvent are Matthews' coefficient (32) and estimated solvent content, respectively, at a given number of protein molecules. *P*(reso) is the normalized probability by high resolution limit based on a recent survey of crystallographic PDB entries (33). The highest *P*(total) is a strong indicator of the preferred solution. Based on cell content analysis, AQP0 crystals most likely contains one molecule per asymmetric unit. On average, protein crystals usually have % solvent content in the range of 40–60%. However, extreme cases have been observed for crystals with solvent contents below 30% and above 70%.

3.8.6. Calculating the Native Patterson MAP

Analysis of the native Patterson map can help identify the presence of noncrystallographic translation in the asymmetric unit. The program Sfcheck analyzes the native Patterson map but currently does not output a Patterson map and peak search results. Here we will use the program FFT (Fast Fourier Transform) (34) to calculate a Patterson map. This step is not essential for AQP0 in this case as there is only one molecule in the asymmetric unit. However, this analysis is recommended for data of unknown structure containing more than one molecule per asymmetric unit and is therefore included here.

1. Select the "Map & Mask Utilities" module and "Generate Patterson map" task. Type in a job title such as "native Patterson" in the "Job title" field. Check the box in front of "Run FFT to generate" and select "Patterson" map in "CCP4" format from drop-down menus. Check the box in front of "List peaks to file" and uncheck the box in front of "Plot default Harker" map sections. In the "MTZ" in field, click "Browse" and select file "2b6o-sf.mtz" from the file window. FP and SIGFP should automatically appear in the drop-down menus for F1 and SigmaF1, respectively. In the "Map" field, the file "2b6o-sf_patterson1.map" should appear. Select "AQP0" from drop-down menu for the map file to be output in the working directory (default in "temporary" directory). The rest on this task window can be left as defaults. Click on "Run" drop-down menu and select "Run now."
2. Examine the output peak files ("2b6o-sf_peaks1.pdb" in orthogonal coordinates and "peaks.ha" in fractional coordinates). In this case, there is only one large peak (peak height 114σ) at 0 0 0, corresponding to the origin peak indicating no noncrystallographic translation. If a noncrystallographic translation is present, additional peak(s) of at least 25% the size of the origin peak should be found. Alternatively, the Patterson map (2b6o-sf_patterson1.map) can be viewed using the program Mapslicer in CCP4i or through a command line.

Sometimes when the input resolution range is too low for calculating a native Patterson map, an extra peak at similar peak height as the origin peak may be observed. This is usually due to the crystallographic translation because the Patterson function is more prone to overlap with a neighboring origin at low resolution. By adjusting the resolution to a higher resolution range, the second peak should disappear. Native Patterson analysis is also implemented into the "MR data analysis" task under the "Molecule Replacement" module. This task runs the programs FFT, Peakmax, Wilson, and Baverage to analyze the data and the search model.

3.8.7. Calculating the Self-Rotation Function

The self-rotation function can indicate the presence of noncrystallographic rotational symmetry in the asymmetric unit. It can often tell the number of molecules in the asymmetric unit and their relative orientation and/or symmetry without the use of a model. Parameters such as resolution range can be adjusted to enhance peak signals. The radius of integration should approximate the diameter of the monomer (twice the radius of gyration) and this value can be estimated from search models. Applying a negative

B-factor to enhance amplitudes can sharpen the data. With automation features in Molrep, the program can calculate its own default setting that often works well. The default value for radius of integration in Molrep is 30 Å. This value should be corrected if a better estimation is available.

1. In the module menu, select the "Molecular replacement" module and select the "Run Molrep – auto MR" task. A task interface appears in a new window. Alternatively, a task "self RT with molrep" under "Analysis" will also lead to the same task window.
2. In the "Job title" field, type "self rotation" as a job title. In the next line, "Do," select "self rotation function" from the drop-down menu (if it has not been selected). Select "MTZ" from the drop-down menu for "Get input structure factors from." In the "MTZ in" field, click on "Browse" on the right and select file "2b6o-sf.mtz" from the file window. Leave the box in front of "Use intensities" unchecked (unless you are using intensities from the input file for the calculation). "FP" and "SIGFP" should automatically appear from the drop-down menus. The rest on this task window can be left blank in order to use default parameters chosen by the program. Click on "Run" drop-down menu and select "Run now."
3. To examine the output files, click on "View Files from Job" on the right side of the user window. The file containing the "srf.molrep_rf" suffix lists all the peaks identified from the self-rotation function. In this case, there is a large peak of 11.20σ at theta = 0, phi = 0, chi = 0 corresponding to the origin peak. Additional peaks are insignificant because their peak heights are much smaller than the origin peak (<2σ). This indicates no NCS rotation in the asymmetric unit. The postscript output file (containing the "rf.ps" suffix in the file name) should also be examined. The plot at chi = 180° can be viewed to identify any twofold axes.

3.8.8. Preparing the Search Model

Many modern MR programs such as Molrep offer an option of inputting the amino acid sequence of the unknown structure together with a homologous structure for the program to generate its own sequence alignment and modify the model to greatly improve the initial model quality. Some automated MR pipelines (Subheading 3.8.12) offer automatic model searches of the Protein Data Bank from the amino acid sequence of the unknown structure. In this section, protocols for manual model modification will be explained in detail.

1. Open the file "1J4N.pdb" using a text editing program of your choice. (For Linux, in a terminal window, go to the working directory and type "nedit 1J4N.pdb" to use the nedit program, or "gedit 1J4N.pdb" to use the gedit program.)

2. Scroll down on the text-editing window, the card “REMARK” in the first column contains experimental information. Under the *B* values section, the mean *B* value (overall) for this structure is 53.20 Å^2, similar to the *B* value from the Wilson plot (42.70 Å^2), indicating that the refined atomic *B*-factors approximate the *B*-factors calculated from the diffraction data well. Noted that for structures determined at lower resolution than 3 Å, the Wilson *B*-factors are not reliable and a large discrepancy is normal. Residues containing atomic *B*-factors in their main chain atoms much larger than the average are highly flexible (as their electron densities are less well-defined) and should be removed from the model.
3. Lines starting with the “HELIX” (for helices and “SHEET” in other PDB files for β-sheets) card contain secondary structure information derived from the coordinates. Since secondary structure is usually conserved in homologous proteins, residues outside defined secondary structure may be removed.
4. The actual coordinates of the AQP1 structure start when the lines start with the “ATOM” card in the first column. In this PDB file, the structure contains only one subunit (chain A) from residue Met 1 to Ser 249 (columns 4 = residue name, column 5 = chain name, column 6 = residue number).
5. If the structure contains hydrogen atoms (usually from NMR spectroscopy or ultra-high resolution X-ray crystallography), the lines containing hydrogen atoms should be removed. This can easily be done with the CCP4 program pdbcur in the “Coordinate Utilities” module and “Edit PDB file” task. This program can also be used to remove alternate conformations, atoms with low/zero occupancy, and anisotropic *U*’s (for high-resolution structures determined by X-ray crystallography where each atom has an additional line in the PDB coordinates for anisotropic *B*-factors).
6. Lines starting with the “HETATM” (for heteroatom) card contain coordinates of nonprotein entities. In this case, the AQP1structure contains ordered BNG (β-nonylglucoside) detergent and ordered water molecules bound to the protein. Since it is unlikely that BNG and water molecules will be present at exactly the same location in AQP0 structure, lines containing HETATM should be removed from the model.
7. Any other lines that do not start with an “ATOM” card are not important for the model preparation purpose and can be deleted. Lines containing CRYST, ORIGX, and SCALE cards may be kept since some programs may require this information to properly interpret the PDB file.
8. In this protocol, we will modify “1J4N.pdb” to create 3 new PDB files for MR comparison. The first file contains the

complete protein coordinates from residues 1 to 249, but everything else is removed. This PDB file is saved as "1J4N-simple.pdb."

9. The second file has flexible loops/termini removed. Residues 1–4 which are not part of the helices and contain high *B*-factors in their main chain atoms (>2 times mean *B*-factor), are deleted. Residues Pro38 to Thr45 are also deleted as these residues are likely to form a flexible loop (high *B*-factor). The PDB file is saved as "1J4N-edited.pdb."

10. Molrep has automated model preparation features for users to have the model modified to poly-alanine or to set *B*-factors to target values (currently fixed at 20 Å^2 or to values related to surface accessibility). Therefore, the model does not need to be edited further from step 8. However, if other MR programs lacking model preparation features are to be used, Gerald Kleywegt's program moleman from the Uppsala Software Factory (URL http://xray.bmc.uu.se/usf/moleman_man.html) is a quick and convenient way to edit the model. In this protocol, the third PDB file containing the poly-alanine model is created using the program Chainsaw in CCP4i. Select the "Molecular Replacement" module in the main CCP4i window, and click on the "Model Generation" box to expand this group for additional tasks and then click on "Create search model." In the "Job title" field, type "Create poly alanine model." In the line "Create search model" select "as polyA model" from the drop-down menu. In the "PDB in" field, click "Browse" on the right and select file "1J4N-edited.pdb" from the file window. The file name "1J4N-edited_chainsaw1.pdb" should automatically appear in the "PDB out" field. Rename this file to "1J4N-edited-polyA.pdb." Click on the "Run" drop-down menu and select "Run now."

11. Alternatively, if a good sequence alignment between the model and the unknown structure (in PIR format) is available, the program Chainsaw (22) can be used for better model preparation by pruning the model atoms based on sequence conservation. To use the program Chainsaw, in the line "Create search model" select "using Chaisaw" from the drop-down menu. Selective modification of nonconserved residues can be performed by selecting "gamma atom," "beta atom," or "last common atom" in the drop-down menu of "prune nonconserved residues to." Chainsaw outputs a coordinate file containing identical residues in the alignment left unchanged and nonidentical residues trimmed to specified atoms in their side chains. Residues that do not align to the sequence are deleted.

3.8.9. Running Molecular Replacement Using Molrep

Since Molrep is an automated MR program, most parameters can be left blank or unchanged from the default values. In some cases, however, default parameters may not be optimal or correct and users are required to set them to more appropriate values in order to get better signal. Normally, users may start with an MR run using the default setting. After inspecting MR results, appropriate parameters should be adjusted or varied for subsequent runs. In this protocol, we will focus on how to run Molrep using the default parameters. Choice of parameters will be discussed when MR results are explained later in the text. In addition, to demonstrate the importance of using a good search model, we will use the different model options below and compare the results:

(a) Unedited model (1J4N-simple.pdb)

(b) Model edited by removing flexible loop and termini (1J4N-edited.pdb)

(c) Poly-alanine model of (b)

(d) Automated model preparation using sequence and unedited model in (a)

1. Select "Molecular Replacement" module from the left side of the CCP4i window and click on "Run Molrep – auto MR" task. A new task interface window will appear.

2. In the "Job title" field, type "MR from 1J4N-simple" as a job title. In the next line, "Do," select "molecular replacement" from the drop-down menu (if it has not been selected) and "performing" select "rotation and translation function." In the line "Get input structure factors from," select "MTZ" from the drop-down menu. Leave the three boxes below unchecked for now. In the "MTZ in" field, click on "Browse" on the right and select file "2b6o-sf.mtz" from the file window. Leave the box in front of "Use intensities" unchecked (unless you are using intensities from the input file for the calculation). "FP" and "SIGFP" should automatically appear from the drop-down menus for FP and SIGFP, respectively.

   (a) In the "Model in" field, click on "Browse" on the right and select file "1J4N-simple.pdb." File name "1J4N-simple_molrep1.pdb" will automatically appear in the "Coords out" field below. The rest on this task window can be left blank for now in order to use default parameters chosen by the program. As default, the *B*-factors of the model are set related to accessibility and the model is shifted to the origin. Click on the "Run" drop-down menu and select "Run now."

   (b) Run Molrep with the file "1J4N-edited.pdb" as the model and change the job title to "MR from 1J4N-edited." The file name "1J4N-edited_molrep1.pdb" will automatically appear in the "Coords out" field. Click on the "Run" drop-down menu and select "Run now."

(c) Run Molrep with the file “1J4N-edited-polyA.pdb” as the model and change the job title to “MR from 1J4N-polyA,” respectively. The file name “1J4N-edited-polyA_molrep1.pdb” will automatically appear in the “Coords out” field. Click on the “Run” drop-down menu and select “Run now.”

(d) Run Molrep using the automated model preparation from the sequence (35). In the same task interface window, type “MR from sequence” in the “Job title” field. Underneath, check the box in front of “Use sequence” and a new section for “Parameter for SEQ” will appear near the bottom of the window. In the “Model in” field, click on “Browse” on the right and select file “1J4N-simple.pdb.” The file name “1J4N-simple_molrep1.pdb” will automatically appear in “Coords out” field below. Rename this file to “1J4N-sequence_molrep1.pdb.” Under the group “Parameter for SEQ,” click “Browse” on the right of the “Seq in” field and select file “2B6O_A.seq.” Click on the “Run” drop-down menu and select “Run now.”

3. By default, Molrep runs MR in a space group defined in the input mtz file. To run MR with all possible space groups, click on the “Infrequently used Parameters” group and in the line “change space group,” select “check all” from the drop-down menu. To test a particular space group, select “define space group” and select a test space group to run from the drop-down menu in the line below.

4. Depending on the CPU, Molrep may take several minutes to run. When the run is finished, the job status in the job database window changes from “Running” to “Finished.” Results from the MR run can be checked by viewing the Log files. Output and Log files can be viewed by clicking on the line corresponding to the desired job and then clicking on the “View files from job” drop-down menu on the right hand side and selecting the appropriate files. Three files should appear on the list of output files: an output coordinates (.pdb) file, a complete search (.doc) file, and a rotation function (.molrep_rf) file. The output coordinate file will be needed for future steps (to check for packing, rigid body, and restrained refinement).

3.8.10. Interpreting Molecular Replacement Solutions from Molrep

1. Begin by checking the Log file from an MR run from step 2a, Subheading 3.8.9. In this run, the program estimated one monomer from the number of atoms in the model corresponding to a volume of the asymmetric unit V_{mol} of 47.5% (important for calculating COMPL). The resolution range used here

is between 22.9 and 2.25 Å. By default, the program detects data anisotropy and corrects it for proper scaling. In this case, the data is highly anisotropic (as shown by Ratio of Eigen values). In this run, the program calculates the “COMPL” and “SIM” parameters as 0.475 and 0.350, respectively. These numbers correspond to model completeness based on asymmetric unit volume (from 0.1 to 1.0) and model similarity to the unknown structure (also from 0.1 to 1.0), which are used to compute B_{off} and B_{add}, respectively. B_{off} and B_{add} control low and high resolution cut off, respectively, for MR and also for scaling of F_{model}. In this case, the model is quite complete (47.5% of the asymmetric unit volume) and has high similarity to the unknown structure (~50%). The values for COMPL and SIM may be entered to override default values. In the Molrep task window, under “The model” group, enter Expect “0.475” fraction completeness of model with “0.5” fraction similarity to input the structure. The program calculates a radius of integration (RAD) of 32.75 Å from the model. If the model is an oligomer (not in this case), the correct radius of integration (twice the radius of gyration) corresponds to the monomer being entered, otherwise the default value will be the radius calculated for the oligomer. To input a search radius, click on the “Infrequently Used Parameters” group and enter the correct number for “Search radius.” The next section includes peaks from cross RF. In addition to plain RF, Molrep uses Rf/sigma to enhance signals and to use for peak ranking. The RF peaks show a big contrast after the first four peaks when the Rf/sigma drops from 5.82–5.65 to 4.57. The next section includes peaks from TF. Molrep calculates Tf/sig and TFcnt (multiplications of different Tf) to enhance peak contrast. In addition, Molrep calculates the packing function (PF) that is an overlapping function of the models (1 = no overlap) which is then used to calculate scor (scoring function = correlation coefficient × packing function). Finally, the program calculates Contrast, the ratio of the top Scor to the mean Scor. Usually Contrast >2.5 indicates a solution. A contrast <1.5 is probably not a solution because it is not significantly different than the mean. The final section summarizes allowed TF results after results containing overlapping models are removed. In this case, Contrast is 4.47 indicating a clear solution.

Rf	TF	theta	phi	chi	tx	ty	tz	TFcnt	Rfac	Scor
1	1	178.44	–144.90	95.53	0.277	0.732	0.356	4.58	0.529	0.543

2. Check the result from step 2b, Subheading 3.8.9 using edited model “1J4N-edited.pdb.” In this case, the contrast is 3.69,

indicating a correct solution. Because the model is improved over the simple model by the removal of flexible loops and termini, the improvement in results is shown by the lower wRfac and higher Scor over the use of the simple model.

Rf	TF	theta	phi	chi	tx	ty	tz	TFcnt	Rfac	Scor
1	1	178.76	–144.90	95.55	0.778	0.235	0.356	4.50	0.523	0.562

However, when the COMPL and SIM parameters are entered in a new run as 0.45 and 0.5, respectively, the result is further improved by the lower wRfac and higher scor with the contrast increased to 4.44.

Rf	TF	theta	phi	chi	tx	ty	tz	TFcnt	wRfac	Scor
1	1	0.76	4.57	174.66	0.235	0.221	0.356	4.29	0.517	0.573

3. Check the result from step 2c, Subheading 3.8.9 using polyalanine model "1J4N-edited-polyA.pdb." In this case, the completeness of the search model suffers from the loss of atoms by conversion of residues to polyalanine. The program calculates a V_{mol} of 33% and expects two monomers, which is incorrect. An incorrect number of monomers may result in an unreliable scoring function. However, the program uses the self-rotation function to determine the actual number of monomers to be used for scoring. In this case, the program found the correct number of a single monomer. The program calculates COMPL and SIM as 0.330 and 0.350, respectively. The RF from this run does not show a big contrast until after 11 peaks, compared to 4 and 3 peaks in (a) and (b), respectively. The TF of the first molecule gives statistics as shown below. Strikingly, the best TF score in this case comes from the 11th and 7th RF peaks (instead of the first peak in the previous two cases) and the second best score is consistent with other MR runs.

Rf	TF	theta	phi	chi	tx	ty	tz	TFcnt	wRfac	Scor
11	1	0.00	0.00	83.34	0.210	0.765	0.360	3.19	0.563	0.498
7	1	178.01	–148.95	96.11	0.210	0.765	0.360	3.89	0.564	0.498

The program continues to search for the second molecule (unless overridden by entering NMON parameter = 1 under "Search Parameters" group and searches for "1" monomers in the asymmetric unit). The program uses the previously calculated RF done for the first monomer to calculate the TF of the second monomer.

Since there is no second monomer in this case, the program cannot find a solution as indicated by the low contrast of 1.42.

4. Check the result from step 2d, Subheading 3.8.9 using the automated model prepared from the sequence and a simple model "1J4N-simple.pdb." In this option, the program calculates a structure-guided sequence alignment. This is essentially a sequence alignment but gaps are not allowed within secondary structure segments and buried residues contribute to the total alignment score more than residues at the surface. The program detects 47.7% identity between the sequence and the model with a gap of nine residues. The model is then modified as follows. First, residues that align with gaps in the sequence are deleted. Second, side chain atoms in the aligned residues of the search model that have no counterpart in the sequence are deleted and only atoms the sequences have in common are kept. Lastly, the residues and atoms of the modified model are renamed and renumbered. The program calculates COMPL and SIM to be 0.429 and 0.477, respectively. The result is much improved over using the simple model in (a) and the polyalanine model in (c) and on par with the edited model in (b). The rotation functions show a big contrast after three peaks from 6.23–6.10 to 4.15. TF scoring gives a contrast of 2.76, indicating a correct solution. The automated model preparation seems to be the best option when a good sequence alignment can be obtained.

Rf	TF	theta	phi	chi	tx	ty	tz	TFcnt	wRfac	Scor
1	1	178.86	−136.33	95.92	0.784	0.241	0.352	3.61	0.521	0.573

5. After the solution is found, a new Molrep run can be performed using the slow MODE parameter. The program default is in the "fast" mode that includes standard RF and TF without rigid body refinement. The "slow" mode uses advanced RF and TF with rigid body refinement to improve MR results and to pave the way for restrained refinement. However, the slow mode generally takes much longer computational time. To change the MODE parameter in the Molrep task window, click on "Infrequently Used Parameters" and in the line "Use," select "advanced RF and TF with rigid body refinement ('slow' mode)" from the drop-down menu. Compared to the fast mode in step 2d, Subheading 3.8.9 the slow mode MR results in improved solution statistics.

Rf	TF	theta	phi	chi	tx	ty	tz	TFcnt	wRfac	Scor
1	1	178.88	−134.54	95.95	0.784	0.241	0.352	3.18	0.451	0.632

3.8.11. Verifying a Probable Solution

After a probable solution has been found, the next step is to check that the solution is indeed correct. In many cases, when Contrast from Molrep is borderline it is possible that the solution may be incorrect. Correct solutions may be hidden among top results but their signal-to-noise ratio may not be high enough to stand out and give good contrast. This usually happens when the search model is poor or MR parameters are not optimal. In this case, additional runs with model modification or different parameters (such as radius of integration) should be tried. In addition, different MR programs such as Phaser (16) (also in CCP4i) should be used. Phaser uses maximum likelihood for scoring instead of using standard RF and TF peak heights and this may help for the correct solution to stand out better. Also, when searching for more than one monomer, Phaser recalculates rotation functions after fixing the first monomer. This process can be slow but often useful to increase the contrast because Phaser modifies F_{obs} by taking the amplitudes of the fixed monomer into account. Furthermore, Phaser can automatically check all space groups, perform rigid body refinement, and output phase information from which density maps can be calculated and examined.

1. A correct solution is likely to give reasonable packing. An easy way to check for packing is to open the output PDB file in molecular rendering programs such as Pymol (www.pymol.org) or Coot (36, 37) that can calculate (and display) neighboring molecules related by crystallographic symmetry. Good packing should not have serious main chain clashes with neighboring molecules while maintaining protein contacts and a clear separation for protein and solvent regions. However, when the model is not perfect, there might be regions in the model that are not present in the unknown structure (due to a different conformation, shorter loops, disordered regions, etc.) and the clashes between molecules are artifacts that can be eliminated by model corrections. It is recommended that several peaks be checked in this way. Sometimes the correct solution can be found in a much lower peak.
2. The correct solution is usually found consistently in different MR runs using different model preparations or parameters and is consistent with the self-rotation function. Check and compare orientations (theta, phi, chi angles) and positions (tx, ty, tz fraction coordinates) of several peaks from different runs.
3. Perform rigid body and restrained refinements and check the refinement *R*-factors (Subheading 4). Molrep outputs results as PDB coordinates that then can be used for refinement using the program Refmac (38) (also part of CCP4i) which automatically calculates phases from the model. However, some other refinement programs may require input phases prior to refinement. In that case, phases can be calculated from the

PDB file using the program SFall (in CCP4i). After restrained refinement, the correct solution will result in a large drop of *R*-factors (especially *R*-free) to below 40%. If *R*-factors stay at high levels, then there is something wrong and could point out that a wrong MR solution has been obtained. However, incomplete models (missing domains or subunits) and model error (different conformation) could also result in high *R*-factors and so the density maps ($2F_{\mathrm{obs}} - F_{\mathrm{cal}}$ and $F_{\mathrm{obs}} - F_{\mathrm{calc}}$) should be examined carefully.

4. Check different space groups. If the space group is wrong, no correct solution should be found. Any probable solutions are basically incorrect and will not refine further. To check, perform MR with different space groups and try basic refinements. The correct space group will give the best *R*-factor following refinement.

3.8.12. Running Molecular Replacement Using Automated MR Pipelines

The success of MR often depends on the choice of search models and how they are prepared for MR searches. The essence of automated program pipelines for MR such as MrBump (28) and Balbes (27) is to take advantage of an increasing number of available structures in the Protein Data Bank to find search models based on sequence similarity to an unknown structure and use all of them for MR searches. The main advantages of MR pipelines are their automation and ease of use. Users only need to provide the sequence of an unknown structure and its amplitude data and the pipelines perform the following three major steps automatically:

1. Database searches from the input amino acid sequence to find as many potential homologous structures (domains, multimers, etc.) as possible. Each program uses several different databases, for example, MrBUMP uses the SCOP database to identify additional search models based on fold similarity.

2. Model preparation. In MrBUMP, a multiple sequence alignment of the potential homologous structures identified in (1) is generated and the structures are ranked and edited. MrBump can generate four different search models for each sequence according to (i) the PDBclip method (removal of water, hydrogen atoms, and alternative conformations); (ii) the Polyalanine method (all side chain truncated to the C_β atom); (iii) the Molrep method (35); and (iv) the Chainsaw method (22). Balbes uses Molrep for model preparation. In addition, both pipelines create additional search models at the domain level. MrBUMP uses the SCOP database to generate additional models truncated to the domain boundary according to domain definitions. Balbes, on the other hand, defines domains based on compactness and truncates the models accordingly. Ensembles of models are also generated as search models.

3. Molecular replacement and restrained refinement. Both pipelines use the programs Molrep (MrBUMP also uses Phaser) for MR, and Refmac (38) for refinement. Each pipeline uses its own definition for solution scoring based on refinement results and suggests the best solution. The pipelines output partially refined models and mtz files containing phases and sigmaA-weighted amplitudes for map calculations.

Running MrBump

1. In the "Molecular Replacement" module, select the "Run MrBUMP" task. A new task interface window will appear.
2. In the "Job title" field, type "Automated MR." In the line "Program Mode," select "Model search and Molecular Replacement" from the drop-down menu. In the "SEQ in" field, click "Browse" on the right and select file "2B6O_A.seq" from the file window. In the "MTZ in" field, click "Browse" and select file "2b6o-sf.mtz." File names "2b6o-sf_mrbump_soln1.mtz" and "2b6o-sf_mrbump_soln1.pdb" should automatically appear in the "MTZ out" and "PDB out," respectively. The rest of the parameters can be left unchanged to use default settings. If desired, the program Phaser can be selected to run MR in parallel or as an alternative to Molrep in some difficult cases.
3. Click on the "Run" drop-down menu and select "Run now."

Running Balbes

1. In the "Molecular Replacement" module, select the "Run Balbes" task. A new task interface window will appear.
2. In the "Job title" field, type "Automated MR." In the line "Do," select "Standard MR" from the drop-down menu. In the "Structure factor file (MTZ or CIF) in" field, click "Browse" on the right and select the file "2b6o-sf.mtz" from the file window. In the "SEQ in" field, click "Browse" and select file "2B6O_A.fas" (Balbes takes input sequence in fasta file format). File names "2b6o-sf_balbes_out1.pdb" and "2b6o-sf_balbes_out1.mtz" should automatically appear in the "Solution PDB" and "Solution HKL," respectively.
3. Click on the "Run" drop-down menu and select "Run now."

4. Methods for Refinement

After a probable MR solution is found, the output coordinates should be refined to check whether or not the solution is correct. Program Refmac (38) in CCP4i can be used to check the refinement after MR by running rigid body refinement followed by restrained refinement cycles. If rigid body refinement has already been performed using MR programs (slow mode in Molrep or Phaser), users can start with restrained refinement directly. Refmac can be used further after manual model building (such as in the program Coot), an automatic model building/rebuilding program such as ARP/wARP (39, 40), or Resolve (41–43) to complete the structure determination process.

4.1. Rigid Body Refinement

1. In the "Refinement" module, select "Run Refmac5" task. A new task interface window will appear.
2. In the "job title" field, type "Rigid body refinement." In the next line "Do," select "rigid body refinement" using "no prior phase information" from the drop-down menus.
3. In the "MTZ in" field, click "Browse" on the right and select the file "2b6o-sf.mtz" from the file window. "FP" and "SIGFP" should automatically appear from the drop-down menus, and the file name "2b6o-sf_refmac1.mtz" appears in the "MTZ out" field. This file name can be changed to a recognizable new name. In the "PDB in" field, click "Browse" on the right and select a Molrep coordinate file such as "1J4N-sequence_molrep1.pdb" from the file window. The file "1J4N-sequence_molrep1_refmac1.pdb" should automatically appear in the "PDB out" field. This file name can be changed to a recognizable new name. By default, Refmac in CCP4i runs 20 cycles of rigid body refinement using all data. The number of cycles and the output resolution can be changed in the "Refinement Parameters" group.
4. If there is more than one identical domain or monomer in the asymmetric unit, it is important to allow them to move independently by treating each domain and/or monomer as a rigid body. To enter domain information, under the "Rigid Domains Definition" group, click on "Add domain Definition" and enter chain id and residue numbers for each domain (or monomer).
5. The rest can be left unchanged to use default parameters. Click on the "Run" drop-down menu and select "Run now" to run the refinement.

4.2. Restrained Refinement

1. In the same task window as Subheading 4.1, type "Restrained refinement" in the "job title" field. In the line "Do," select "restrained refinement" using "no prior phase information" from the drop-down menus.
2. Enter input files as described in Subheading 4.1. By default, Refmac in CCP4i runs 10 cycles of rigid body refinement using all data. To change the number of refinement cycles and the resolution range, click on the "Refinement Parameters" group and enter the desired parameters.
3. By default, Refmac automatically determines a weight matrix (experimental data vs. ideal geometry) for every refinement cycle. However, a defined number can be entered to override

the default. Under the "Refinement parameters" group, uncheck the box in front of "Use automatic weighting," and in the field "use weighting term" below, enter a number (usually between 0.01 and 10). A small number should be used for low-resolution data since the refinement uses less weight for experimental data and restrains the model to tighter geometry. An optimal weight matrix can be determined by trying different weights until a suitable r.m.s.d. for bond lengths and bond angles from ideal geometry is found.

4. If there are multiple monomers (or domains) related by NCS, it is useful to restrain them. To set up NCS restraints, under "Setup Non-Crystallographic Symmetry (NCS) Restraints" group, click "Add NCS restraint." In the drop-down menus, select chains that need to be restrained together and enter the residue numbers to define range. Select a level of restraints for the main chain and side chain from the drop-down menu. "Tight," "medium," and "loose" restraints allow for up to 0.05, 0.5, and 5 Å deviation, respectively. Tight to medium NCS restraint should be selected at the beginning of the refinement. Later when it becomes clear that each monomer should be allowed to deviate more from each other, appropriate NCS restraints should be used and the refinement R-free should drop further.

5. Click on the "Run" drop-down menu and select "Run now."

6. A drop in the R_{work} and R_{free} value often indicates good progress in the refinement and R values below 40% can be expected. If the R values do not decrease during the refinement, one needs to check whether or not the solution is indeed correct. One possibility is that the chosen space group is incorrect and MR should be tried with different space groups. Another possibility is that there are serious clashes between protein subunits. In addition, factors such as model incompleteness (missing parts of the model) and coordinate errors (large r.m.s.d between the model and the target structure) also result in large R values. Therefore after a macro cycle of the geometry restrained refinement, the σ_A-weighted $2F_{\mathrm{obs}}-F_{\mathrm{cal}}$ and $F_{\mathrm{obs}}-F_{\mathrm{cal}}$ density maps (output by the Refmac program) should be examined for positive and negative densities (in a graphic program such as Coot) and the model should be edited according to the density maps that are obtained from the refinement. New atoms should be added in the positive density regions as additional residues or modified amino acid side chains, while residues in the negative density regions should be deleted or revised. Several cycles of model building followed by the geometry restrained refinement should be carried out until the protein model is complete.

7. If the data resolution is better than 3 Å, additional density for water, lipid, and ligand molecules may show up in the later stage of the refinement when the protein model is nearly

complete and the *R* values are low. Cycles of water and ligand adding to the model followed by the geometry restrained refinement should be performed to complete the structure determination.

Acknowledgments

Research in the Gonen laboratory is supported by the American Diabetes Association Award # 1-09-CD-05 and by the National Institutes of Health R01GM079233 and U54GM094598 as well as the Howard Hughes Medical Institute.

References

1. Henderson R, Unwin PN (1975) Three-dimensional model of purple membrane obtained by electron microscopy. Nature 257:28–32
2. Fujiyoshi Y (1998) The structural study of membrane proteins by electron crystallography. Adv Biophys 35:25–80
3. Gonen T, Cheng Y, Sliz P, Hiroaki Y, Fujiyoshi Y, Harrison SC, Walz T (2005) Lipid-protein interactions in double-layered two-dimensional AQP0 crystals. Nature 438:633–638
4. Hite RK, Li Z, Walz T (2010) Principles of membrane protein interactions with annular lipids deduced from aquaporin-0 2D crystals. EMBO J 29:1652–1658
5. Mitsuoka K, Hirai T, Murata K, Miyazawa A, Kidera A, Kimura Y, Fujiyoshi Y (1999) The structure of bacteriorhodopsin at 3.0 A resolution based on electron crystallography: implication of the charge distribution. J Mol Biol 286:861–882
6. Tani K, Mitsuma T, Hiroaki Y, Kamegawa A, Nishikawa K, Tanimura Y, Fujiyoshi Y (2009) Mechanism of aquaporin-4's fast and highly selective water conduction and proton exclusion. J Mol Biol 389:694–706
7. Crowther RA, Henderson R, Smith JM (1996) MRC image processing programs. J Struct Biol 116:9–16
8. Gipson B, Zeng X, Stahlberg H (2007) 2dx_merge: data management and merging for 2D crystal images. J Struct Biol 160:375–384
9. Philippsen A, Schenk AD, Signorell GA, Mariani V, Berneche S, Engel A (2007) Collaborative EM image processing with the IPLT image processing library and toolbox. J Struct Biol 157:28–37
10. Ceska TA, Henderson R (1990) Analysis of high-resolution electron diffraction patterns from purple membrane labelled with heavy-atoms. J Mol Biol 213:539–560
11. Gonen T, Sliz P, Kistler J, Cheng Y, Walz T (2004) Aquaporin-0 membrane junctions reveal the structure of a closed water pore. Nature 429:193–197
12. Navaza J (1994) Amore—an automated package for molecular replacement. Acta Crystallogr A50:157–163
13. Trapani S, Navaza J (2008) AMoRe: classical and modern. Acta Crystallogr D64:11–16
14. Vagin A, Teplyakov A (1997) MOLREP: an automated program for molecular replacement. J Appl Crystallogr 30:1022–1025
15. Vagin A, Teplyakov A (2010) Molecular replacement with MOLREP. Acta Crystallogr D66:22–25
16. McCoy AJ (2007) Solving structures of protein complexes by molecular replacement with Phaser. Acta Crystallogr D63:32–41
17. McCoy AJ, Grosse-Kunstleve RW, Adams PD, Winn MD, Storoni LC, Read RJ (2007) Phaser crystallographic software. J Appl Crystallogr 40:658–674
18. Collaborative Computational Project, Number 4 (1994) The CCP4 suite: programs for protein crystallography. Acta Crystallogr D50:760–763
19. Brunger AT (2007) Version 1.2 of the crystallography and NMR system. Nat Protoc 2:2728–2733
20. Brunger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS, Read RJ, Rice LM, Simonson T, Warren GL (1998) Crystallography & NMR system: a new software suite for macromolecular structure determination. Acta Crystallogr D54:905–921

21. Potterton E, Briggs P, Turkenburg M, Dodson E (2003) A graphical user interface to the CCP4 program suite. Acta Crystallogr D59:1131–1137
22. Stein N (2008) CHAINSAW: a program for mutating pdb files used as templates in molecular replacement. J Appl Crystallogr 41:641–643
23. Vagin A, Teplyakov A (2000) An approach to multi-copy search in molecular replacement. Acta Crystallogr D56:1622–1624
24. Tong L (2001) How to take advantage of non-crystallographic symmetry in molecular replacement: 'locked' rotation and translation functions. Acta Crystallogr D57:1383–1389
25. Tong L, Rossmann MG (1990) The locked rotation function. Acta Crystallogr A46:783–792
26. Tong LA (1996) The locked translation function and other applications of a Patterson correlation function. Acta Crystallogr A52:476–479
27. Long F, Vagin AA, Young P, Murshudov GN (2008) BALBES: a molecular-replacement pipeline. Acta Crystallogr D64:125–132
28. Keegan RM, Winn MD (2008) MrBUMP: an automated pipeline for molecular replacement. Acta Crystallogr D64:119–124
29. Sui HX, Han BG, Lee JK, Walian P, Jap BK (2001) Structural basis of water-specific transport through the AQP1 water channel. Nature 414:872–878
30. 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 Crystallogr D55:191–205
31. French S, Wilson K (1978) Treatment of negative intensity observations. Acta Crystallogr A34:517–525
32. Matthews BW (1968) Solvent content of protein crystals. J Mol Biol 33:491–497
33. Kantardjieff KA, Rupp B (2003) Matthews coefficient probabilities: improved estimates for unit cell contents of proteins, DNA, and protein-nucleic acid complex crystals. Protein Sci 12:1865–1871
34. Teneyck LF (1973) Crystallographic fast Fourier-transforms. Acta Crystallogr A29:183–191
35. Lebedev AA, Vagin AA, Murshudov GN (2008) Model preparation in MOLREP and examples of model improvement using X-ray data. Acta Crystallogr D64:33–39
36. Emsley P, Cowtan K (2004) Coot: model-building tools for molecular graphics. Acta Crystallogr D60:2126–2132
37. Emsley P, Lohkamp B, Scott WG, Cowtan K (2010) Features and development of Coot. Acta Crystallogr D66:486–501
38. Murshudov GN, Vagin AA, Dodson EJ (1997) Refinement of macromolecular structures by the maximum-likelihood method. Acta Crystallogr D53:240–255
39. Cohen SX, Ben Jelloul M, Long F, Vagin A, Knipscheer P, Lebbink J, Sixma TK, Lamzin VS, Murshudov GN, Perrakis A (2008) ARP/wARP and molecular replacement: the next generation. Acta Crystallogr D64:49–60
40. Perrakis A, Harkiolaki M, Wilson KS, Lamzin VS (2001) ARP/wARP and molecular replacement. Acta Crystallogr D57:1445–1450
41. Terwilliger TC (2003) Improving macromolecular atomic models at moderate resolution by automated iterative model building, statistical density modification and refinement. Acta Crystallogr D59:1174–1182
42. Terwilliger TC (2003) Automated main-chain model building by template matching and iterative fragment extension. Acta Crystallogr D59:38–44
43. Terwilliger TC (2003) Automated side-chain model building and sequence assignment by template matching. Acta Crystallogr D59:45–49