Structural Biology Practical Course

Gene Center
Ludwig-Maximilians-University
Feodor-Lynen-Str. 25
81377 Munich

March, 27th – April, 13th, 2017

Prof. Dr. K.-P. Hopfner hopfner@genzentrum.lmu.de
Prof. Dr. R. Beckmann beckmann@genzentrum.lmu.de

Coordinator
Dr. D. Kostrewa kostrewa@genzentrum.lmu.de

Schedule First Week

Date	Morning	Afternoon
Monday, 27.3.2017	Introduction & Aims Protocols Flowchart of Structure Solution Introduction to Crystallisation Setup of Crystallisation Plates	Introduction to Unix/Linux Bioinformatics Tools
Tuesday, 28.3.2017	Introduction to Crystal Mounting Check Crystallisation Plates & Fish Crystals	Introduction to Radiation Damage Introduction to Data Processing Processing of MAD Data Sets
Wednesday, 29.3.2017	Introduction to the Patterson Function Introduction to Isomorphous Phasing & Density Modification Heavy Atom Substructure Solution Calculation of Initial MAD phases	Model Building
Thursday, 30.3.2017	Introduction to Electron Density Maps & Refinement Model Building & Refinement	Model Building & Refinement
Friday, 31.3.2017	Anomalous Difference Density Map Preparation of Table & Figures	Preparation of Table & Figures

Table of Contents

1 Introduction

In this three-week course, you will carry out all major steps of a modern protein crystal structure determination. You will come across most of the techniques used in modern protein X-ray crystallography and in Cryo-EM.

The X-ray crystallographic structure determination for the first week is based on the following publication:

- G. Lipps, A. Weinzierl, G. von Scheven, C. Buchen, P. Cramer: "Structure of a bifunctional DNA primase/polymerase", *Nature Struct. Mol. Biol.*, 11, 157-162 (2004)

1.1 Background

Genome replication generally requires primases, which synthesize an initial oligonucleotide primer, and DNA polymerases, which elongate the primer. Primase and DNA polymerase activities are however combined in novel replicases from archaeal plasmids, such as pRN1 from *Sulfolobus islandicus*. The crystal structure of the pRN1 primase/polymerase (prim/pol) domain shows a central depression lined by conserved residues. Mutations on one side of the depression reduce DNA affinity. On the opposite side of the depression cluster three acidic residues and a histidine, which are required for primase and DNA polymerase activity. One acidic residue binds a manganese ion, suggesting a metal-dependent catalytic mechanism. The structure shows no similarity to DNA polymerases, but is distantly related to archaeal/eukaryotic primases, with corresponding active site residues. It is proposed that archaeal/eukaryotic primases and the prim/pol domain have a common evolutionary ancestor, a bifunctional replicase for small DNA genomes.

Replication of cellular DNA requires helicases to unwind DNA, primases to synthesize short RNA oligonucleotide primers, and DNA polymerases that elongate the primers. Primases fall in three classes, archaeal/eukaryotic, bacterial/dnaG-type, and viral primases. Structures of primases from the archaeon *Pyrococcus furiosus* (*Pfu*), and from *E. coli* differ strongly from each other and are unrelated to known DNA polymerase structures. Structures of DNA polymerases of families A, B and Y have revealed a resemblance to a right hand, with fingers, thumb and palm subdomains. The fingers and thumb domains are structurally diverse, but the fold of the palm domain is conserved, except for the DNA polymerase family X.

Recently, the replication protein ORF904 of the archaeal plasmid pRN1 was proposed to be the founding member of a novel DNA polymerase family. ORF904 harbours both primase and polymerase activity in its N-terminal region, and a helicase activity in its C-terminal region. Primase and DNA polymerase activities prefer deoxynucleoside triphosphates (dNTPs) as substrates, enabling the protein to polymerize DNA *de novo*. The sequence of the N-terminal region of ORF904 is unrelated to known primases or polymerases.

2 Bioinformatics Tools

Biochemical or biophysical studies on a protein of interest always start with "database mining", a search for published data about the protein of interest with the help of databases and various internet tools. In this course, a couple of hours are devoted to get to know the most important databases and how to use them. Important data about the prim/pol domain will be extracted and correlated with the experimental results. These operations can be done at any computer with the help of internet-browsers (e.g. Firefox). Print out the most important results for the structural analysis. Feel free to use the time to explore the available bioinformatics tools for the analysis of additional proteins of your choice! Most bioinformatics tools are available through our Gene Center Toolkit website http://toolkit.genzentrum.lmu.de.

2.1 Learning goals

Using the prim/pol protein sequence, you will use databases and bioinformatics tools for finding homologous protein sequences, aligning protein sequences for comparison, and assigning putative biochemical functions. You will predict the secondary structure and disordered regions of that protein, and get an idea about domain organization. By application of these toolboxes to other proteins of your choice you can deepen your knowledge and you can develop a feeling when to use which tool.

2.2 Protein sequences – UniProt

The most extensive and best database for protein sequences is UniProt, available at http://www.uniprot.org. It is a merger of the former Swiss-Prot, TrEMBL and PIR databases.

"The Universal Protein Resource (UniProt) is a comprehensive resource for protein sequence and annotation data. The UniProt databases are the UniProt Knowledgebase (UniProtKB), the UniProt Reference Clusters (UniRef), and the UniProt Archive (UniParc). The UniProt Metagenomic and Environmental Sequences (UniMES) database is a repository specifically developed for metagenomic and environmental data.

UniProt is a collaboration between the European Bioinformatics Institute (EBI), the Swiss Institute of Bioinformatics (SIB) and the Protein Information Resource (PIR). Across the three institutes

close to 150 people are involved through different tasks such as database curation, software development and support."

Exercise: Find the UniProt data entry for the large subunit of DNA Primase from *Pyrococcus furiosus*. Save the amino acid sequence in FASTA format into a text document.

2.3 Sequences and alignments – NCBI

The central database for molecular biology in the USA is the National Center for Biotechnology Information (NCBI) available at http://www.ncbi.nlm.nih.gov/. The NCBI offers fantastic tools about proteins, genes, literature search and biological functions. The authors regard themselves as following:

"Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease."

At the NCBI website, you can run BLAST (Basic Local Alignment Search Tool) to identify similar protein or DNA sequences in the databases. This is especially useful when a new protein sequence has been found and the question arises if there are known similar proteins with similar functions.

Exercise: Go to the NCBI website and search the "Protein" database for the following sequences according to the GenBank accession numbers: 10954551 (pRN1 ORF904), 11137542 (pHEN7 homolog), 10954590 (pRN2 homolog). Please search also for the sequence of the distantly related bacteriophage protein Sfi11 (9634995). Save all these sequences in FASTA format. For each sequence, perform a BLAST search under "Analyze this sequence" → "Run BLAST".

Which proteins with sequence similarity do you find? Which part(s) of the protein show high sequence similarity?

2.4 Multiple sequence alignments – ClustalW & T-Coffee

A good algorithm for the alignment of protein (and DNA) sequences is ClustalW (Clustal Omega). The authors wrote (Nucleic Acids Research, 22, 4673-4680):

"The simultaneous alignment of many nucleotide or amino acid sequences is now an essential tool in molecular biology. Multiple alignments are used to find diagnostic patterns to characterise protein families; to detect or demonstrate homology between new sequences and existing families

of sequences; to help predict the secondary and tertiary structures of new sequences; to suggest oligonucleotide primers for PCR; as an essential prelude to molecular evolutionary analysis. The majority of automatic multiple alignments are now carried out using the "progressive" approach of Feng and Doolittle. The sensitivity of the commonly used progressive multiple sequence alignment method has been greatly improved for the alignment of divergent protein sequences."

A more recent algorithm for multiple sequences is T-Coffee, which is described on Wikipedia:

"T-Coffee (Tree-based Consistency Objective Function For alignment Evaluation) is a multiple sequence alignment software using a progressive approach.[1] It generates a library of pairwise alignments to guide the multiple sequence alignment. It can also combine multiple sequences alignments obtained previously and in the latest versions can use structural information from PDB files (3D-Coffee). It has advanced features to evaluate the quality of the alignments and some capacity for identifying occurrence of motifs (Mocca). It produces alignment in the aln format (Clustal) by default, but can also produce PIR, MSF and FASTA format. The most common input formats are supported (FASTA, PIR)."

Both Clustal Omega and T-Coffee are available at our Gene Center Toolkit website http://toolkit.genzentrum.lmu.de.

Exercise: Perform a multiple sequence alignment of the three prim/pol protein sequences with ClustalW and T-Coffee. Copy the 300 N-terminal amino acids of the three sequences in FASTA format into the query windows. Are there any differences in the results? How does the alignment change if the first 40 amino acids are removed? Identify critical residues for DNA-binding and activity according to the publication by Lipps et al. Are these amino acids conserved? Mark the critical residues for metal binding (zinc, manganese). Perform an alignment with the prim/pol domain of pRN1 and the distantly related protein Sfi11. What is the degree of conservation? Which critical amino acids are conserved, which are not? Try to get an alignment of the prim/pol domain of pRN1 with the Pfu primase. Do you find any homology?

2.5 Secondary structure prediction – PSIPRED

PSIPRED is one of the best secondary structure prediction programs.

"PSIPRED is a simple and accurate secondary structure prediction method, incorporating two feed-forward neural networks which perform an analysis on output obtained from PSI-BLAST (Position

Specific Iterated - BLAST). Using a very stringent cross validation method to evaluate the method's performance, PSIPRED 2.6 achieves an average Q3 score of 80.7%."

PSIPRED is available at the website http://bioinf.cs.ucl.ac.uk/psipred/.

Exercise: Perform a secondary structure prediction of the pRN1 prim/pol. Please compare the secondary structure prediction with the real secondary structure (see Fig. 1 in the Lipps et al. paper). Why is the prediction of α-helices usually better than the prediction of β-sheets (think about the 3D structure)?

2.6 Prediction of disorder - DisEMBL

A great challenge in the proteomics and structural genomics era is to predict protein structure and function, including the identification of those proteins that are partially or wholly unstructured. Disordered regions in proteins often contain short linear peptide motifs (e.g. SH3-ligands and targeting signals) that are important for protein function.

"DisEMBL is a computational tool for prediction of disordered/unstructured regions within a protein sequence. As no clear definition of disorder exists, we have developed parameters based on several alternative definitions, and introduced a new one based on the concept of "hot loops", i.e. coils with high temperature factors. Avoiding potentially disordered segments in protein expression constructs can increase expression, foldability and stability of the expressed protein. DisEMBL is thus useful for target selection and the design of constructs as needed for many biochemical studies, particularly structural biology and structural genomics projects."

DisEMBL is available at the website http://dis.embl.de/.

Exercise: Submit the prim/pol domain, including the N-terminus, to the disorder-prediction DisEMBL (disable Java output). You can download the PostScript file and convert it to a pdf file with the command `"ps2pdf"`. We know that the first 40 residues of ORF904 are disordered. Are they predicted to be disordered? Which criterium predicts the disordered N-terminus best?

2.7 Modular proteins, domains & structure prediction – Pfam, HHpred, HHblits

Large proteins are frequently modular, i.e. they consist of different domains with distinct structures and functions. Classical transcription factors for example consist of a DNA-binding domain and an activation domain. An excellent tool to predict the domain architecture from the protein sequence is Pfam.

"The Pfam database is a large collection of protein families, each represented by multiple sequence alignments and hidden Markov models (HMMs). Proteins are generally composed of one or more functional regions, commonly termed domains. Different combinations of domains give rise to the diverse range of proteins found in nature. The identification of domains that occur within proteins can therefore provide insights into their function."

Pfam is available at http://pfam.sanger.ac.uk/. Important: Pfam can only predict domains that have been annotated already!

Another sensitive tool for domain and structure prediction is our Gene Center tool HHPred, available at http://toolkit.genzentrum.lmu.de/hhpred.

"The primary aim in developing HHpred was to provide biologists with a method for sequence database searching and structure prediction that is as easy to use as BLAST or PSI-BLAST and that is at the same time much more sensitive in finding remote homologs. In fact, HHpred's sensitivity is competitive with the most powerful servers for structure prediction currently available.

HHpred is the first server that is based on the pairwise comparison of profile hidden Markov models (HMMs). Whereas most conventional sequence search methods search sequence databases such as UniProt or the NR, HHpred searches alignment databases, like Pfam or SMART. This greatly simplifies the list of hits to a number of sequence families instead of a clutter of single sequences. All major publicly available profile and alignment databases are available through HHpred."

The most recent tool for protein structure prediction is our Gene Center tool HHblits, available at http://toolkit.genzentrum.lmu.de/hhblits.

“HHblits is the first iterative method based on the pairwise comparison of profile Hidden Markov Models. In benchmarks it achieves better runtimes than other iterative sequence search methods such as PSI-BLAST or HMMER3 by using a fast prefilter based on profile-profile comparison. Furthermore, HHblits greatly improves upon PSI-BLAST and HMMER3 in terms of sensitivity/selectivity and alignment quality.“

Exercises: Predict the domain architecture and structure of the three prim/pol protein sequences from the NCBI database. Which domains are predicted? Which amino acids belong to the prim/pol domains?

2.8 Literature search – PuBMed

The complete original scientific biomedical literature can be found at PubMed.

"PubMed comprises over 20 million citations for biomedical literature from MEDLINE, life science journals, and online books. PubMed citations and abstracts include the fields of medicine, nursing, dentistry, veterinary medicine, the health care system, and preclinical sciences. PubMed also provides access to additional relevant Web sites and links to the other NCBI molecular biology resources.

PubMed is a free resource that is developed and maintained by the National Center for Biotechnology Information (NCBI), at the U.S. National Library of Medicine (NLM), located at the National Institutes of Health (NIH)."

PubMeD is available at http://www.ncbi.nlm.nih.gov/pubmed/.

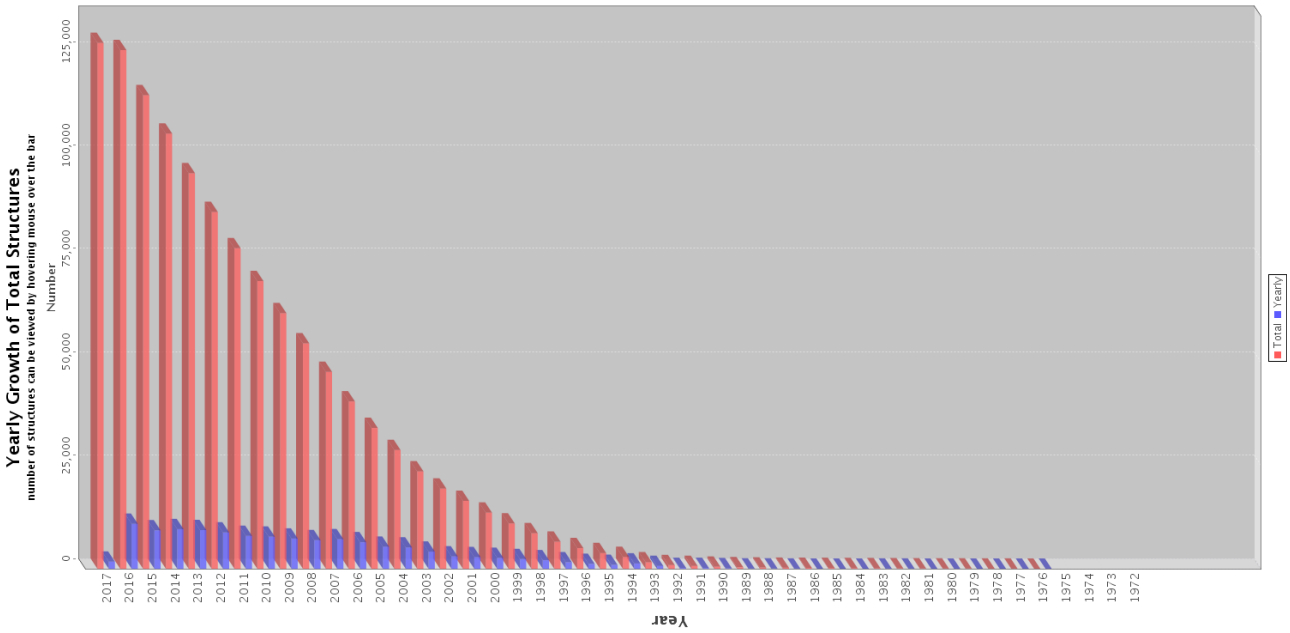
Exercise: Find the original publication about the cloning and characterization of the pRN1 prim/pol domain from Lipps and colleagues (EMBO Journal 2003). Find the original publication about the elucidation of the structure of the archaeal DNA primase from Augustin and colleagues (Nat. Struct. Biol. 2001). Please notice the links to related articles on the right. Find and list 2-3 recent reviews about the structure and function of DNA polymerases and primases.

2.9 Three-dimensional structures – PDB

In the Protein Data Base (PDB) all three-dimensional structures of macromolecules, which were determined by X-ray crystallography or nuclear magnetic resonance (NMR), are archived.

"The PDB archive contains information about experimentally-determined structures of proteins, nucleic acids, and complex assemblies. As a member of the wwPDB, the RCSB PDB curates and annotates PDB data according to agreed upon standards. The RCSB PDB also provides a variety of tools and resources. Users can perform simple and advanced searches based on annotations relating to sequence, structure and function. These molecules are visualized, downloaded, and analyzed by users who range from students to specialized scientists."

By spring of 2017, the PDB (www.pdb.org) contains more than 125000 three-dimensional structures of macromolecules. The following graph shows the growth of deposited structures since 1972.



A structure file in the PDB consists of the header, which includes all information on the experimental conditions during structure determination, and the actual structural data in form of atom records with coordinates x,y,z, and other parameters like occupancy and B-factor.

Exercise: Find the 3D structures of the prim/pol domain and the Pfu DNA Primase. How many prim/pol structures are available? Find the following information for each structure: Which method was used to solve the structure? To which resolution and R-factors was the structure refined? Which hetero-atoms are present in the structure? Download the PDB-files for the prim/pol structures, and open them in an editor. Read the header lines to get an impression which kind of information is stored along with the atomic coordinates.

2.10 Calculation of simple molecule parameters - ProtParam

There are many tools available to calculate different parameters such as molecular weight or the theoretical isoelectric point of a protein from the sequence information. Different tools calculate the molecular weight differently, therefore significant differences can occur. Often mass spectrometry is more accurate. ProtParam (http://web.expasy.org/protparam/) is a useful tool from ExPASy.

"The protein pI is calculated using pK values of amino acids, which were defined by examining polypeptide migration between pH 4.5 to 7.3 in an immobilised pH gradient gel environment with 9.2 M and 9.8 M urea at 15º C or 25º C. Prediction of protein pI for highly basic proteins is yet to be studied and it is possible that current predictions may not be adequate for this purpose. The buffer capacity of a protein will affect the accuracy of its predicted pI, with poor buffer capacity

leading to greater error in prediction. Because of this, pI predictions for small proteins can be problematic. Protein MW is calculated by the addition of average isotopic masses of amino acids in the protein and the average isotopic mass of one water molecule."

Exercise: Calculate the number of amino acids, the molecular weight and the isoelectric point for the prim/pol full length amino acid sequence, for the N-terminal 255 amino acids, and for the fragment that is described in the Lipps et al. paper.

2.11 Detection of similar folds – Dali

The Dali server (http://www.bioinfo.biocenter.helsinki.fi/dali_server/start) is a service for comparing protein structures in 3D. You submit the coordinates of a query protein structure and Dali compares them against those in the PDB. A multiple alignment of structural neighbours is returned. In favourable cases, comparing 3D structures may reveal biologically interesting similarities that are not detectable by comparing sequences. If you want to know the structural neighbours of a protein already in the Protein Data Bank, you can find them in the FSSP database.

The Dali site compares a structure that you have solved against the database of all known folds and sends back (within hours–days) a list of similar 3D structures, with scores, root mean square deviations of Cα atoms in Å and rotation and translation matrices that are required for superposition of the database structure onto your structure – so you can compare them all in 3D!

Exercise: Submit the native prim/pol domain structure that you have just downloaded from the PDB database to the Dali server to find similar protein folds. Which structurally similar folds are found?

3 Crystallization & Cryo-Crystallography

Crystals are necessary for structure determination by X-ray crystallography. Crystals are three-dimensional ordered structures that can be described by a repetition of an identical unit cell. Through the large number of unit cells in a crystal, the X-ray diffraction signal is amplified such that it can be detected and measured. Obtaining diffraction quality crystals is still a major obstacle in protein crystallographic research.

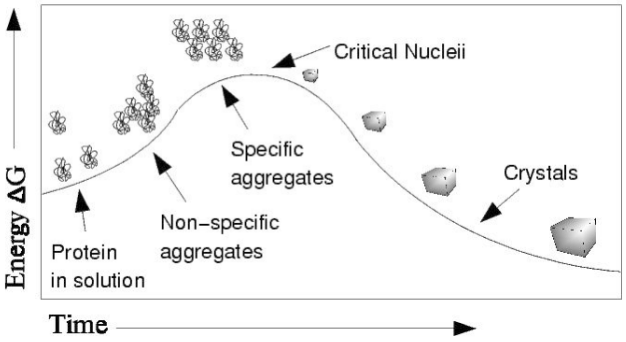
Cryo-crystallography is the main method to preserve a macromolecular crystal as long as possible during data collection by reducing the effects of radiation damage.

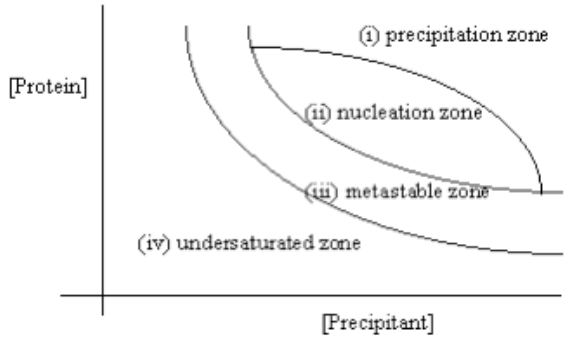
3.1 Learning goals

You will check crystallization plates for crystals and learn how to "fish" crystals with a small nylon loop for cryo-cooling in liquid nitrogen for future data collection under cryo-conditions.

3.2 Crystallization theory

Crystallisation in general is a thermodynamically driven reaction. The driving force for crystallisation is that a crystalline arrangement has a lower free energy than an irregular arrangement. But in order for a protein to crystallize it must overcome an energy barrier analogous to that for conventional chemical reactions. The critical nucleus, the starting point of crystallisation, corresponds to the higher energy intermediate.





Protein crystallization occurs when the concentration of protein in solution is greater than its solubility limit, *i.e.*, the protein solution is supersaturated. The probability of nucleation increases proportional to the degree of supersaturation. A too high supersaturation leads to precipitation. This

can be represented in a phase diagram by dividing the supersaturated zone into regions of increasing probability of nucleation and precipitation.

3.3 Crystallization methods

Crystallization experiments often go through two distinguishable phases: a screening step, which is usually followed by an optimization step. Screening establishes which conditions produce promising crystals, and optimization refers to the fine-tuning of those initial conditions to obtain well-diffracting crystals for data collection.

Some basic things you should check before starting

Purity of the protein solution is the most important prerequisite for crystallizability, because contaminants lead to crystal growth defects. The protein in solution should be mono-disperse, which means that the protein is free of non-specific oligomers and aggregates (→ dynamic light scattering, gel filtration).

Other things to consider: Is the protein correctly folded (activity tests)? Is it fresh (proteins break down with time and the mixture becomes heterogeneous)? Does the protein need to be kept reduced (cysteins)? Does the protein need to be stabilized by additives (*e. g.* salt) to stay in solution?

The most important methods of growing protein crystals

Vapour Diffusion (Sitting and Hanging Drop Method)

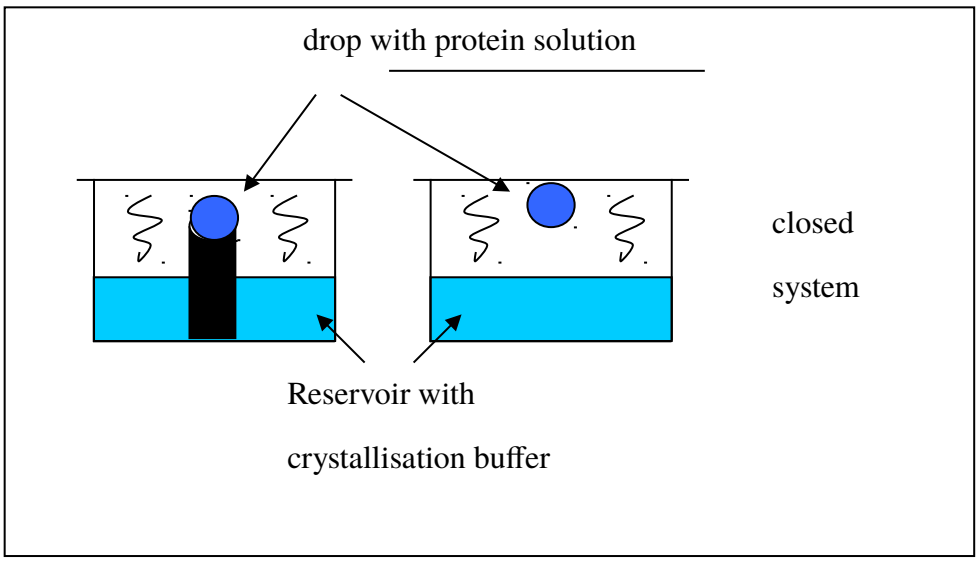
This is probably the most common way to grow crystals. A drop of protein solution is suspended over a reservoir containing buffer and a higher concentration of precipitant. To reach vapour pressure equilibrium, water diffuses from the drop to the reservoir solution, increasing the supersaturation in the drop, ideally leading to optimal crystal growth conditions.

Microbatch crystallization

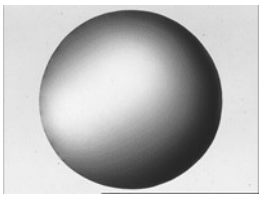
A drop with a mixture of protein solution and precipitant is put under inert oil. Eventually, crystallization nuclei form and crystals grow. One potential problem is an unwanted interaction of the protein with the oil phase.

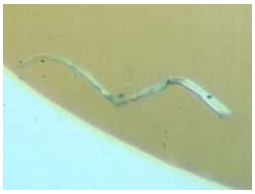
Dialysis

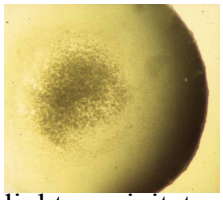
Here, the protein solution and a solution containing the precipitant are separated by a semi-permeable membrane. Supersaturation is achieved by diffusion of the precipitant into the protein solution.

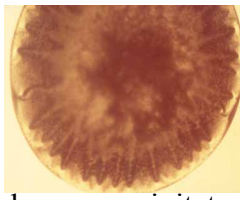


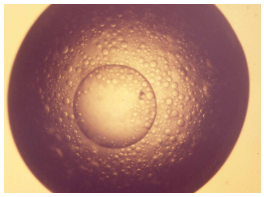
Typical crystal growth takes between several hours and several weeks. To check whether crystals grow, you should control the crystallization drops with a microscope over time. The following pictures give you some examples of what you may see:

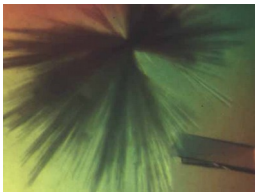

clear drop


fibre

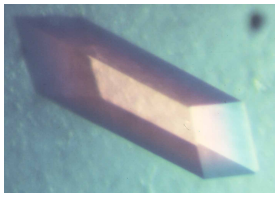

light precipitate


heavy precipitate


phase separation


needles


2d crystal plates


3d crystal

3.4 Optimization – screening around the initial crystallization conditions

Once a promising initial crystallization condition is found, attempts are made to improve the crystallization to get well-diffracting crystals. This is done by screening around the initial

crystallization conditions. This can include varying the precipitant concentration, pH, protein concentration, temperature and crystallization method (e.g. sitting/ hanging drop, drop size, ...). Another optimization attempt is to add additives like glycerol, divalent cations (Mg^{2+},..), alcohols, sugars, … further one can try seeding methods to obtain bigger crystals. The practical part is done with lysozyme.

3.5 Practical part

Stock solutions (provided):

• 100 mg/ml lysozyme in 50 mM sodium acetate (NaAc) buffer pH 4.5

• 500 mM NaAc buffer pH 4.5

• 4 M NaCl

• 50% PEG MME 5000

We will use a two-dimensional grid matrix, varying the NaCl concentration in the columns as shown in the table below, and the PEG concentration in the rows. Please, prepare 1 ml of the following crystallization conditions using the provided stock solutions:

	1	2	3	4	5	6
A	50 mM NaAc pH 4.5 0.6 M NaCl	50 mM NaAc pH 4.5 0.8 M NaCl	50 mM NaOAc pH 4.5 1.0 M NaCl	50 mM NaOAc pH 4.5 1.2 M NaCl	50 mM NaOAc pH 4.5 1.4 M NaCl	50 mM NaOAc pH 4.5 1.6 M NaCl
B	50 mM NaOAc pH 4.5 0.6 M NaCl 10% PEG MME 5000	50 mM NaOAc pH 4.5 0.8 M NaCl 10% PEG MME 5000	50 mM NaOAc pH 4.5 1.0 M NaCl 10% PEG MME 5000	50 mM NaOAc pH 4.5 1.2 M NaCl 10% PEG MME 5000	50 mM NaOAc pH 4.5 1.4 M NaCl 10% PEG MME 5000	50 mM NaOAc pH 4.5 1.6 M NaCl 10% PEG MME 5000
C	50 mM NaOAc pH 4.5 0.6 M NaCl 20% PEG MME 5000	50 mM NaOAc pH 4.5 0.8 M NaCl 20% PEG MME 5000	50 mM NaOAc pH 4.5 1.0 M NaCl 20% PEG MME 5000	50 mM NaOAc pH 4.5 1.2 M NaCl 20% PEG MME 5000	50 mM NaOAc pH 4.5 1.4 M NaCl 20% PEG MME 5000	50 mM NaOAc pH 4.5 1.6 M NaCl 20% PEG MME 5000
D	50 mM NaOAc pH 4.5 0.6 M NaCl 30% PEG MME 5000	50 mM NaOAc pH 4.5 0.8 M NaCl 30% PEG MME 5000	50 mM NaOAc pH 4.5 1.0 M NaCl 30% PEG MME 5000	50 mM NaOAc pH 4.5 1.2 M NaCl 30% PEG MME 5000	50 mM NaOAc pH 4.5 1.4 M NaCl 30% PEG MME 5000	50 mM NaOAc pH 4.5 1.6 M NaCl 30% PEG MME 5000

The screening will be done by the hanging drop vapour diffusion method as follows:

1. Fill 500 µl of the crystallization buffer into the reservoir
2. Mix 4 µl of the protein solution with 4 µl crystallization buffer from the reservoir on the cover slide
3. Gently put the cover slides on their respective reservoirs to form a closed system
4. Check the crystallization setups under the microscope now and after one day

3.6 Cryo-crystallography

One of the obstacles to structure determination of macromolecules by X-ray crystallography is radiation-induced crystal decay. Crystal decay appears to be a result of X-ray radiation forming free radicals, and of thermal damage to the macromolecule and, subsequently, to the lattice. Radiation and thermally induced crystal decay can be considerably reduced by collecting data from crystals that are cryo-cooled to temperatures around 100 K.

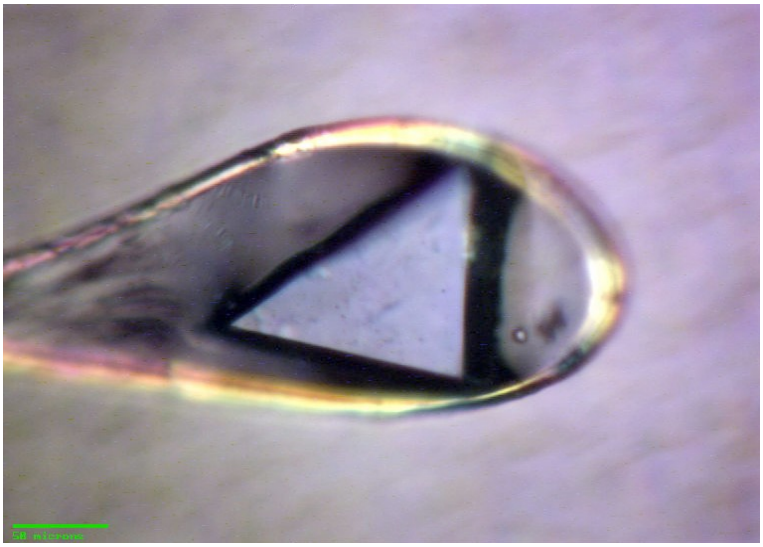
The speed of crystal-cooling is crucial for the success of any cryo-technique. Slowly lowering the temperature causes formation of ice crystals within or around the protein crystal lattice. Ice formation degrades the diffraction from the protein lattice in two ways: by contributing to the diffraction pattern and by shearing the crystal lattice. To prevent ice formation, rapid flash-cooling of the crystal is needed. The mounted crystal is either plunged into liquid nitrogen or quickly brought into a cold nitrogen stream. Flash-cooling causes the aqueous solution in and around the crystal to go into an amorphous glass-like form, a process known as vitrification, instead of forming ice-crystals. To increase the rate of success of flash-cooling, usually, cryo-protectants must be introduced to the mother liquor. The cryo-protectant disturbs ice formation in the mother liquor by slowing the nucleation of ice-crystals and raising the viscosity of the solution. Ideally, the cryo-protectant should not increase the crystal mosaicity. Typical cryo-protectants are glycerol, alcohols, low molecular weight PEGs or sugars. Typical concentrations to prevent ice formation range from 5% to 50% (v/v).

When searching for a cryo-protectant, one should first consider the mother liquor composition. In general, the best choice of cryo-protectant will be one that most closely resembles the composition of the mother liquor. For instance, if the crystal grows in PEG, then ethylene glycol, the "monomer" of PEG, would be a good first choice. One may then try glycerol and even low molecular weight PEG. If the crystal grows in low MPD concentration, a higher MPD concentration would be the first option and so on.

3.7 Crystal mounting with loops

In order to collect diffraction data using cryo-crystallography, the crystals have to be mounted with small nylon loops. The pre-mounted crystals can be stored under liquid nitrogen until they are transferred for measurement from the liquid nitrogen to the nitrogen cryo-stream with special tools like cryo-tongs.

If no suitable cryo-conditions for a crystal can be found, or if a control-measurement at room temperature is required, the crystal can be mounted with special loops where small transparent plastic-hoods protect the crystal from drying-out.



3.8 Practical part

You will check your crystallisation plates for suitable crystals and take some nice pictures for the protocol. You will then practice to fish crystals with small nylon loops, and cryo-cool them in liquid nitrogen.

4 Radiation Damage & Data Collection Strategy

4.1 Learning goals

You will have a look at a typical home source X-ray generator (with X-rays switched off), an imaging plate detector, and a typical protein diffraction pattern. You will learn the basic principles of radiation damage and how to minimize it during data collection.

4.2 Data collection parameters

If you have grown well-diffracting crystals, you want to extract as much information with best possible statistics from a diffraction experiment. To this end, the experimenter has to make a number of strategic decisions before collecting diffraction data. These include:

- What X-ray source and which wavelength should be used?
- How long should each image be exposed?
- Which oscillation range per frame gives reasonable number of frames and zero overlaps?
- Which angular range gives maximal completeness?
- Which resolution limit should be applied?

1. **Source and Wavelength** The choice of the x-ray source will depend largely on what is available. A synchrotron source will be required in most cases to collect high-resolution data of protein crystals. The choice of wavelength in the range of 0.9-1.3 Å is restricted to the element-specific absorption-edges when collecting anomalous data. Otherwise, shorter wavelengths have the benefit of reduced air and sample absorption and supposedly induce less radiation damage.
2. **Exposure per frame** This depends on the strength of the x-ray source and the diffraction power of the crystal. Ideally, spots at maximum resolution should have an average intensity of I/sig(I) > 2 and even the strongest spots should not overload the detector. Exposure times of approx. 1 sec per frame are realistic at state-of-the-art synchrotron beamlines.
3. **Oscillation width ΔΦ** If ΔΦ per frame is too wide, the number of spots per frame increases, therefore the likelihood that two spots overlap increases, too. Certain crystal

orientations are particularly prone for overlaps. Additionally, fine ΔΦ-slices need shorter exposure times per frame, reducing the overall background, since background intensity is proportional to exposure time. Typical ΔΦ values range between 0.1° and 0.5° per frame.

4. **Total oscillation range** To minimize data collection time and radiation damage, the total angular range, which gives maximum completeness with the minimal number of frames has to be found. This should be done using the strategy task of the data processing software.

5. **Resolution limits and detector distance** It is advisable to move the detector back as far as possible, so that highest-resolution spots end up close to the edge of the detector. Doing so, the background, which is proportional to 1/distance² can be reduced significantly.

6. **Use cryo-cooling of the crystal!**

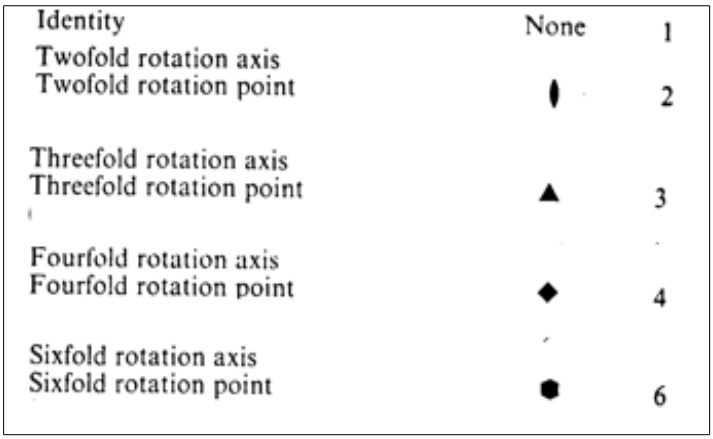
5 Symmetry

Crystals are built by the regular translational repetition in three dimensions of a basic building block, called the *unit cell*. The regular translational repetition is called the *crystal lattice*. Each unit cell can be composed of a single molecule or molecular complex (protein, DNA, …) or by multiple copies of it, which are related by symmetry-operations (see below). The minimal molecule or molecular complex that is needed to build the whole unit cell by symmetry-operations is called *asymmetric unit*.

5.1 Learning goals

You will learn the basic symmetry elements, crystal lattices and some example space groups that occur in macromolecular crystallography.

5.2 Symmetry-elements of crystals



Rotational Symmetry:

Identity	rotation by 360°
2-fold axis	rotation by 180°
3-fold axis	rotation by 120°
4-fold axis	rotation by 90°
6-fold axis	rotation by 60°

No other rotational symmetry could lead to shapes that fill a space completely (try pentagons, heptagons, octagons ...)

Mirror Symmetry

Inversion Symmetry

Protein crystals are built up by chiral molecules (L-amino acids) and therefore can only show rotational symmetry elements, but neither mirror symmetry nor inversion symmetry.

5.3 The seven crystal systems

A three-dimensional unit cell is defined by three axes ***a***, ***b*** and ***c***, and three angles between these unit cell axes ***α***, ***β*** and ***γ,*** defined as follows:

α the angle which is formed between ***b*** and ***c***

β the angle which is formed between ***a*** and ***c***

γ the angle which is formed between ***a*** and ***b***

There are only seven crystal systems that can fill a three-dimensional space:

triclinic: no restrictions on ***a***, ***b***, and ***c***; no restrictions on ***α***, ***β***, and ***γ***

monoclinic: no restrictions on ***a***, ***b*** and ***c***; no restrictions on ***β; α*** = ***γ*** = 90°

orthorhombic: no restrictions on ***a***, ***b*** and ***c***; ***α*** = ***β*** = ***γ*** = 90°

tetragonal: ***a*** = ***b***, no restrictions on ***c***; ***α*** = ***β*** = ***γ*** = 90°

trigonal: ***a*** = ***b,*** no restrictions on ***c***; ***α*** = ***β*** = 90°, ***γ*** = 120°

hexagonal: ***a*** = ***b***, no restrictions on ***c***; ***α*** = ***β*** = 90°***, γ*** = 120°

cubic: ***a*** = ***b*** = ***c***; ***α*** = ***β*** = ***γ*** = 90°

5.4 The 32 crystal classes

If we combine possible symmetry elements with the seven crystal systems, we get 32 crystal classes. Of these, only 11 crystal classes without mirrors and inversion centers are possible for protein crystals:

crystal system	*symbol*	*symmetry*
triclinic:	1	identity
monoclinic:	2	one 2-fold axis parallel ***b***
orthorhombic:	222	three 2-fold axis perpendicular to each other
tetragonal:	4	one 4-fold axis parallel ***c***
	422	one 4-fold axis parallel ***c*** two 2-fold axis: both perpendicular to ***c*** one 2-fold axis parallel ***a***, the other face-diagonal
trigonal:	3	one 3-fold axis parallel ***c***
	32	one 3-fold axis parallel ***c*** one 2-fold axis perpendicular to ***c***

hexagonal:	6	one 6-fold axis parallel ***c***
	622	one 6-fold axis parallel ***c***
		one 2-fold axis parallel ***a***, the other face-diagonal
cubic:	23	one 2-fold axis parallel ***a***
		one 3-fold axis room-diagonal
	432	one 4-fold axis parallel ***a***
		one 3-fold axis room-diagonal
		one 2-fold axis face-diagonal

The symmetry operations along the unit cell axes ***a***, ***b***, and ***c*** act on the atomic positions within a unit cell. Atomic positions are given relative to the unit cell axes as *fractional coordinates* (x, y, z), each ranging from 0 to 1 within a unit cell.

Let's assume that we have only a twofold axis parallel to ***b***. This twofold symmetry creates for an atom at position (x, y, z) a symmetry-equivalent atom at position (-x, y, -z):

(x, y, z) → (-x, y, -z)

Consequently, the size of the asymmetric unit in a monoclinic crystal is 1/2 of the volume of the unit cell.

As an exercise, find all symmetry-equivalent positions of (x, y, z) for point group 222.

How big is the size of the asymmetric unit?

5.5 Translational symmetries

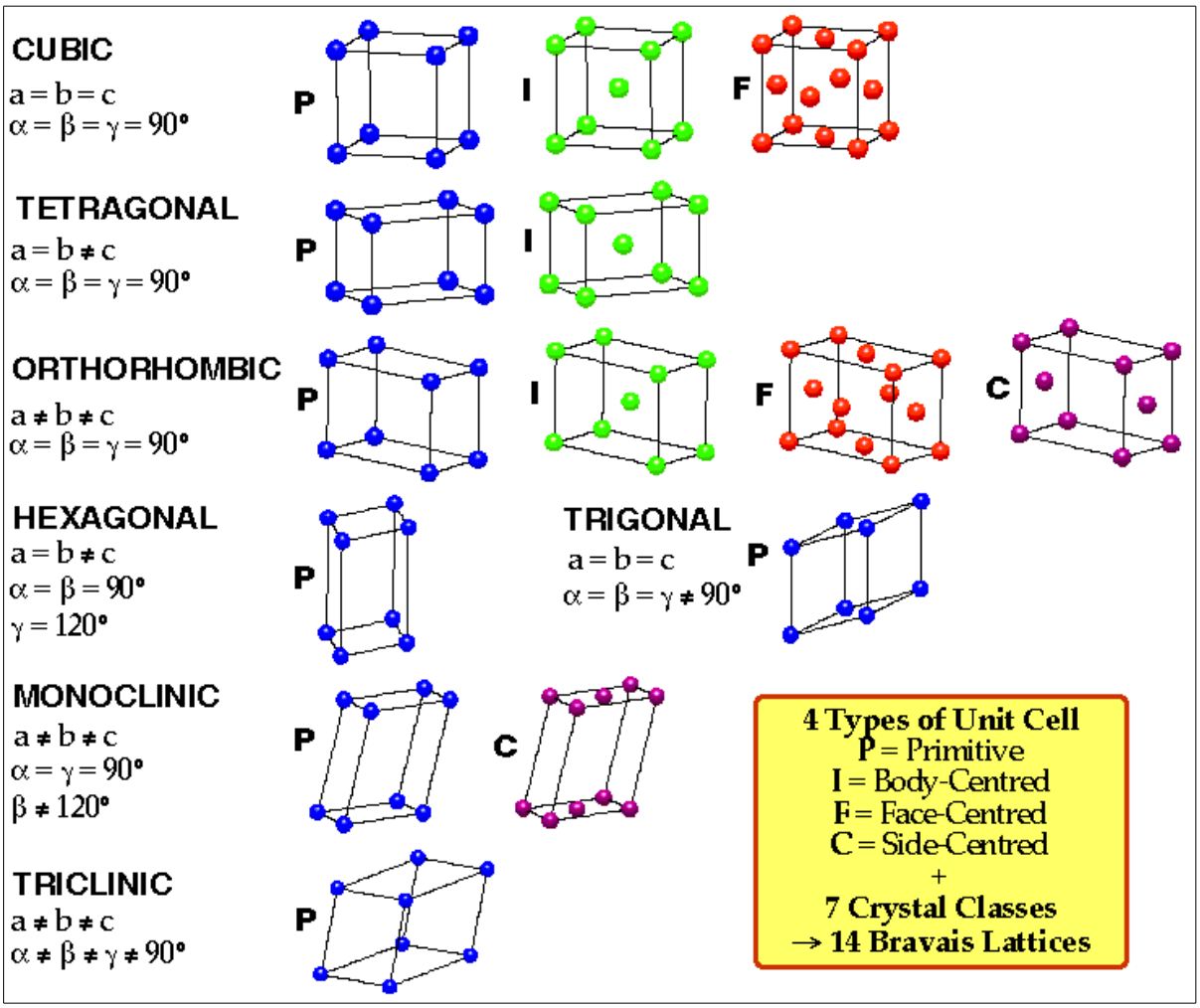
In a crystal, there are different possible ways of regular translational repetitions of the asymmetric unit, resulting in different lattice types. In a *primitive lattice*, the only translational repetitions are along ***a***, ***b*** and ***c***. Therefore, the lattice points lie only on the corners of the unit cell. In non-primitive lattices, additional lattice points exist. They are summarized as the five different *Bravais centers*:

Bravais center:

type	*symbol*	*additional translational operation*
primitive:	*P*	none

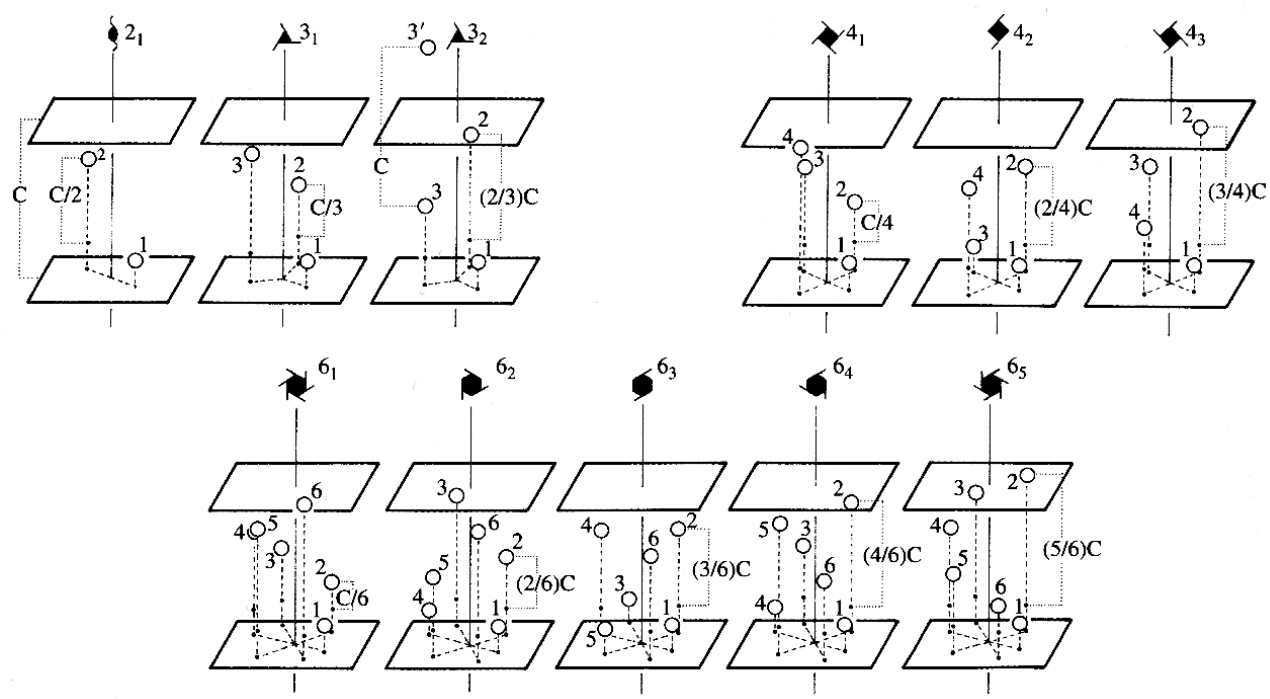
body centered:	*I*	translation of 1/2 along ***a***, ***b*** and ***c***: (x,y,z) → (x+1/2, y+1/2, z+1/2)
face centered:	*C*	translation of 1/2 along ***a*** and ***b***: (x,y,z) → (x+1/2, y+1/2, z) C is the face perpendicular to ***c; i***n principle, face centers *A* and *B* would also exist, but they can be transformed to *C*
	F	*A*, *B*, and *C* centered at once translation of 1/2 along ***a*** and ***b***: (x,y,z) → (x+1/2, y+1/2, z) translation of 1/2 along ***a*** and ***c***: (x,y,z) → (x+1/2, y, z+1/2) translation of 1/2 along ***b*** and ***c***: (x,y,z) → (x, y+1/2, z+1/2)
hexagonal centered	*H*	translational components of: (x+2/3, y+1/3, z+1/3) and of: (x+1/3, y+2/3, z+2/3)

The following 14 Bravais lattices exist:



Screw axes

Screw axes are combined rotational and translational symmetry operations along the rotation axis. The translational component has to be an integer fraction of the rotation order. For instance, a 4-fold axis can have a translational component of 1/4, 1/2 (2/4), and 3/4. The standard notation of the corresponding screw axes is 4_1, 4_2 and 4_3, respectively, with the fractional translation component notated as a subscript. The following screw axes exist:



If we look, for example, at 4_1 and 4_3, we see, that they are mirror images of each other. 4_1 shows a right hand turn, 4_3 a left hand turn. They are *enantiomorphous* to each other. An inversion operation would transform one into the other. Now try and find the other enantiomorphous pairs.

5.6 Space groups

If all symmetry operators with all translational components are combined, we get totally 230 different *space groups*. If we combine only rotational and translational symmetries, we get 65 space groups without mirrors or inversion centers. Only these space groups are possible in protein crystallography.

Each space group is characterized by a Hermann-Mauguin symbol which is defined as follows: there are four character positions: first one letter code for the lattice type, then three number codes for the rotational components along different directions.

Examples:

P422

The first letter code characterizes the primitive Bravais center. The three numbers characterize the rotational symmetries, a 4-fold axis and two 2-fold axes. As you are already familiar with the point groups, you will recognize the tetragonal point group 422. Further we have no screw axis, because no translation period is stated as subscript. In case we would have a 4_1 screw axis parallel ***c***, the Hermann-Mauguin symbol would be

$P4_122$

Remember, the point group is still 422 as point groups do not include any translational components.

$P4_322$

would be the enantiomorphous space group.

Other examples:

$P6_322$

$P2_12_12$

Here are some conventions that you should remember:

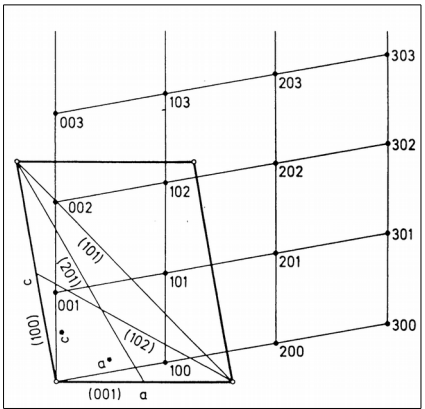
1. *monoclinic:* Usually, the so-called 2nd setting is used. This means, that the 2-fold axis has to be parallel to the ***b*** axis.

2. *orthorhombic:* Only the *C* center is allowed according to conventions. If one axis is different to the other axes, then the unique axis has to be the ***c*** axis and the rule $\boldsymbol{a} < \boldsymbol{b} < \boldsymbol{c}$ is not valid, anymore. For instance $P2_122$ is not a conventional setting, use $P222_1$, instead. The same is true for $P2_12_12$.

3. *higher systems:* the highest symmetry has always to be parallel to the ***c*** axis.

5.7 The reciprocal space

As the real space and diffraction pattern are reciprocal to each other, we will define a reciprocal space that allows to understand a diffraction pattern. From a common origin, draw lines perpendicular to lattice planes. The length of these normals



should be reciprocal to the lattice plane distances. The endpoints will form the basis vector set for the reciprocal space.

Every point, which actually gives rise fore a reflection in the diffraction pattern, is characterized by the Miller indices hkl (note without brackets). *h* gives you the subdivision factor for ***a***, *k* for ***b***, and *l* for ***c***, respectively. During indexing of a diffraction pattern, you assign to every reflection a corresponding Miller index.

Translational components in real space within a unit cell, like for screw axes, lead to systematic absences of reflections in reciprocal space. Using the knowledge about these systematic absences can help you to determine the space group of a protein crystal. One exception are enantiomorphous space groups. According to Friedel's law, the reciprocal space contains an inversion symmetry center. Since enantiomorphous space groups can be transformed into each other by an inversion, you can not distinguish between enantiomorphous space groups from the diffraction pattern in reciprocal space.

5.8 Rules for systematic absences

All systematically absent reflections for an assumed space group should have observed intensities close to zero. Any expected systematically absent reflection with a large intensity contradicts the assumption about this space group. Therefore, this assumption is very likely wrong. Here is a list of expected systematic absent reflections related to the underlying symmetry:

Bravais center

$h + k = 2n$	*C* face centered
$k + l = 2n$	*A* face centered
$h + l = 2n$	*B* face centered
$h + k + l = 2n$	*I* centered
$h + k = 2n$	*F* centered
$h + l = 2n$	
$k + l = 2n$	
$-h + k + l = 3n$	hexagonal centered, obverse setting
$h - k + l = 3n$	hexagonal centered, reverse setting

Screw axis

h00	$h = 2n$	2_1, 4_2

h00	$h = 4n$	$4_1, 4_3$
0k0	$k = 2n$	$2_1, 4_2$
0k0	$k = 4n$	$4_1, 4_3$
00l	$l = 2n$	$2_1, 4_2, 6_3$
00l	$l = 3n$	$3_1, 3_2, 6_2, 6_4$
00l	$l = 4n$	$4_1, 4_3$
00l	$l = 6n$	$6_1, 6_5$

5.9 Space group determination

In principle you should now be able to determine the space group from undistorted diffraction patterns.

Here are some guidelines:

1. Search for the rotational symmetries in the given main sections through the reciprocal space. In these sections, one index of the three Miller indices is constant. Now try to assign the correct point group. Take care to stick to conventions.

2. You assigned a Miller index to every reflection. Now you can check for systematic absences. First for Bravais centers, because you see them all over a image. Later for systematic absences of screw axes. They are always found on a line going through the center (origin).

3. Now you can set the symbol for the Bravais center in front of the point group symbol and assign every screw component to the axes as subscript in the point group symbol. Finally, check if you carefully followed all conventions. If not, you have to rename the directions and thereby you reindex the diffraction pattern.

Well done, you've determined the space group!

6 Derivatisation

Common phasing techniques use the initial phases that are derived from data from heavy atom derivative crystals (SAD, MIR), or anomalous scatterers (MAD). One method for heavy-atom derivatisation is substitution of methionine by seleno-methionine. It offers a general method for introduction of anomalous scatterers into over-expressed proteins. A second method for heavy metal derivatisation is soaking of crystals with appropriate heavy atom compounds.

6.1 Learning goals

You will learn the basic principles of heavy atom derivatisation that can be used in isomorphous replacement phasing techniques.

6.2 Substitution of methionine by seleno-methionine

Preparation of seleno-methionine-containing protein is relatively easy to perform. For producing seleno-methionine-labelled proteins you have to use an *Escherichia coli* strain, which is auxotrophic for methionine. A special medium, e.g. LeMaster medium, which contains seleno-methionine, is then used for expression. Because the *E. coli* strain is auxotrophic for methionine, the bacteria will incorporate the seleno-methionine from the medium into the protein.

Seleno-methionyl proteins are much more sensitive to oxidation than natural proteins. If selenium atoms are on the surface of the protein molecule, they can alter protein hydrophobicity and solubility. Usually, seleno-methionyl proteins are more hydrophobic and less soluble. These properties require some modifications to the normal purification. To avoid oxidation of seleno-methionine all buffers should be degassed. Buffers should include a reducing reagent such as dithiothreitol (DTT) and a chelator such as Ethylene Diamine Tetraacetic Acid (EDTA) to remove traces of metals that could catalyse oxidation.

6.3 Heavy atom derivatives

Derivatization by soaking

Heavy metal derivatives can be prepared by soaking crystals in the appropriate heavy atom containing crystallisation buffer or by co-crystallization.

Heavy atom derivatisation is typically performed by either adding a heavy atom solution to the drop containing the crystal or by transferring a crystal from the mother liquor drop to a stabilizing solution containing the heavy atom. The solvent channels in macromolecular crystals allow heavy atom reagents to diffuse into the crystals and provide access to the protein in the crystal lattice. Heavy atom binding sites can be both on the surface and the interior of proteins.

Typical concentrations of heavy atom range between 0.1 mM to 100 mM depending on the pH, temperature, crystallization reagent and heavy atom. A good starting concentration for the heavy atom compound is 1 mM in the presence of the mother liquor and the crystal. The soaking time for heavy atom binding depends on a number of variables such as the heavy atom, concentration of the heavy atom, solubility of the heavy atom, temperature, and the crystallization reagent. Soaking time can vary from less than one hour to weeks. For initial screening, typically, a few hours or overnight are sufficient.

If, after soaking, the crystal appearance is unchanged from the native crystal and no changes are observed in the diffraction pattern, then one might try to increase the soaking time and the concentration of the heavy atom. If, after soaking, the crystal cracks or the diffraction pattern changes too much (different unit cell constants, lower maximum resolution), then one could try to decrease the heavy atom concentration and use shorter soaking times.

Derivatization by co-crystallization

In some cases, one can first derivatize the protein with the heavy atom and then crystallize it. This procedure is less frequently used since the procedure may alter the crystallization condition or produce crystals that are not isomorphous to the native crystals, anymore, because the binding of the heavy atom may change intermolecular contacts in the crystal lattice. However, the method can be useful when soakings with heavy atoms fail.

Crystallization reagents and heavy atoms

Crystallization reagents (specifically salts) and buffers are a potential ligand source for heavy atoms. Complex formation of buffer reagents with the heavy atom may precipitate the heavy atom or compete with the reaction of the heavy atom with the protein. Heavy atom compounds are usually less soluble in high salt buffers. Hence, high salt concentrations are not ideal for heavy atom reactions with macromolecules. Polyethylene glycol does not react with most heavy atom compounds and is favourable for heavy atom reactions. TRIS, phosphate, citrate, ammonium sulfate

(above pH 6), β-mercaptoethanol, and DTT may interfere with heavy atom binding in some instances.

Classification of heavy atoms

Class A (e.g. Rb, Cs, Mg, Ca, Sr, Ba, Pb, lanthanides, UO_2)

Class A metals have a preference for hard ligands such a carboxylates and other oxygen containing groups. Class A metals do not polarize well and bind electronegative, hard ligands such as F, OH, H_2O, phosphate and carboxylate in an electrostatic fashion. Interactions between the macromolecule and Class A metals are weakened in high ionic strength reagents and mother liquors. Class A metals can form insoluble hydroxides at alkaline pH levels (> 7.5). Phosphates and citrates compete for Class A metals in solution, but NH_3 does not.

Class B (e.g. Pt, Hg, Os, Au, Ag, Ir, Ca)

Class B metals have a preference for soft ligands such as sulfur, nitrogen and halides. The Class B metals are polarizable. They do not work well at low pH since many of their ligands such as histidine and cysteine are protonated at low pH and thus less reactive. However, Pt is an exception since Pt can interact with methionine and disulphide bridges at low pH. In the presence of ammonium sulfate and pH above 7, the Class B metals do not function well since NH_3 will compete for the heavy atom (an exception is Pt). Phosphates and hydroxides are less a problem with Class B metals since they are hard ligands and do not bind tightly to Class B metals.

Class B metals can form stable heavy atom complexes with macromolecules through covalent (Hg), anionic, cationic, or hydrophobic interactions in conditions.

Amino acid reactivity with heavy atoms

Sometimes the choice of heavy atoms, the appropriate pH, and mother liquor composition can be made from examining the amino acid composition of the macromolecule. The following reference is recommended for learning more about the specific reactivities of amino acids with heavy atoms:

- Petsko, G.A., Preparation of Isomorphous Heavy-Atom Derivatives, in Wyckoff, H.W., Hirs, C.H.W. and Timasheff, S.N. ed. (1985). Methods in Enzymology, Vol. 114. Diffraction Methods for Biological Macromolecules Part A. Academic Press, Orlando, pp. 147-156.

Heavy atoms are very toxic, so prepare yourself before you prepare derivatives!

7 Indexing, Integration and Scaling

A diffraction data set forms an image of the three-dimensional reciprocal space. This 3D image is recorded as a series of 2D diffraction images, each of them representing a different, curved slice of reciprocal space. Ideally, a complete dataset, consisting of up to several hundred images (= frames, batches), fully covers the reciprocal space up to a certain maximum resolution.

During the indexing, integration and scaling process, all information about crystal lattice parameters, position and intensity of the reflections is extracted from the set of diffraction images (several GBytes) into a text file containing hkl, intensities and their experimental errors (approx. 1 MByte). Hence, the term "data reduction" is frequently used for this process.

7.1 Learning goals

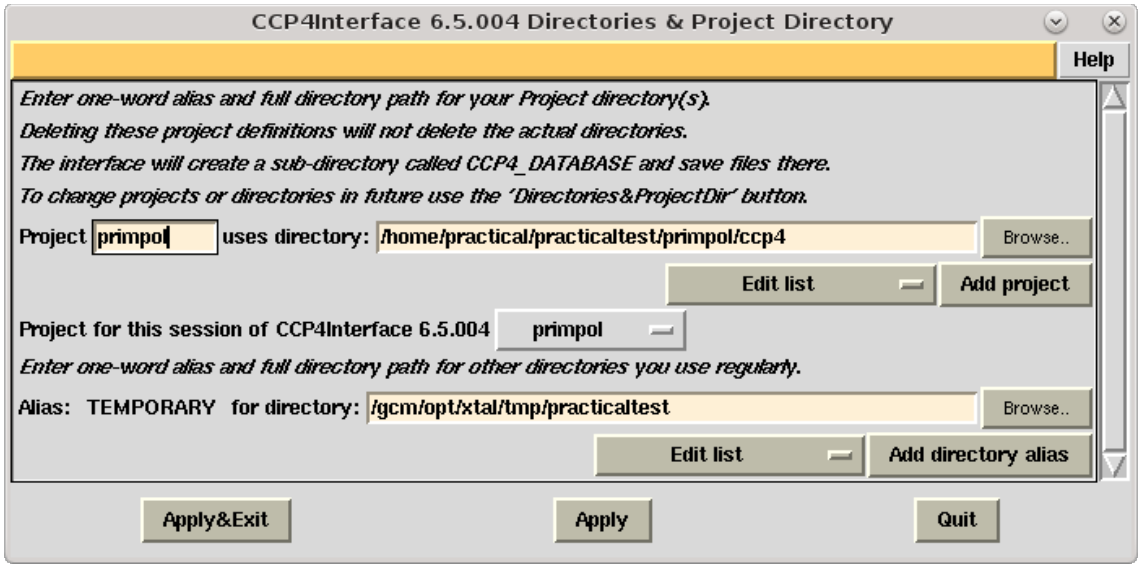
You will define a new project using the CCP4 program suite, which is used for data processing and later for refinement. During data processing, you will assign a lattice to the observed diffraction images (indexing), integrate the observed intensities, determine the proper space group and convert the final intensities to amplitudes. This will be done for four MAD data sets that will be used in phasing.

7.2 Prim/Pol MAD data sets

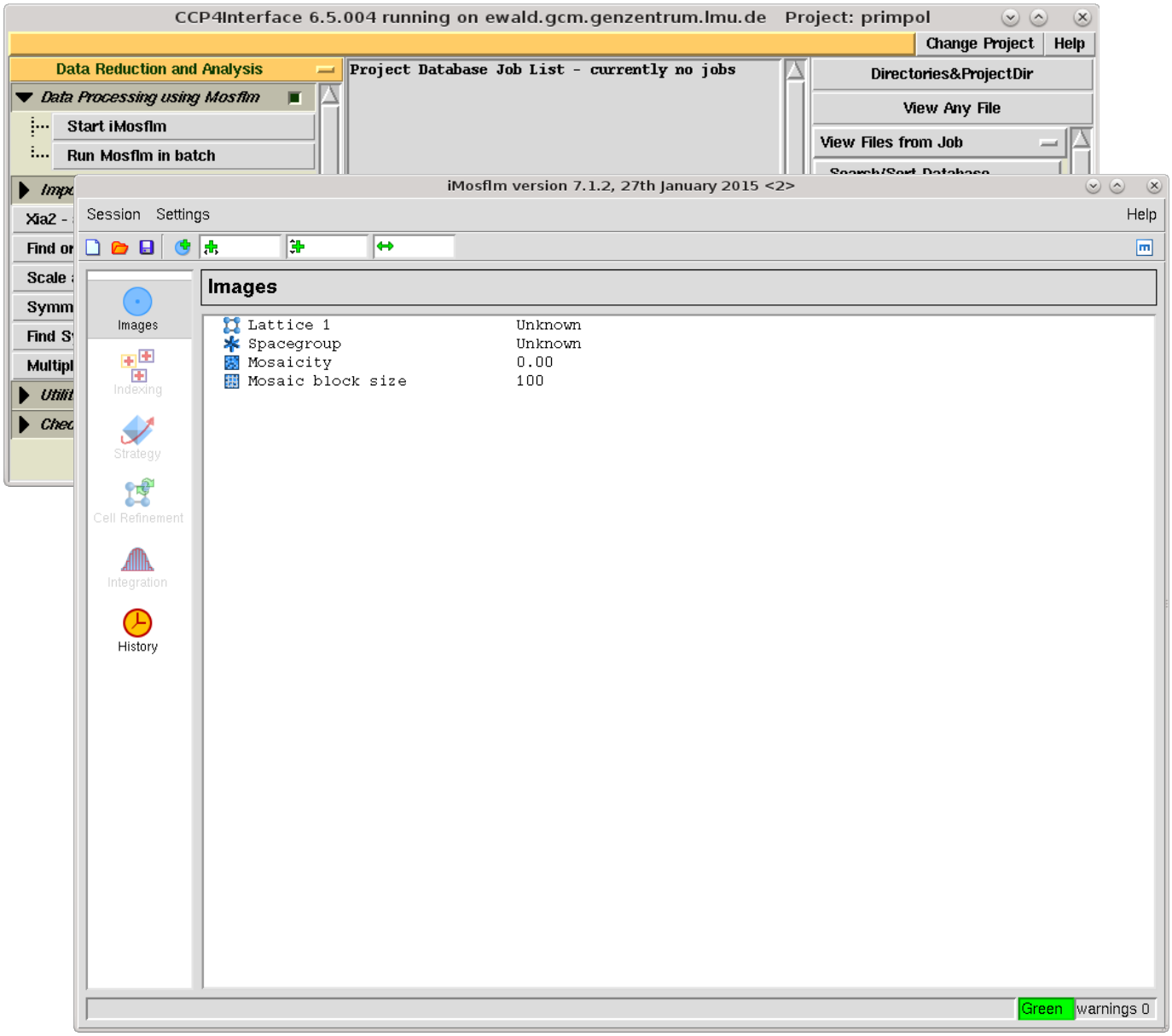
Data Set	Wavelength	Distance	Direct Beam	Start	ΔΦ	#Frames	Comment
pc303_1	0.97931 Å	124 mm	80.9/80.7 mm	50°	0.5°	200	peak
pc303_2	0.97974 Å	124 mm	80.9/80.7 mm	50°	0.5°	200	inflection
pc303_3	0.95372 Å	128 mm	80.9/80.7 mm	50°	0.5°	200	high energy remote
pc303_4	1.28312 Å	99 mm	80.9/80.7 mm	50°	0.5°	200	low energy remote

7.3 Define a new CCP4 project and start data processing

In a terminal, type `"ccp4i &"` to open the CCP4 graphical user interface (GUI). Set up your new project:



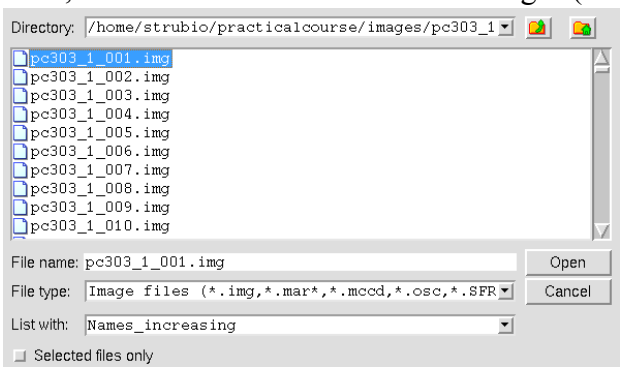
Choose the module "Data Reduction", open "Data Processing using Mosflm" and "Start iMosflm".

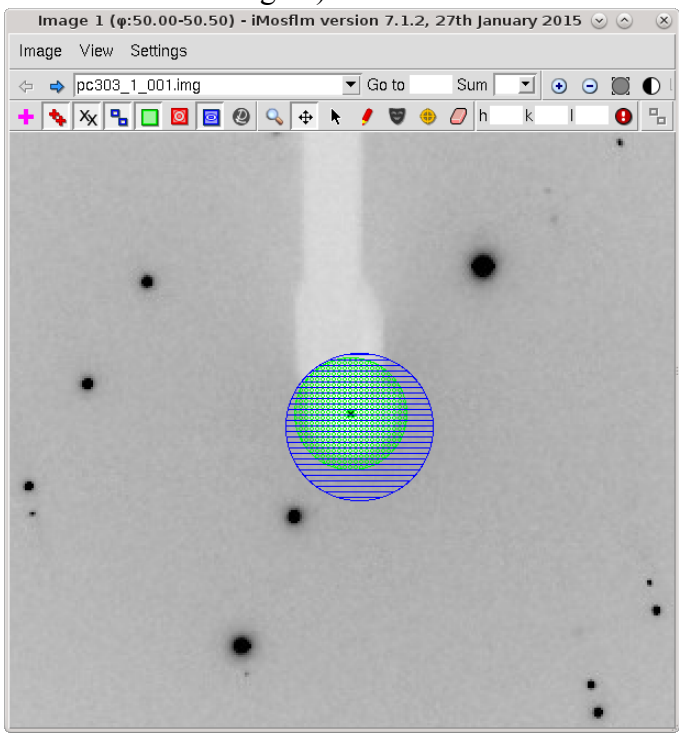


7.4 Getting the lattice parameters and first orientation

In order to assign a Miller index properly to each spot on a diffraction image, the unit cell dimension and space group plus the orientation of the crystal in space must be known. Fortunately, an automated software package is capable of finding these parameters from one or a few diffraction images ("*autoindexing*").

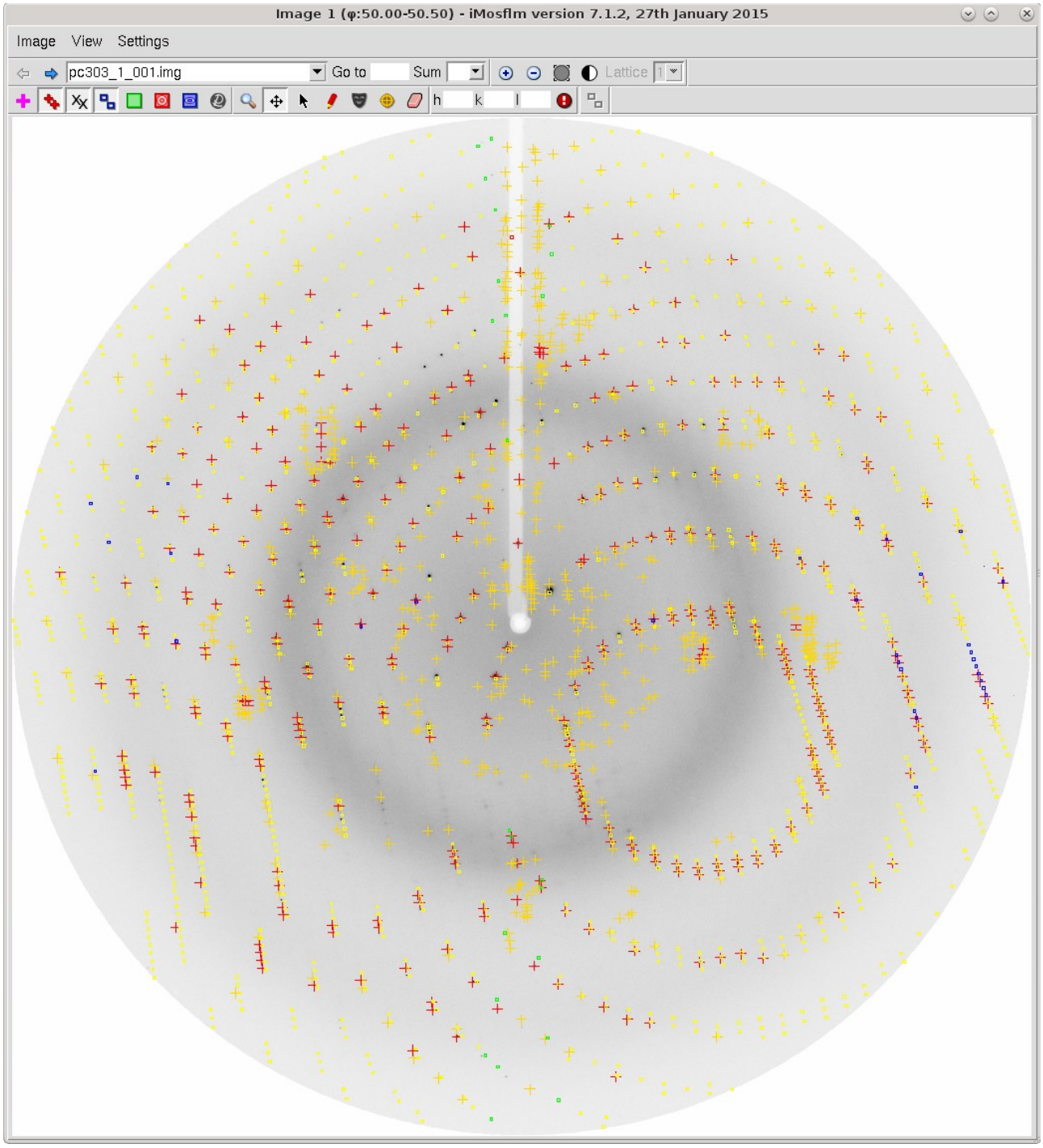
First, tell iMosflm where to find the images ("Session" → "Add Images"):



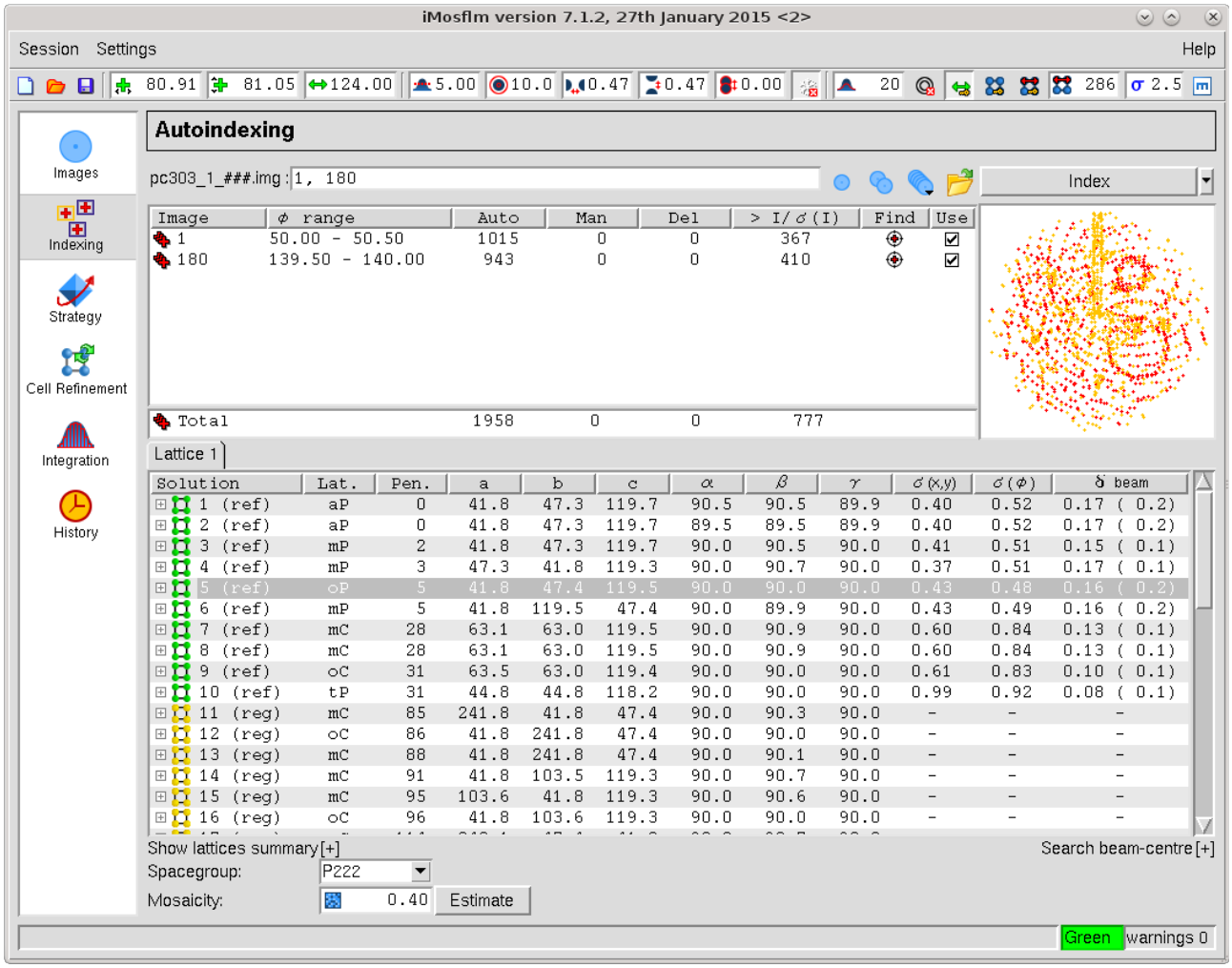


Then, set a proper mask to exclude the beamstop shadow (green button). Set a reasonable low resolution limit (blue button), and move the mask for the beamstop holder to the "North" (red button).

The first step in autoindexing is the *peak search*, which chooses spots (ideally a few hundred strong ones) to be used for indexing. This is done with the "Indexing" task in iMosflm. You will see the first diffraction image with crosses marking the spots that were found (red, above intensity threshold; yellow, below intensity threshold).



The second step is the mapping of diffraction maxima identified by the peak search onto reciprocal space, *i. e*, assigning hkl indices. This gives an estimation of unit cell parameters and possible space groups along with three parameters defining the orientation of the crystal and the exact diffractometer geometry. The squares mark the positions of predicted spots according to the chosen lattice (yellow, partial; blue, fully; red, overlap; green, too close to spindle axis). A summary of the indexing with the choice of the most probable lattice is given. The user can choose a different lattice.

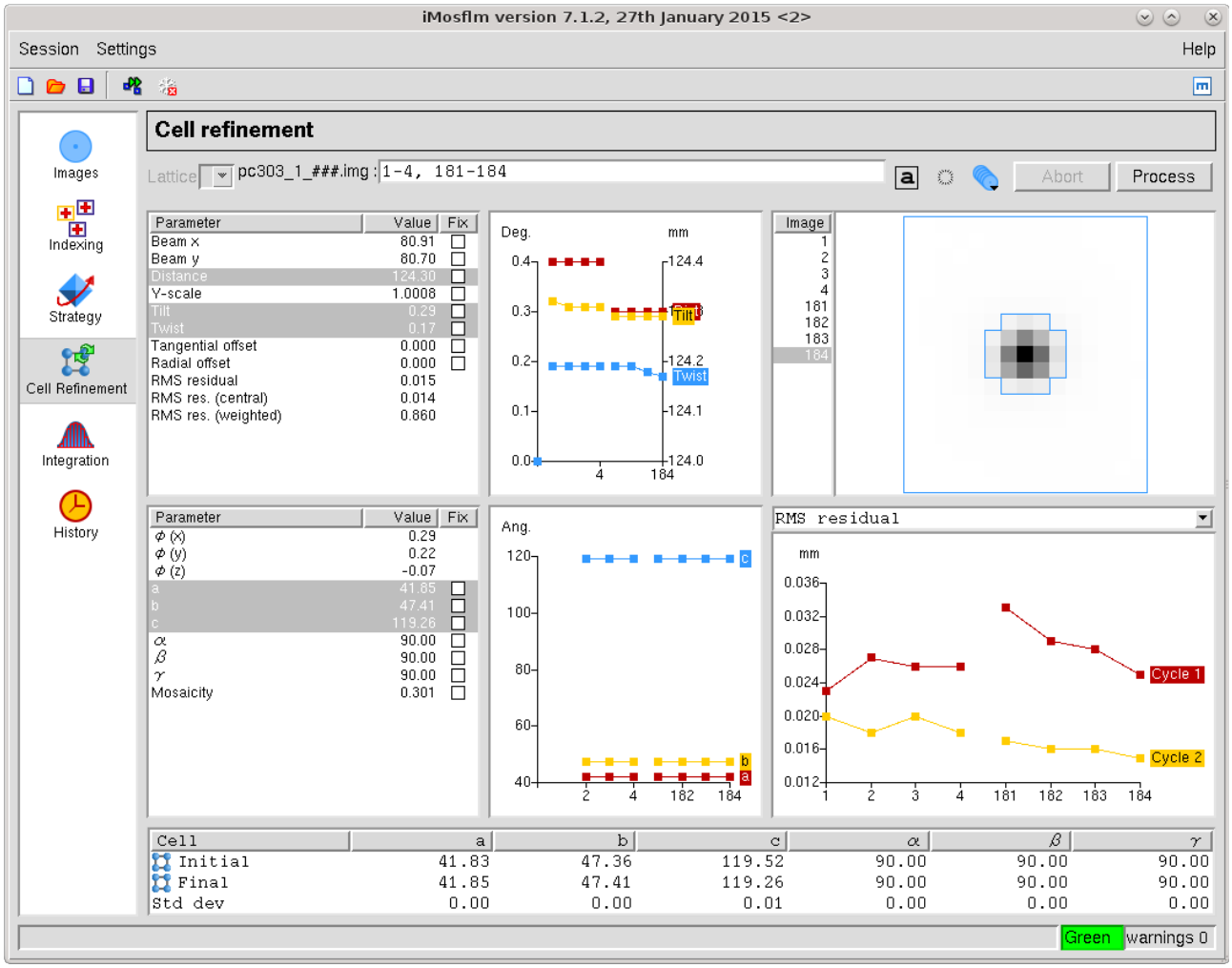


Once the user decides for a particular lattice, the restrictions imposed by this lattice (*e. g.* $\alpha = \beta = \gamma = 90°$ in orthorhombic) are applied and the initial crystal and diffractometer parameters are refined so that the predicted reflections superimpose with measured reflections as good as possible.

But first, we look at the "Strategy" task to predict, based on our lattice choice, how to optimally collect a complete data set on the shortest possible route, and thus in the shortest possible time, in order to reduce the detrimental effects of radiation damage (here, the data set has been collected already, but in a real situation, the strategy would help the user to optimally collect the data set).

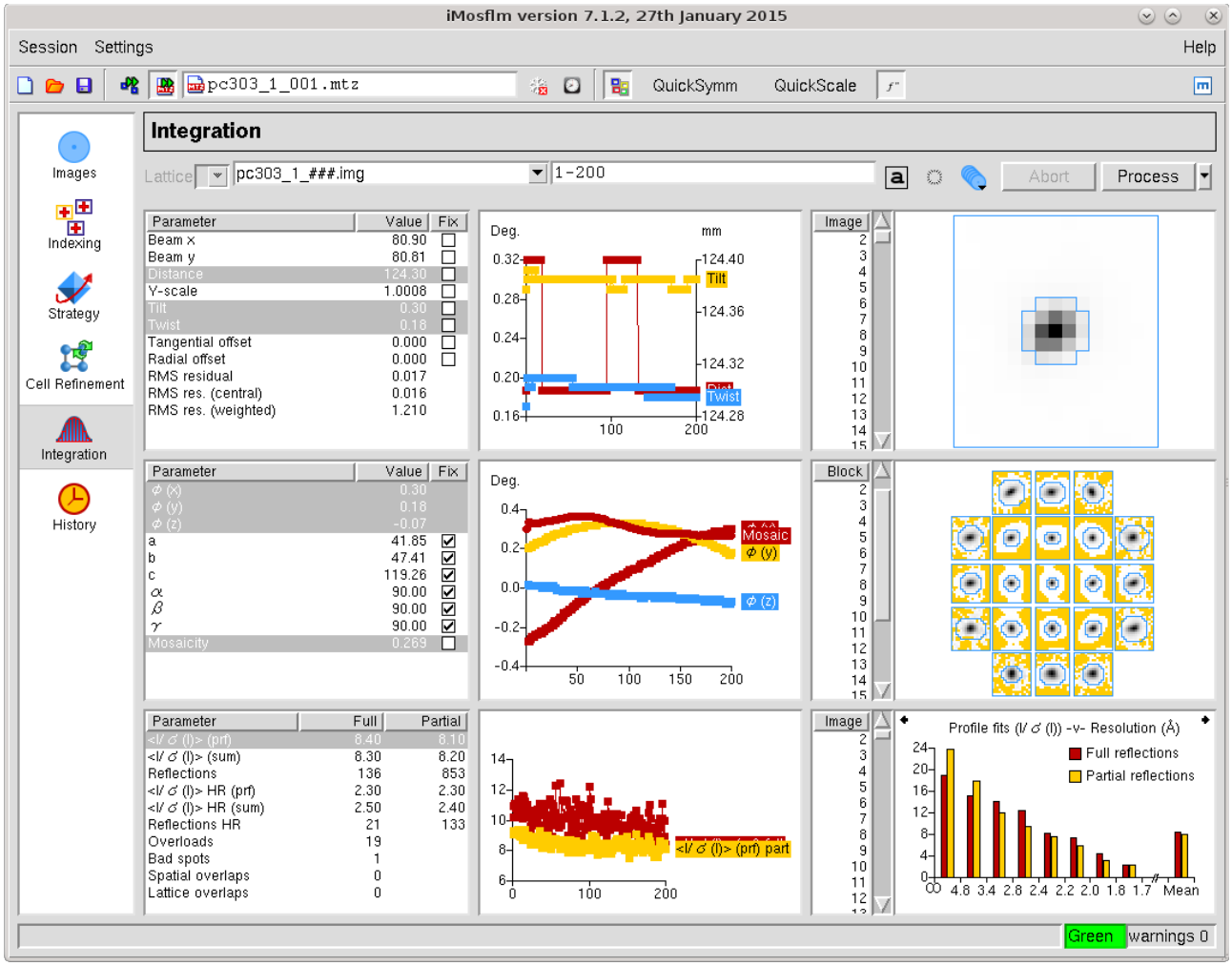
Please, check the effect on the completeness by "grabbing" the start and stop Φ angles in the pie chart.

Next, we refine the initial unit cell parameters and diffraction experiment parameters by selecting the "Cell refinement" task for two wedges of a few degrees that are 90° apart:



7.5 Integration

Now that the expected positions of reflections on the detector are known after indexing and refinement, the intensity of each reflection on each of the diffraction images has to be determined. This is done by selecting the "Integration" task. The software automatically sums up the pixel values under each spot and subtracts the detector background estimated by the background in a neighbouring area. The crystal orientation and diffractometer parameters are refined for every diffraction image. You may activate the button "Show predictions on images during processing".

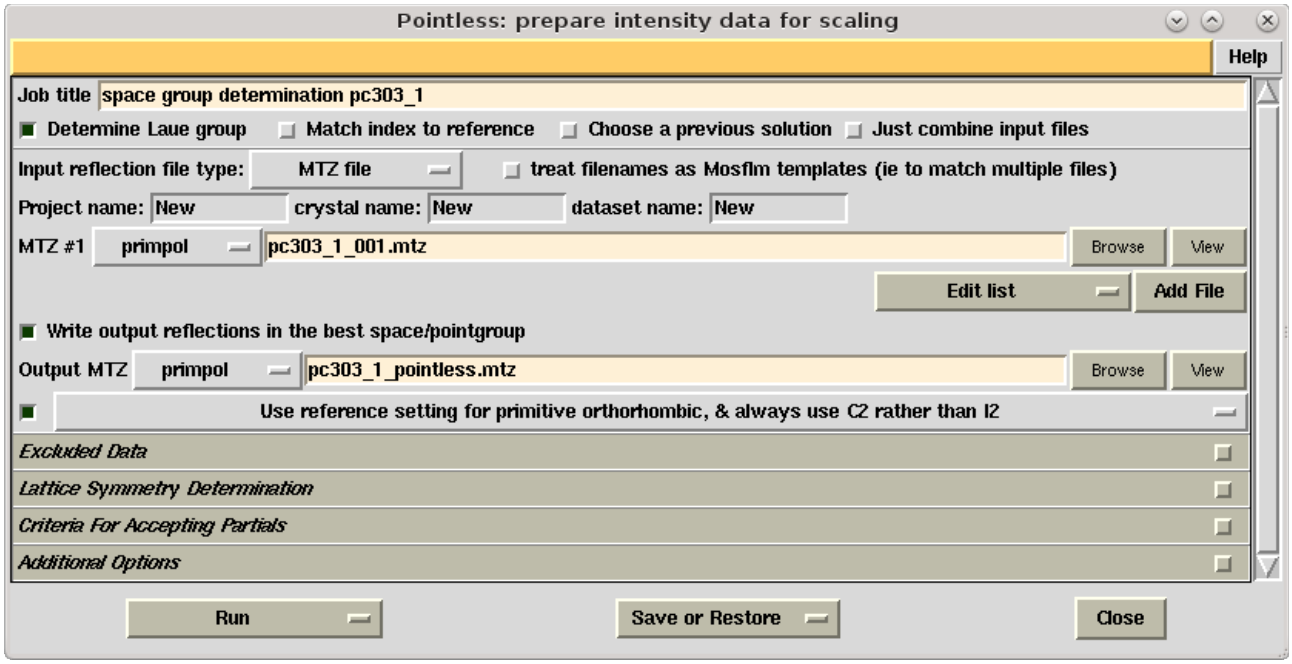


Finally, save the iMosflm session giving explicitly the extension ".mos". This session can be later reloaded for detailed inspection of the processed images, with blue boxes marking full reflections, yellow boxes partial reflections, and green and red boxes for rejected reflections.

Process all four data sets before space group determination and scaling.

7.6 Space group determination

For the primitive orthorhombic lattice, we have to assign one of the four possible space groups, P222 (No. 16), $P222_1$ (No.17), $P2_12_12$ (No. 18), $P2_12_12_1$ (No. 19). This is done by looking at extinction rules for axes reflections. For screw axes, only even axes reflections are present, which means for a 2_1-axis along *a*: (h,0,0)=2n, a 2_1-axis along *b*: (0,k,0)=2n, and a 2_1-axis along *c*: (0,0,l)=2n. In addition, we may have to re-index the reflections to change the order of axes according to the orthorhombic convention. The check for systematic absences, space group assignment and re-indexing is done from the CCP4 GUI with the task "Find or Match Laue Group" and choosing the option "Use reference setting for primitive orthorhombic ...".



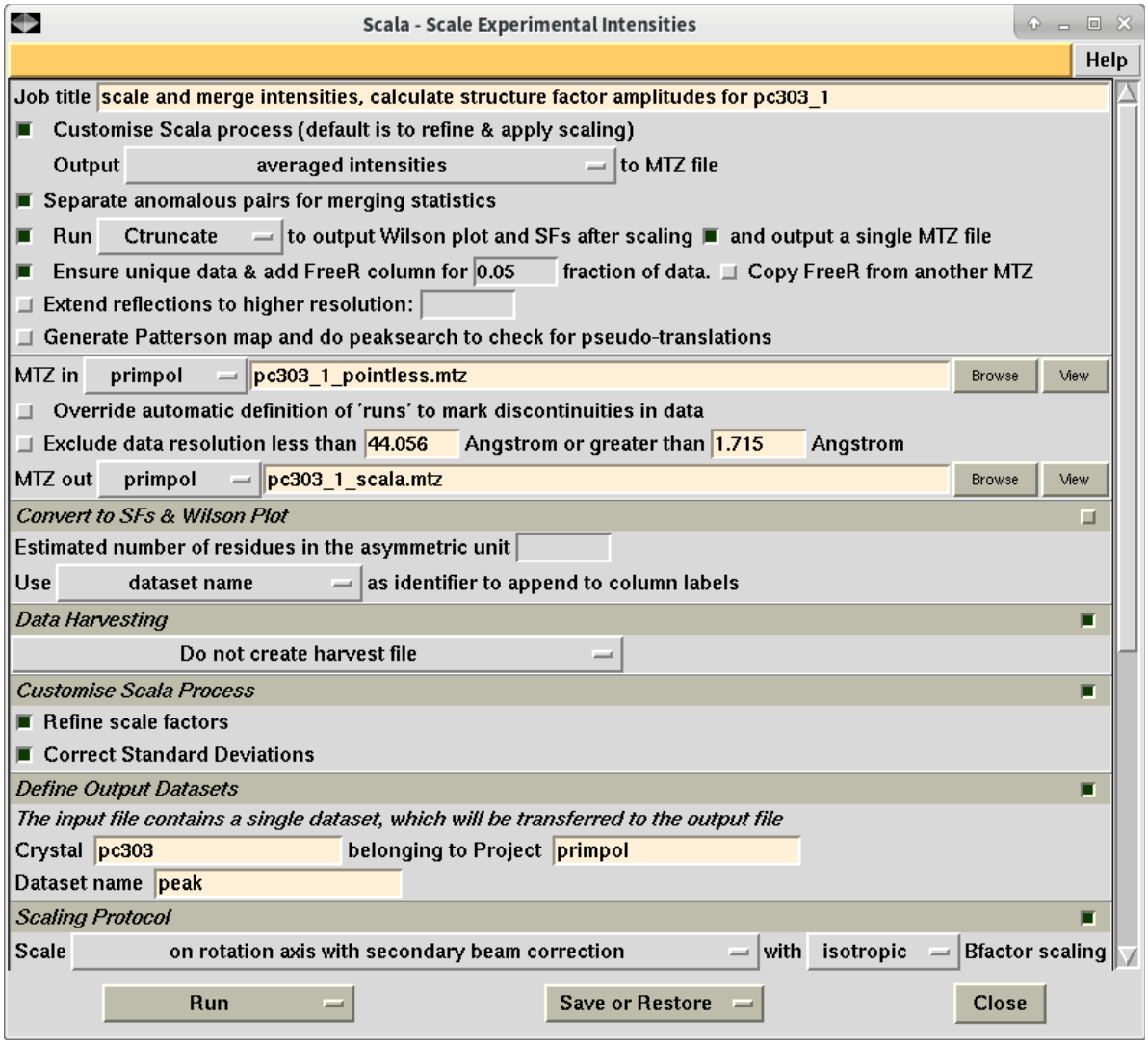
Please, do the space group determination for each data set and note the correct space group and any re-indexing that was necessary.

7.7 Scaling

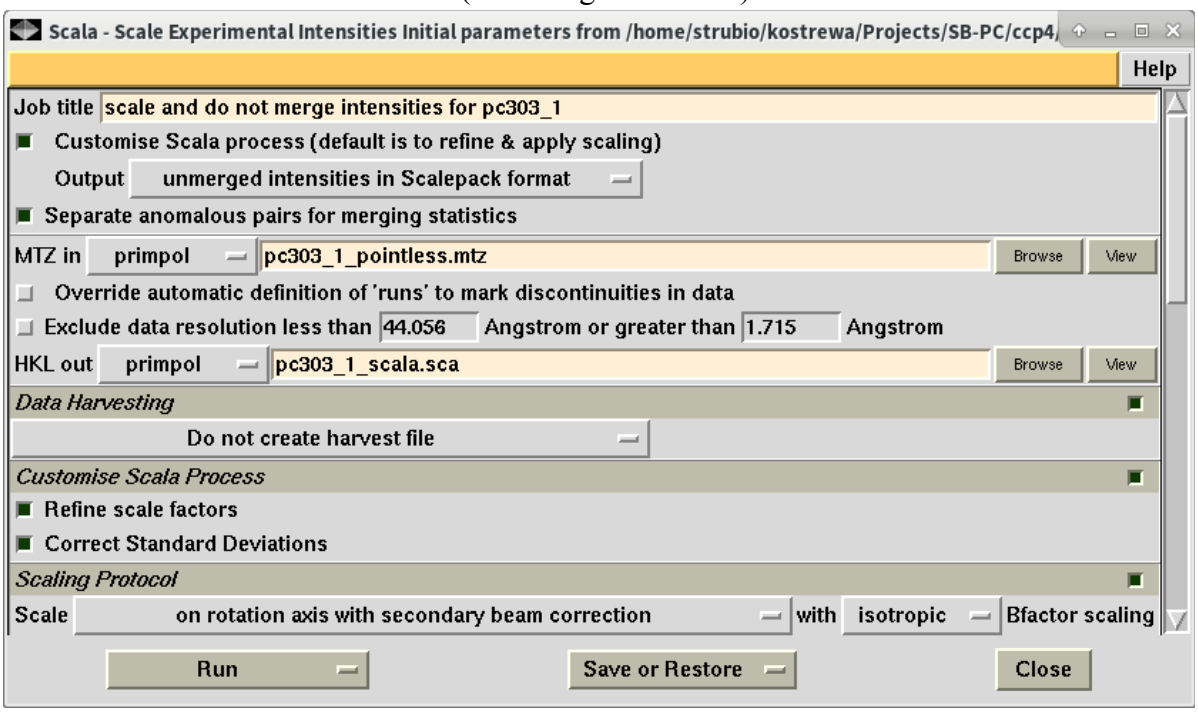
The scaling routine finds reflections on all frames of the dataset with hkl indices that should theoretically have the same intensity. In practice, fluctuations in exposed crystal volume, beam intensity, mechanic imperfections of the hardware, *etc.*, cause experimental errors. The scaling software tries to put the intensities of symmetry-equivalent reflections on the same scale as a function of the rotation angle. In addition, a simple empirical absorption correction is applied. Finally, the standard deviations of the intensities are corrected to reflect the differences between symmetry-equivalent intensities. This paves the way for statistics to estimate the quality of the dataset. We will run the CCP4 program SCALA for scaling.

Please, run for each data set "Scale and Merge Intensities" in the CCP4 GUI *twice*:

1. For each data set, use the option "Output averaged intensities to MTZ file" to get the symmetry-averaged merged intensities and structure factor amplitudes, which will be later used for refinement.


Scala - Scale Experimental Intensities
Help
Job title scale and merge intensities, calculate structure factor amplitudes for pc303_1
Customise Scala process (default is to refine & apply scaling)
Output averaged intensities to MTZ file
Separate anomalous pairs for merging statistics
Run Ctruncate to output Wilson plot and SFs after scaling and output a single MTZ file
Ensure unique data & add FreeR column for 0.05 fraction of data. Copy FreeR from another MTZ
Extend reflections to higher resolution:
Generate Patterson map and do peaksearch to check for pseudo-translations
MTZ in primpol pc303_1_pointless.mtz Browse View
Override automatic definition of 'runs' to mark discontinuities in data
Exclude data resolution less than 44.056 Angstrom or greater than 1.715 Angstrom
MTZ out primpol pc303_1_scala.mtz Browse View
Convert to SFs & Wilson Plot
Estimated number of residues in the asymmetric unit
Use dataset name as identifier to append to column labels
Data Harvesting
Do not create harvest file
Customise Scala Process
Refine scale factors
Correct Standard Deviations
Define Output Datasets
The input file contains a single dataset, which will be transferred to the output file
Crystal pc303 belonging to Project primpol
Dataset name peak
Scaling Protocol
Scale on rotation axis with secondary beam correction with isotropic Bfactor scaling
Run
Save or Restore
Close

2. For each data set, use the option "Output unmerged intensities in Scalepack format" to preserve the individual unmerged intensities for later selenium substructure-solution with the SHELX programs. *Please, note*: the file extension of the output scalepack file must be ".sca". Remove all other extensions (this a bug in the GUI)!



7.8 Evaluation of data quality

Each data processing software usually produces a number of plots and tables with statistical parameters. The most important ones to assess the quality of a dataset are discussed here.

1. The signal-to-noise ratio of the measured intensities, I/sig(I), is the most important criterion to assess the maximum resolution of a dataset. It decreases with resolution, and a common rule is that the highest acceptable resolution shell still has an I/sig(I) ≈ 2.

2. The Wilson-B-factor represents the decrease of diffraction intensities with resolution due to the fall-off of atomic scattering factors and internal crystal disorder:

$$I = I_0 \, e^{-\frac{1}{2} B (\sin\Theta/\lambda)^2}$$

Typically, the Wilson-B-factor has values around 20 Å² for crystals diffracting to high resolution, and increases to values > 100 Å² for crystals diffracting to low resolution. The usual way to determine the Wilson-B-factor is the Wilson plot (see Scaling & Merging output).

3. The symmetry-R-factor shows the agreement of intensities that should be the same due to crystallographic symmetry:

$$R_{\mathrm{sym}} = \frac{\sum_{hkl} \sum_{i} |I_i(hkl) - \langle I(hkl) \rangle|}{\sum_{hkl} \sum_{i} I_i(hkl)}$$

   Typically, the symmetry R-factor increases with resolution and it is customary to compute them for a number of resolution shells. Excellent crystals give data sets with symmetry R-factors as low as 0.05 (also stated as "5%"). Since the intensities decrease rapidly at high resolution, symmetry R-factors can increase to values > 1 in the highest resolution shells.

4. The completeness is the ratio of the number of measured unique reflections to the number of theoretically expected unique reflections. Useful datasets should have a completeness >95%. If the completeness is < 90%, this might indicate that some serious problem was present during data collection.

5. The redundancy (or multiplicity) represents the average number of measurements of a reflection and all of its symmetry-related reflections. The higher the redundancy, the more reliable is the average intensity measurement. High redundancy is particularly important for anomalous data, since anomalous differences are very small and close to the noise level.

Protocol/Documentation

Please, extract for each data set the data processing statistics for the merged intensities out of the log-files of the first SCALA runs. This will be used for a table summarizing the most important parameters of each dataset. As a guideline, you may use Table 2 in the original publication. Compare your results with the published values. Include plots of I/sigma and Rmerge against resolution in your protocol.

8 Patterson Function

8.1 Learning goals

You will learn which information the Patterson function provides and how to use this information to solve the heavy atom substructure of a derivative data set. You will learn how to construct a cross-table from the space group symmetry elements and the advantage of using Harker sections. Solving the heavy atom substructure is a pre-requisite to calculate initial macromolecular phases.

8.2 Patterson function: what is it good for?

A Patterson function is very similar to an electron density equation except that all the amplitudes are squared and the phases are "0". The peaks in the map do not correspond to centers of electron density, but mark inter-atomic vectors relating those centers.

$$P(u,v,w) = \frac{1}{V} \sum_{h,k,l} |F(h,k,l)|^2 \, e^{-2\pi i(hu+kv+lw)}$$

Please, remember that $|F(h)|^2$ is simply the measured intensity I(h).

An equivalent representation is the convolution of the atomic structure with itself:

$$P(u,v,w) = \sum_{i=1}^{n} \rho(x_i, y_i, z_i)\rho(x_i + u, y_i + v, z_i + w)$$

The Patterson function P(u,v,w) includes all possible inter-atomic vectors in the crystallographic unit cell. The height of each Patterson peak is proportional to the atomic numbers of the atoms that give rise to the corresponding inter-atomic vector. The number of peaks in the Patterson function is the square of the numbers of all the atoms in the unit cell. An example of a Patterson function is shown in the following figure:

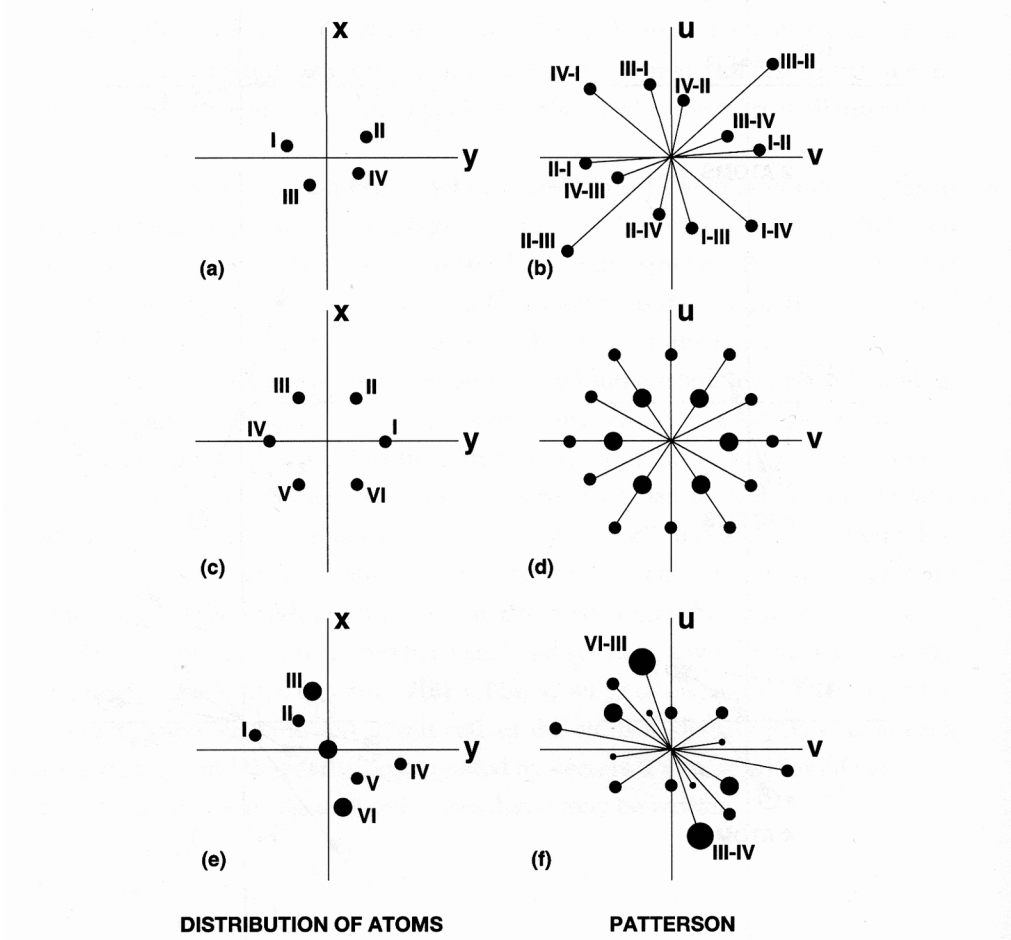


FIGURE 7.4 In *(a)* is a two-dimensional structure composed of four atoms, and in *(b)* its corresponding Patterson map. The four atoms give rise to $\mathbf{n}(\mathbf{n}-1) = 12$ vector peaks, plus the origin peak, which is not shown. The Patterson contains a center of symmetry even though the real structure does not. If one were presented with the Patterson map on the right, the objective, called deconvolution, is to deduce a set of atomic positions whose interatomic vectors satisfy all of the Patterson peaks, but require no others. A set of six atoms with a center of symmetry is shown in *(c)* along with the corresponding Patterson map in *(d)*. Some vectors are identical, though they relate completely different atoms. This gives rise to overlap in Patterson space and produces peaks of multiple weight, as well as a reduction in the total number. Note that the symmetry present in real space is preserved in Patterson space. In *(e)* is a dyad-related pair of three atom structures, containing one atom considerably more electron dense than the others. The corresponding Patterson map is in *(f)*. The vectors between the symmetry-related heavy atoms are outstanding in the Patterson because of the greater weight of their corresponding peaks. The heavy atom Patterson vectors precisely fix the coordinates of the heavy atoms in real space.

8.3 Interpretation of a Patterson function

If the real unit cell just contains a few atoms, the Patterson function of this object may be solved by generating a few trial structures. This is not possible with thousands of atoms in macromolecular structures where many similar Patterson vectors fall into the same Patterson space, and where the resolution of the Patterson peaks is limited by the non-atomic resolution of the diffraction data.

However, Patterson techniques are very useful for locating a few heavy atoms from structure factor differences.

The following two features of the Patterson function are very useful in macromolecular crystallography:

1. As P(u,v,w) is the product of the atomic numbers of the atoms at both ends of the interatomic vector, peaks between heavy atoms will result in much larger peaks. For example: C-C: 6x6=36, C-Br: 6x35=210, Br-Br: 35 x 35 = 1225, Hg-Hg: 80x80=6400.

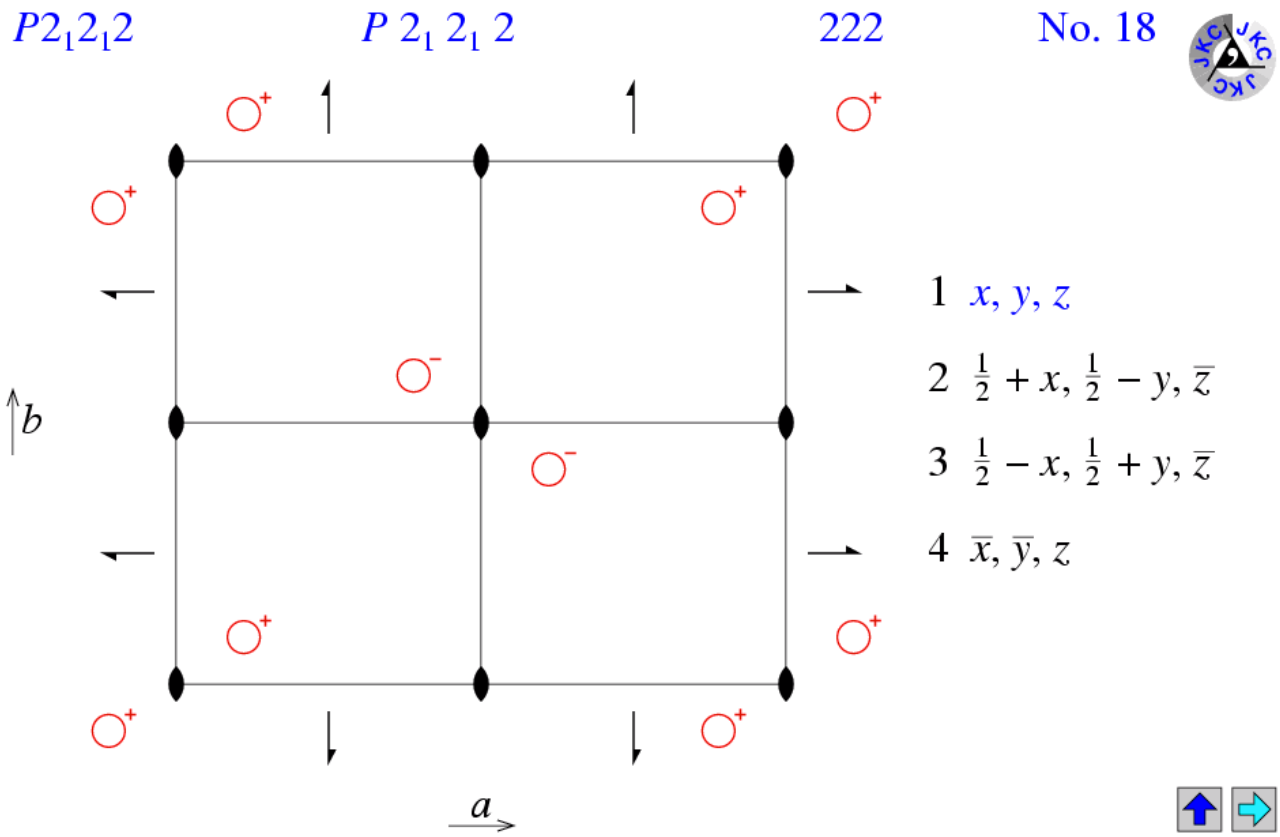
2. Symmetry-related atoms in a unit cell can result in vectors lying on a plane in Patterson space, called *Harker sections*. For example, a crystallographic twofold symmetry along the ***b*** axis generates for each atom at position (x,y,z) a symmetry-equivalent atom at position (-x,y,-z). The interatomic vectors of all symmetry-equivalent atom pairs fall on the Harker section (u,v,w)=(2x,0,2y). The peaks on the Harker sections can be used to reveal the actual coordinates (x,y,z) of the heavy atoms in the unit cell.

In practice:

For heavy atom sub-structure solution, Patterson functions are computed from isomorphous differences ($\Delta F = |F_{deriv}| - |F_{native}|$) or anomalous differences ($\Delta F = |F^+| - |F^-|$) to reduce the noise from other interatomic vectors in the unit cell.

To find out, which Harker sections belong to a given space group, simply look at all possible difference vectors between symmetry-equivalent positions (you may invert the signs of the results) and write them into a cross-table.

Space group diagram for $P2_12_12$ (No. 18)



Cross-table to get Harker sections for $P2_12_12$

9 Macromolecular Phasing

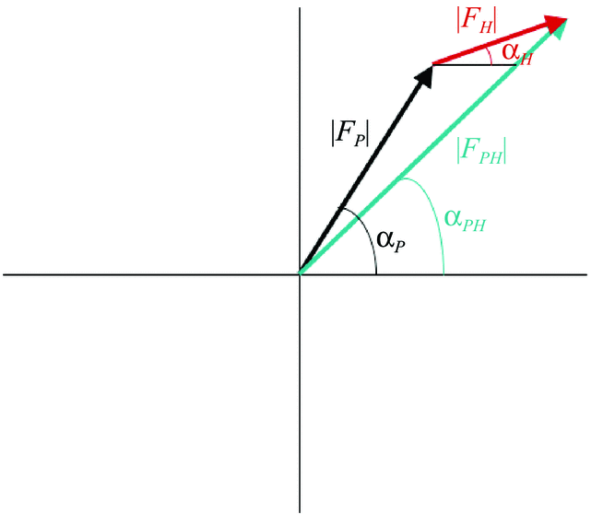
In order to solve a macromolecular crystal structure, the heavy atom substructure must be solved, and initial phases must be calculated using the heavy atom positions and the observed structure factor amplitudes. Here, common isomorphous phasing techniques using the known heavy atom substructure are shortly explained for SIR(AS), MIR(AS), SAD and MAD. Phasing by molecular replacement will be explained in a later chapter.

9.1 Learning goals

You will solve the heavy atom substructure. You will calculate initial protein phases and determine the correct hand of the heavy atom sites. You will estimate the solvent content and the expected number of molecules in the asymmetric unit, and you will use density modification techniques to improve the initial protein phases.

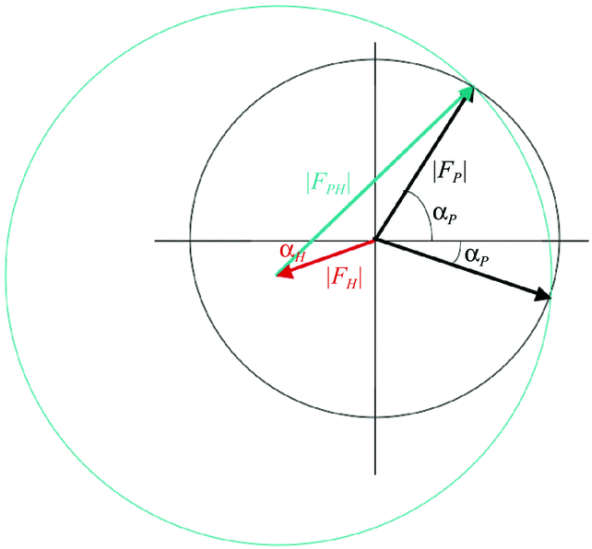
9.2 Single isomorphous replacement, SIR(AS)

In single isomorphous replacement (SIR), a native data set without heavy atoms and a derivative data set with heavy atoms are collected. The heavy atom contribution leads to changes in the structure factors of the derivative, which is shown in the following Argand diagram (F_P, native structure factor; F_{PH}, derivative structure factor):



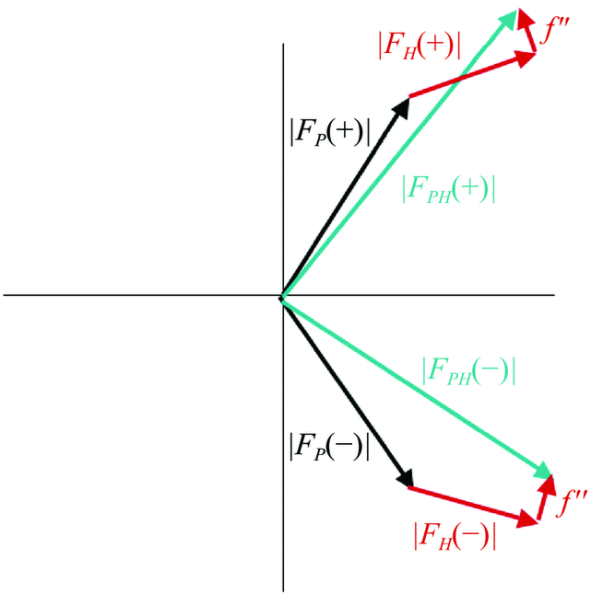
If the phases are unknown, all we know is that the ends of the structure factors lie on circles with radius $|F_P|$ and $|F_{PH}|$, respectively. If the substructure of the heavy atoms can be solved, the structure

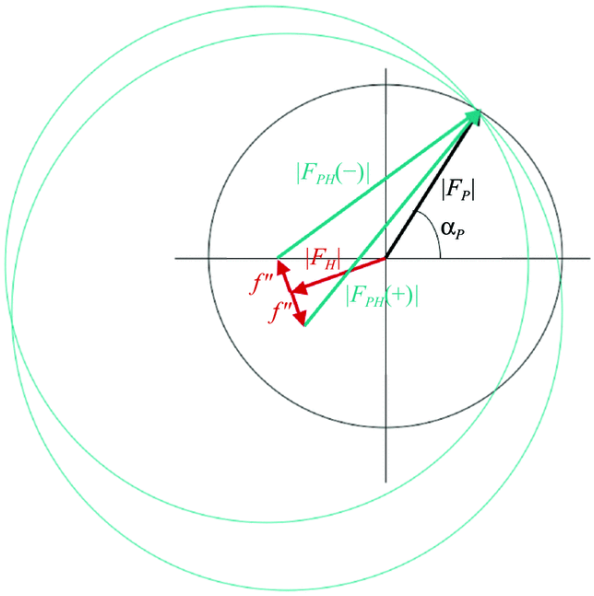
factor of the heavy atoms can be calculated and the unknown phases of the native structure factor can be restricted to usually two values, which is shown in the following Harker diagram. The heavy atom structure factor F_H has to start at the native circle and end at the derivative circle. In the Harker diagram, the center of the derivative circle is shifted by $-F_H$, and the intersection of the two circles gives the two possible solutions for the native phase.



The Harker diagram represents a simple way of identifying possible phases.

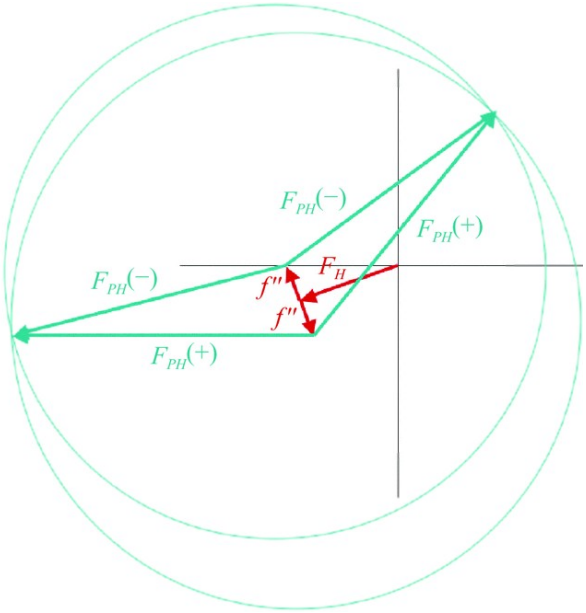
If the anomalous signal from the heavy atom can be used, the protein phase can be unambiguously determined using diffraction data from the native protein and from one derivative. This method is called single isomorphous replacement with anomalous scattering, SIRAS. The Argand and Harker diagram for SIRAS are shown below.





9.3 Single anomalous dispersion, SAD

Calculating phases by using the difference in anomalous scattering (f'') at one wavelength from one diffraction data set with an anomalous scatterer is called single anomalous dispersion, SAD. Although, it is generally not possible to solve the phase ambiguity if only two experimental measurements are available, in case of SAD, there is a slightly higher probability that the protein phase has a value closer to the phase of the anomalous scatterer. The Harker diagram for SAD is shown on the right.

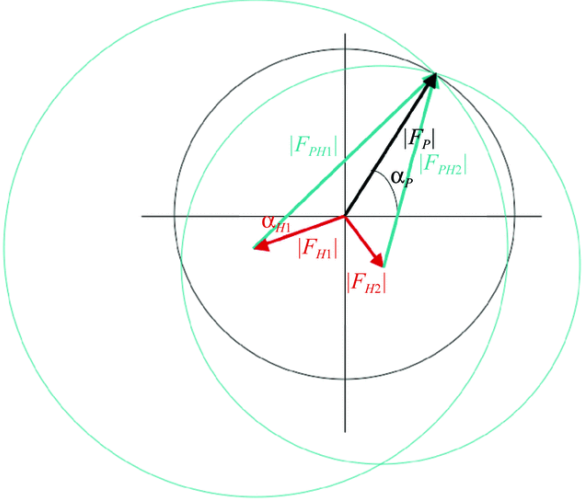


Modern approaches will use probabilistic methods to determine the initial phases and their reliability (maximum likelihood phasing - used in SHARP and PHASER). These phase distributions are improved by density-modification procedures. Of the two possible enantiomers of the anomalous scatterer sub-structure, only one will result in an interpretable electron density.

Often for SAD, a naturally binding anomalous scatterer like zinc or iron, can be directly used for phasing, without the need to get a heavy atom derivative. Sometimes, even the small anomalous signal of the sulfur atoms that occur in cysteine and methionine amino acids can be used for successful SAD phasing.

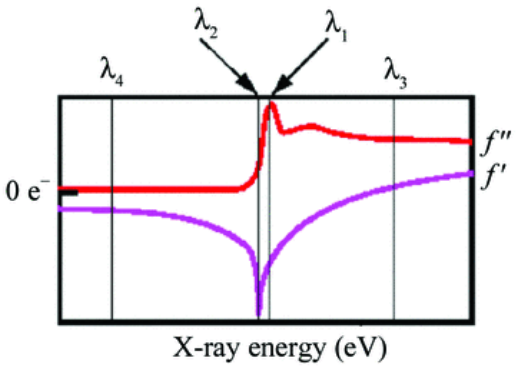
9.4 Multiple isomorphous replacement, MIR(AS)

In multiple isomorphous replacement, a native data set and data sets of two or more different heavy atom derivatives are used for phasing. This leads to a unique phase solution for the unknown native phase, as shown in the Harker diagram on the right. If the anomalous effects of the heavy atoms are measured, the resulting solution for the native phase should be even clearer.

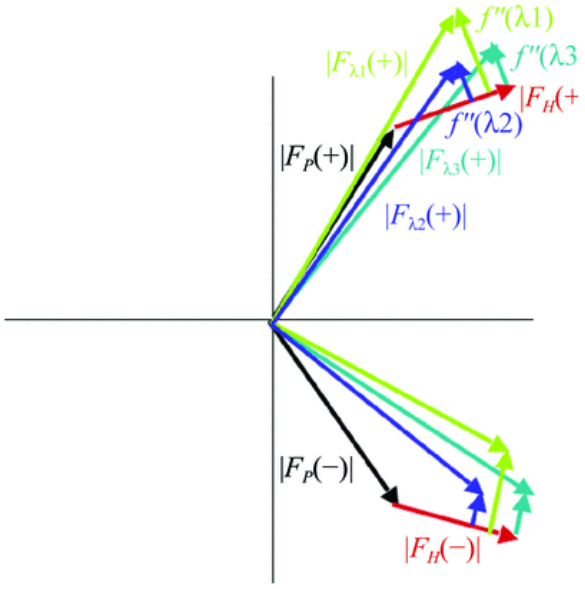


9.5 Multi-wavelength anomalous dispersion, MAD

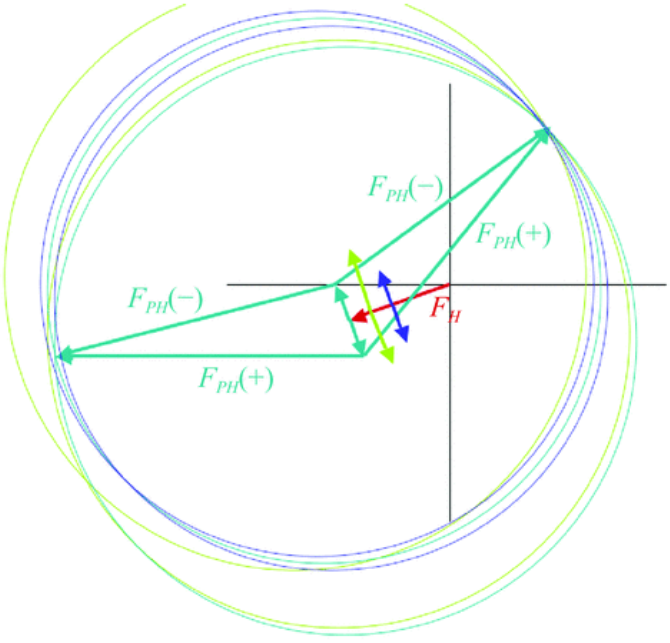
Using the isomorphous scattering f' and anomalous scattering f'' contribution of an anomalous scatterer at different wavelengths for phasing is called multi-wavelength anomalous dispersion, MAD. In a typical MAD experiment, diffraction data are collected at three different wavelengths, a peak wavelength at maximum f'', an inflection wavelength at minimum f', and a remote wavelength at the high-energy side of the peak. Sometimes, a fourth data set is collected at the low-energy side of the peak. Altogether, we have three to four data sets, each with different isomorphous and anomalous differences that result in a unique determination of the unknown phase. In combination with seleno-methionine substitution, MAD is probably the most powerful phasing method in protein crystallography. It is necessary to measure the exact wavelength of the peak with a fluorescence spectrum at the synchrotron, since the peak position depends on the chemical environment of the anomalous scatterers. A typical spectrum for seleno-methionine substituted protein and the resulting Argand and Harker diagrams for MAD are shown below.


λ2
λ1
λ4
λ3
f″
f′
0 e⁻
X-ray energy (eV)

(*a*)


f″(λ1)
|Fλ1(+)|
f″(λ3
|FH(+
f″(λ2)
|FP(+)|
|Fλ3(+)|
|Fλ2(+)|
|FP(−)|
|FH(−)|

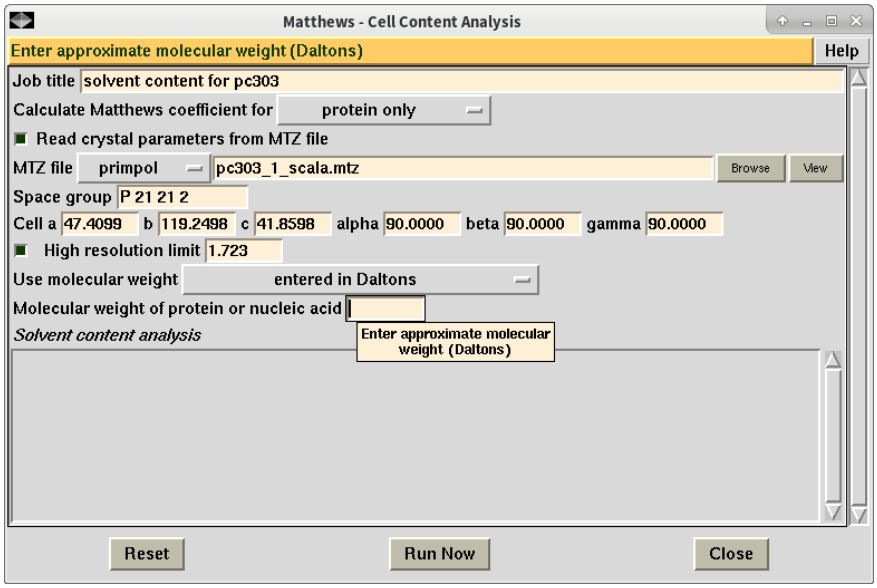
(*b*)


FPH(−)
FPH(+)
FPH(−)
FH
FPH(+)

(*c*)

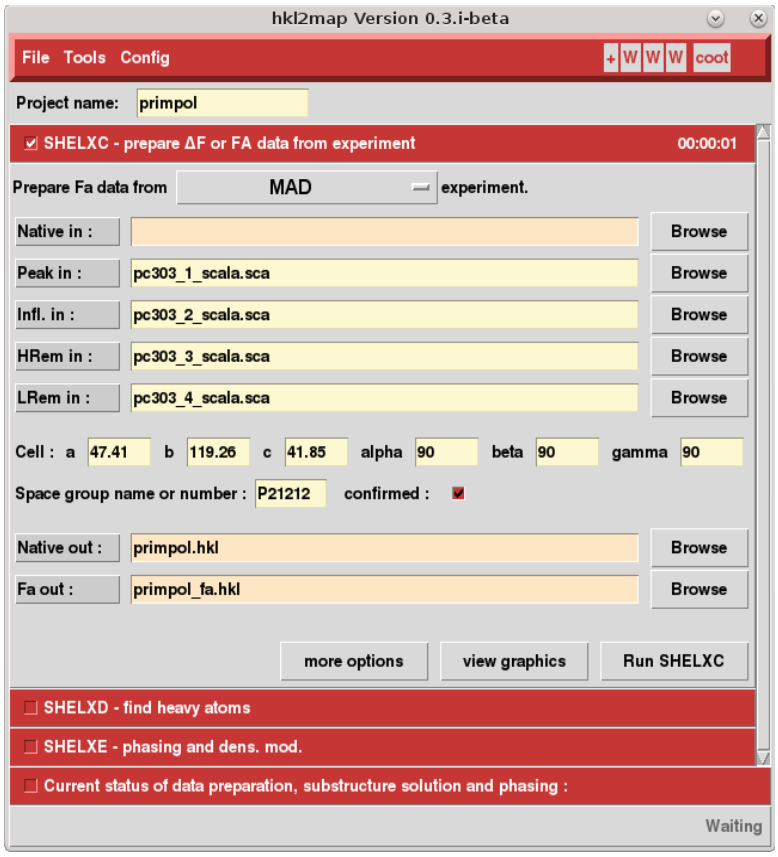
9.6 Solving the heavy atom substructure

First, we will estimate the expected solvent content using the CCP4 GUI by choosing the module "Density Improvement" and the task "Cell Content Analysis". Which solvent contents are possible?



We will then use the programs SHELXC, SHELXD and SHELXE from the SHELX suite (http://shelx.uni-ac.gwdg.de/SHELX/) to prepare the data, locate the selenium atoms, calculate initial phases and improve the phases by density modification. All SHELX programs can be called in a command line. The read input parameters from text files and write their results to various output files. Here, we will use a GUI for the SHELX programs with nice input masks and output diagrams by calling `"hkl2map &".` Make screenshots of important results, since the hkl2map session cannot be saved!

SHELXC prepares the data for heavy atom substructure solution. SHELXC reads data sets preferably as unmerged intensities, puts those data sets on a common scale, analyses anomalous differences and writes out the scaled data sets and an input file for use with SHELXD. Tell SHELXC the type of experimental phasing and the names of the input files.

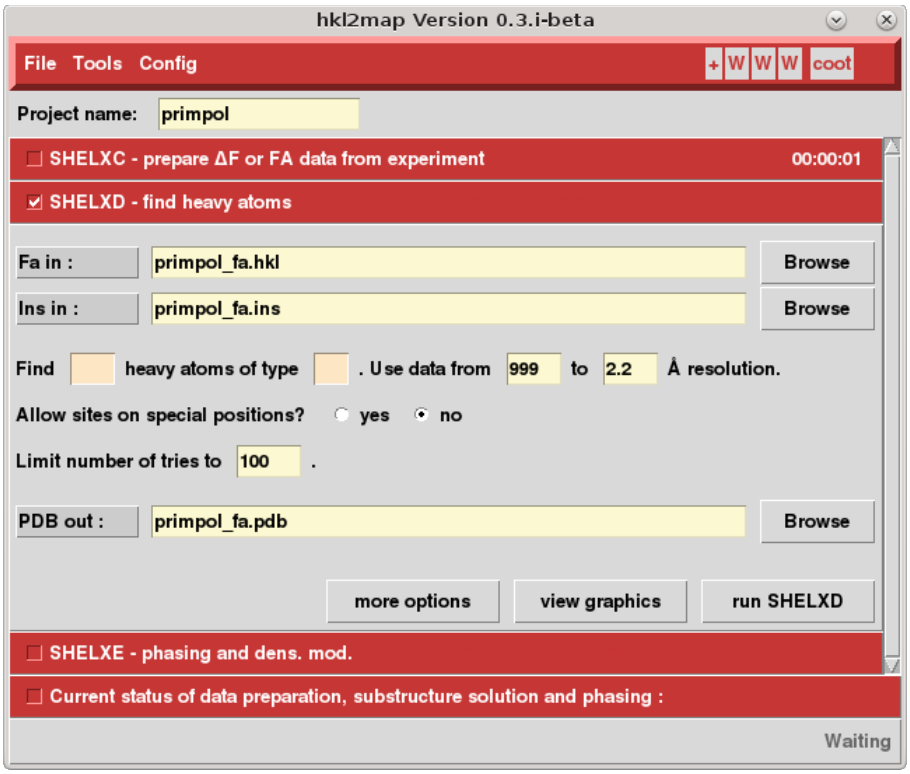


With "view graphics", you can visually inspect the results of every SHELX run. Either make snapshots of these graphics, or save them as PostScript plots.

For SHELXC, the most useful graphics is the strengths of the anomalous signals in your data sets, "<d"/sig> *vs.* Resolution", which should be above ≈1.2. Which data set has the strongest anomalous signal? At which maximum resolution would you cut the data?

The next step in the heavy atom substructure solution is SHELXD, which uses Harker vectors for initial heavy atom sites, and a cycling between phase shifts in reciprocal space and peak-picking in real space to solve the heavy atom substructure.

Tell SHELXD how many heavy atoms and which element you expect, the maximum resolution to accept data, and how many random trials it should do to find these atoms. How many heavy atoms do you expect?



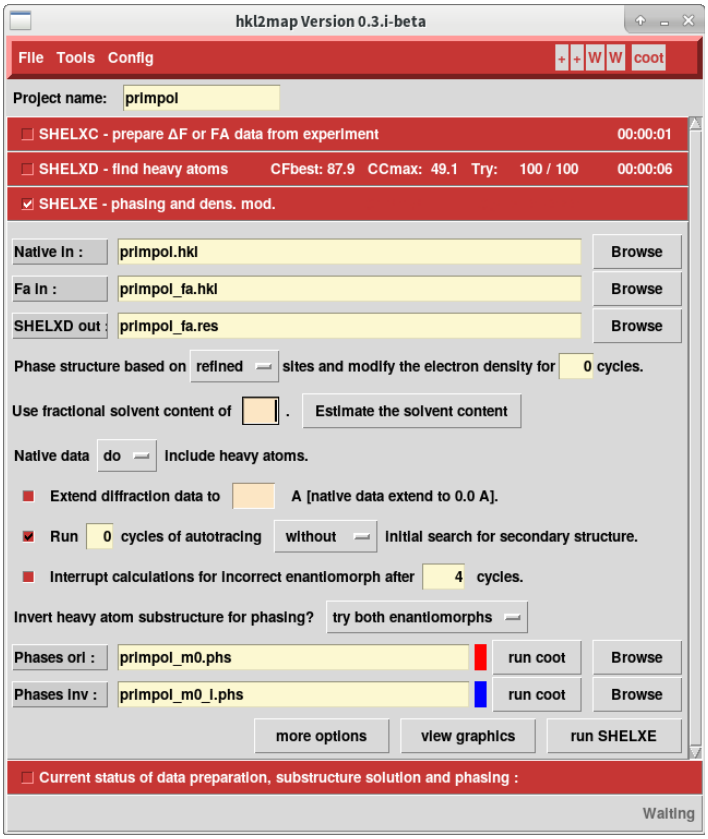
SHELXD produces various plots that you should all inspect. Very useful are the correlation coefficients of calculated *versus* observed heavy atoms structure factors for all and weak data, "CCall vs. CCweak", for each trial solution. Trials with large CCall (> 30) *and* large CCweak (> 15) are likely correct solutions.

Another very useful plot are the relative occupancies of the heavy atoms of the best solution, "Site Occupancy *vs.* Peak Number". A steep falloff either indicates a marked change in site occupancy or the limit of the true number of heavy atoms found. Occupancy values below 0.2 usually indicate noise peaks.

9.7 Calculating initial phases and map improvement by density modification

We use SHELXE in a first run for refinement of the heavy atom substructure from SHELXD and calculation of initial phases, and in a second run for phase improvement using solvent-flattening and automatic main chain tracing. Since it is not clear at this stage, whether the original heavy atom substructure or its inverted substructure leads to correct phases, SHELXE calculates and writes out phases and figure-of-merits for both hands.

In the first run, use SHELXE *only* for refinement of the heavy atom positions from SHELXD and calculation of initial phases by specifying "0" electron density modification and autotracing cycles. You must enter the solvent content from the previous cell content analysis:



We will now visually check, which of the two hands is the correct one by looking at the resulting electron density maps. However, if there is a clear difference between both hands, only phases for the better hand will be written out in this step. You can check the appearance of the electron density maps calculated with initial phases by clicking on the respective buttons "run coot" in SHELXE. In the Coot window, display the unit cell with "Draw" → "Cell & Symmetry" → "Show Unit Cells": "Yes". Change the contour level of the electron density map with the mouse wheel to a value around 2.5 rmsd and increase the map radius with "Edit" → "Map Parameters" → "Map Radius" to "50".

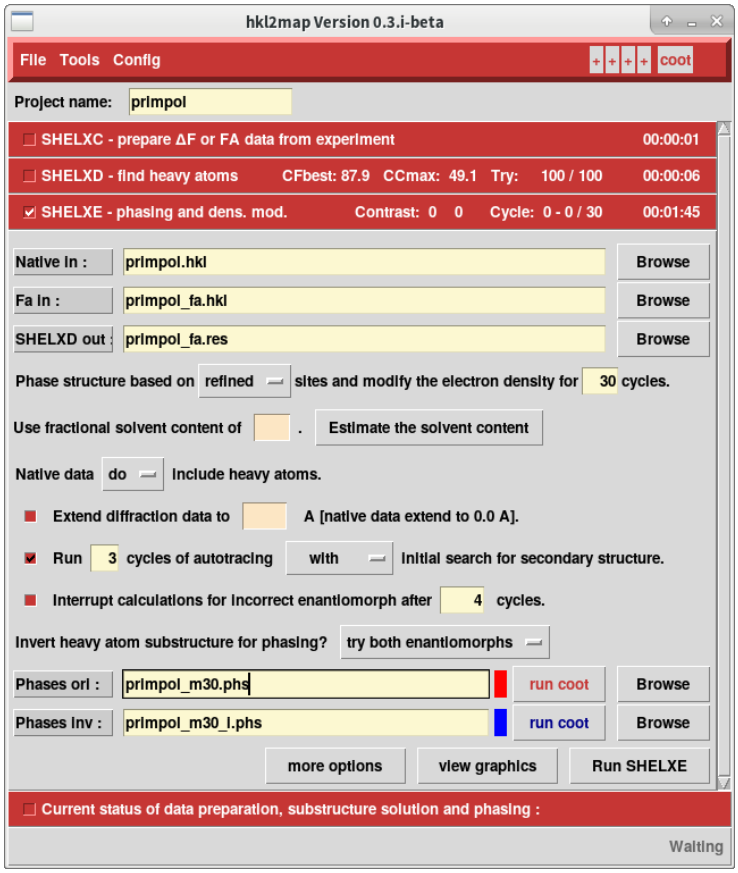
Can you clearly distinguish between protein and solvent regions in the electron density map for one of the two hands?

Make a nice picture for the protocol with "Draw" → "Screenshot" → "Simple" and give it a new output file name.

All density modification algorithms seek to improve phases by imposing restrictions on the density in real space, followed by using the phases of the modified map to modify the initial phases. The most important density modification technique is *solvent flattening*. During solvent flattening, the electron density outside a protein mask is replaced by a flat constant density. A variation of this technique is called solvent flipping. Here, the deviations of the electron density map from a flat constant density outside a protein mask are inverted, resulting in an even more effective reduction of these peaks. SHELXE uses another variant of solvent-flipping, which modifies the electron density map depending on the local variance within a sphere of 2.42 Å, the so-called "sphere of influence". If the variance is high, the pixel probably belongs to protein and the electron density is retained, if it is low, the pixel probably belongs to solvent and the electron density is flipped.

In addition, SHELXE can do automatic main chain tracing, further improving the phases by combination with the partial model phases.

In the second run, use SHELXE with 30 cycles density modification and 5 cycles automatic chain tracing with secondary structure search for getting an initial template for your model building and refinement:



Which of the two hands has higher values in the output graphics "Contrast vs. Cycle"?

9.8 Comparing initial electron density maps

First, you can quickly check, which of the two hands gave better results in the density modification step and thus is the correct one by clicking on the respective "run coot" buttons in SHELXE. Make some nice pictures for the protocol.

Second, for the correct hand, compare the initial electron density map without density modification with the electron density map after density modification.

Start Coot in a terminal with the command `"coot &"` and open the autotracing model for the correct hand with "File" → "Open Coordinates". Open both the initial electron density map and the electron density map after density modification with "File" → "Open MTZ ..." and choosing the *.phs files for the correct hand with 0 cycles and with 30 cycles density modifiaction. To better distinguish between the two maps, you can colour them as you like with "Display Manager", then for each map with "Properties" → "Colour". Do you see differences in quality with and without density modification? Again, make some nice pictures for the protocol.

It is now a good opportunity to check the heavy atoms. Click on the "Go To Atom" button next to "Display Manager" , open the chain "Z" and center on the first selenium atom. With the space bar,

you can center to the next selenium atom. How many selenium atoms are there? Do you have an explanation, why the number of selenium atoms is different to the expected number?

Check the packing of your model in the crystal:

- change the display of your initial model in "Display Manager" to "C-alphas/Backbone"
- display the unit cell with "Draw" → "Cell & Symmetry" → "Show Unit Cells?": "Yes"
- still in "Cell & Symmetry", display symmetry equivalent molecules with "Master Switch: Show Symmetry Atoms?": "Yes", and "Symmetry by Molecule" → "Display as CAs"
- increase the "Symmetry Atom Display Radius" to a larger value (≈ 50 Å)

Look at the packing of your molecule in the crystal. Can you see, how the molecules build the crystal? Do you see the solvent channels? Make some nice pictures for yourself and for your protocol.

A short Coot tutorial is given at the end of this script.

10 Electron Density Maps

When everything has been done to get the best possible starting phases, the time has come to examine the structure in the form of an electron density map. An electron density map is the Fourier transform of the structure factors (including phases). Depending on the resolution of the diffraction data and the quality of the starting phases, the initial electron density map may be readily interpretable in terms of atomic coordinates (built into the electron density map as amino acids). It is usually necessary to iterate many times between model building using a computer graphics program and refinement (see below).

10.1 Types of maps

During model building and refinement, one will use various maps, which differ in the use of observed and calculated structure factor amplitudes, F_{obs} and F_{calc}, respectively. The most frequently used maps are briefly described here:

F_{obs} map

The observed structure factor amplitudes $|F_{obs}|$ are used together with the best available phases α_{best} (e.g. from MAD phasing). An F_{obs} map might be very noisy and hard to interpret. In particular, when the phases are calculated (e.g. by molecular replacement), this type of map is subject to model bias. F_{obs} maps can be improved by weighing each coefficient by the confidence in its phase (figure-of-merit, *m*):

unweighted: $$\rho(\vec{r}) = \frac{1}{V} \sum_{h,k,l} |F_{\mathrm{obs}}(\vec{S})| \, e^{i\Phi^{\mathrm{calc}}} \cdot e^{-2\pi i \vec{r}\vec{S}}$$

weighted: $$\rho(\vec{r}) = \frac{1}{V} \sum_{h,k,l} m|F_{\mathrm{obs}}(\vec{S})| \, e^{i\Phi^{\mathrm{calc}}} \cdot e^{-2\pi i \vec{r}\vec{S}}$$

F_{calc} map

This is the least useful map for crystallographic purposes: the calculated amplitudes derived from a model are phased with the calculated phases from the same model and you get back exactly what you put in (*model bias*).

$$\rho(\vec{r}) = \frac{1}{V} \sum_{h,k,l} |F_{\mathrm{calc}}(\vec{S})| \, e^{i\Phi^{\mathrm{calc}}} \cdot e^{-2\pi i \vec{r}\vec{S}}$$

$F_{obs} - F_{calc}$ or difference map

The difference Fourier map using the coefficients F_{obs}–F_{calc} and Φ_{calc}, where Φ_{calc} is the calculated phase from a model, is very useful in terms of information content: in this map, there are positive peaks at locations where electron density in the model is missing, and negative peaks at locations where too much electron density was put into the model. This map is especially useful for finding corrections to the current model, e.g. looking for missing amino acids, detecting movements etc.

$$\rho(\vec{r}) = \frac{1}{V} \sum_{h,k,l} (|F_{\mathrm{obs}}(\vec{S})| - |F_{\mathrm{calc}}(\vec{S})|) \, e^{i\Phi^{\mathrm{calc}}} \cdot e^{-2\pi i \vec{r}\vec{S}}$$

$2F_{obs} - F_{calc}$ map

The $2F_{obs}$–F_{calc} map can be seen as the sum of a F_{obs} map and a F_{obs}–F_{calc} map. It contains information from both the model map and the difference map.

$$\rho(\vec{r}) = \frac{1}{V} \sum_{h,k,l} (2|F_{\mathrm{obs}}(\vec{S})| - |F_{\mathrm{calc}}(\vec{S})|) \, e^{i\Phi^{\mathrm{calc}}} \cdot e^{-2\pi i \vec{r}\vec{S}}$$

Sigma-A-weighted maps: $2mF_{obs} - DF_{calc}$ & $mF_{obs} - DF_{calc}$

Here, the map coefficients are weighted with figure-of-merits *m* and *D* from a statistical sigma-A analysis. The resulting coefficients are $2mF_{obs}$–DF_{calc}, Φ_{calc}, and mF_{obs}–DF_{calc}, Φ_{calc}. Effectively, these figure-of-merits down-weight high resolution terms for models that are very incomplete and erroneous, leading to less model-biased electron density maps, which are easier to interpret. Sigma-A weighted electron density maps are the preferred maps for iterative model building and refinement.

$$\rho(\vec{r}) = \frac{1}{V} \sum_{h,k,l} (2\,m|F_{\mathrm{obs}}(\vec{S})| - D|F_{\mathrm{calc}}(\vec{S})|) \, e^{i\Phi^{\mathrm{calc}}} \cdot e^{-2\pi i \vec{r}\vec{S}}$$

$$\rho(\vec{r}) = \frac{1}{V} \sum_{h,k,l} (m|F_{\mathrm{obs}}(\vec{S})| - D|F_{\mathrm{calc}}(\vec{S})|) \, e^{i\Phi^{\mathrm{calc}}} \cdot e^{-2\pi i \vec{r}\vec{S}}$$

Anomalous difference Fourier map

This map uses the observed anomalous differences F^{+}_{obs}–F^{-}_{obs} and calculated phases Φ_{calc}. The result is a very simple map with positive peaks at positions of anomalous scatterers. It is a very useful map to localize heavy atoms after initial phases have been obtained.

$$\rho(\vec{r}) = \frac{1}{V} \sum_{h,k,l} (|F^{+}_{\mathrm{obs}}(\vec{S})| - |F^{-}_{\mathrm{obs}}(\vec{S})|) \, e^{i\Phi^{\mathrm{calc}}} \cdot e^{-2\pi i \vec{r}\vec{S}}$$

11 Model Building & Refinement

11.1 Learning goals

Based on the autotracing main-chain model from SHELXE, you will assign the amino acid sequence to the electron density map, mutate the poly-alanines to the correct amino acids, renumber the residues and rename the chain. You will then refine your first model and inspect the resulting electron density maps to build missing parts, correct any errors, and put water molecules into reasonable electron density peaks. You will iterate over model building and refinement until convergence (almost).

Finally, you will calculate an anomalous difference density map and locate the active site Mn^{2+} using model phases and observed synchrotron data after a manganese ion soak.

11.2 Model building

After getting an initial electron density map, the next step towards a final structure of a biological macromolecule is to interpret the electron density map. A model is built into the electron density map that explains as good as possible both the observed electron density and any other chemical knowledge of the macromolecule (e.g. sequence of amino acids, reasonable stereochemistry, water, ions,...). If the resolution and quality of the phases are sufficiently good, map interpretation is usually straightforward.

For electron density map interpretation, we use the computer graphics program Coot that you've already used for inspecting the initial electron density maps. Coot is a relatively new program and has become the most popular graphics program in crystallography, because it is relatively easy to use. Coot provides the following features:

- 3D stereo visualization
- Display of continuous maps
- Model building tools
- Model refinement tools
- Validation tools
- Various model and electron density display features

11.3 Refinement

The basic idea of crystallographic refinement is quite simple: we want to optimise the simultaneous agreement of an atomic model both with the observed data and with known chemical information. It can be formulated as a chemically restrained non-linear optimisation problem.

The agreement between calculated model structure factors and observed structure factors is usually represented by an R-factor defined as:

$$R = \frac{\sum ||F_{\mathrm{obs}}| - |F_{\mathrm{calc}}||}{\sum |F_{\mathrm{obs}}|}$$

During refinement, the positional parameters and temperature factors of all non-hydrogen atoms are adjusted to minimize both the difference between calculated and observed structure factor amplitudes and the deviation from ideal stereochemical parameters. In X-ray crystallography, hydrogen atoms can only be observed with diffraction data at atomic resolution (1.2 Å or better). At lower resolutions, hydrogen atoms are usually added at fixed, non-refined relative positions to their bonded non-hydrogen atoms (riding atom model), or are not included at all.

For a well-defined optimisation, the problem should be highly over-determined, i.e. there should be more observations than adjustable parameters. Unfortunately, this is usually not the case for protein structure refinement.

For a typical protein crystal with an average solvent content of ~50% and four parameters per atom (x, y, z and B), the observations-to-parameters ratios are:

Resolution	**Observations/Parameters**
3.5 Å	0.5
3 Å	0.8
2.5 Å	1.4
2 Å	2.8
1.5 Å	6.2

At about 2.8 Å, the ratio of observations to parameters is "1". Below this resolution, we could fit the observations with to an arbitrarily good value (overfitting) that does not reflect the quality of the model! In addition, for a well-behaved stable refinement, we need a ratio greater than "1". In simple

words, a low ratio of observations to parameters can lead to overfitting where the model has significant errors but spuriously good agreement with the observations.

What can we do to improve this situation?

- We go back to the bench and try to get better diffracting crystals, so that we get higher resolution data. (Which might be extremely difficult and time consuming, if possible at all)
- We reduce the number of refined parameters. This can be done, for instance, by refining group-B-factors for whole amino acids instead of their individual atoms, or in case of the presence of non-crystallographic symmetry, by forcing the individual copies of our molecule to have identical geometry and B-factors.
- We compensate the lack of observations by including more information in form of ideal stereochemical parameters which are known from small molecule structures determined at atomic resolution:
  - bond lengths
  - bond angles
  - planar groups
  - van der Waals contacts
  - preferred torsion angles

In general, there are two way how to improve the poor ratio of observations to parameters:

- We can reduce the number of refined parameters in form of *constraints*, where we fix certain parameters at their expected values.
- We can increase the number of observations in form of *restraints*, where we introduce ideal stereochemical parameters which should be reached within a given uncertainty (expressed as standard deviation around the ideal value or as energy terms).

The large solvent channels in protein crystals are filled with disordered solvent molecules (water) and can be modelled as a flat electron density with only a few parameters.

11.4 Cross-validation – the free R-factor

During refinement, usually the decrease of the crystallographic R-factor is monitored as a measure for success. However, it can be shown that due to over-fitting in the refinement, the crystallographic R-factor can reach surprisingly low values even for incorrect protein models. Therefore, the *free R-factor* was introduced as a safer measure against over-fitting and as a general guide for choosing optimum refinement protocols. Here, the observed data set is divided into a *working set* and a *test set*. The test set is a random selection of typically 5% of the observed reflections. Refinement is carried out using only the working set, for which the *working R-factor* is calculated. The test set is never used in refinement, but the refinement progress is monitored by calculating the *free R-factor* for this test set. This separation of the observed data into a working set and a test set and the monitoring of the two R-factors, R_{work} and R_{free}, is called *cross-validation*.

If the refinement really improves the model, both the R_{work} and the R_{free} will decrease. But if the refinement decreases the R_{work} because of over-fitting, the R_{free} will increase. Controlling the values and behaviour of both R-factors is essential throughout the refinement process!

11.5 Temperature factor, B-factor

The temperature factor or B-factor can be thought of as a measure of how much an atom oscillates or vibrates around the position specified in the model. Atoms at side-chain termini are expected to exhibit more freedom of movement than main-chain atoms. This movement distribute each atom over a small region of space.

Diffraction is affected by this variation in atomic position and is expressed as a temperature factor, or B-factor, for each atom, and is usually included as a parameter in the refinement. From the temperature factors computed during refinement, we observe which atoms in the molecule have the largest freedom of movement, and we gain some insight into the dynamics of our largely static model. In addition, adding the effects of motion to our model makes it more realistic and hence more likely to describe the observed data.

$$B = 8\pi^2\langle u^2\rangle \approx 79\langle u^2\rangle$$

If the temperature factor B is purely a measure of the thermal motion of an atom, then in the simplest case of a purely harmonic thermal motion with equal magnitude in all directions (called isotropic vibration), B is related to the magnitude of vibration as given in the following examples:

If the refined B-factor of an atom is 80 $Å^2$, the total mean-square displacement of that atom due to vibration is ≈1.0 $Å^2$, and the root-mean-square (rms) displacement is also ≈1.0 Å. If the B value is 20 $Å^2$, the corresponding rms displacement is ≈0.5 Å. If the B value is 5 $Å^2$, the corresponding rms displacement is ≈0.25 Å. But the B values obtained for most proteins are too large for describing purely thermal motion and include other sources of disorder as well.

11.6 Maximum likelihood

Most modern refinement programs use the concept of *maximum likelihood*:

- the basic idea is quite simple: the best model is most consistent with the observations
- consistency is measured statistically, by the probability that the observations would be made, given the current model
- if the model is changed to make the observations more probable, the model gets better and the likelihood goes up
- the probabilities have to include the effects of all sources of error, including errors in the model. But as the model gets better, the errors get smaller and the probabilities get sharper, which also increases the likelihood
- least-squares is a special case of maximum likelihood where the model is almost complete and correct, and errors in the data are only statistical

11.7 Building the model with Coot

- start "`coot &`" in a terminal
- open the autotracing main chain model from SHELXE with the correct hand with "File" → "Open Coordinates"
- open the electron density map with "File" → "Open MTZ ...", and choose the corresponding *.phs file
- switch on the display of symmetry-equivalent atoms with "Draw" → "Cell & Symmetry"

- try to assign the prim/pol amino acid sequence to the electron density map; very helpful are the selenium atoms (why?), large aromatic residues, especially tryptophane, and glycines
- mutate the poly-alanine model either by clicking on "Mutate & AutoFit" for single residues, or with "Calculate" → "Mutate Residue Range" for stretches of residues
- for manually fitted residues, do some real-space refinement including neighbouring residues; please, reduce the relative weight for the fit to the electron density to a value around "10" by clicking on "R/RC" before real-space refinement
- for adding missing residues at the end of chain fragments, you can simply click on "Add residue"
- you can also try to let Coot fit stretches of missing residues automatically with "Calculate" → "Fit Loop"
- if an α-helix or β-strand is missing in the electron density, you can add it by centering on that electron density, followed by "Calculate" → "Other Modelling Tools" → "Place Helix Here" or "Place Strand Here"; you can merge your α-helix or β-strand into your main molecule with "Calculate" → "Merge Molecules"
- for each fragment, check and correct the residue numbers with "Calculate" → "Renumber Residues"; please, think before renumbering - check for existing numbers in the same chain!
- after the numbering is correct, you can rename the chain identifiers to "A" with "Calculate" → "Change Chain IDs"; here, you must give the old chain name and start and end residue numbers
- skip parts that you can't interpret, yet
- please, note: SHELXE builds incomplete alanines at the ends of the chain fragments; you have to first delete any incomplete alanine and then add a residue in its place
- please, remember: this protein contains *seleno*-methionine!
- save your modified model with "File" → "Save Coordinates" frequently!
- Finally, delete the selenium atoms in the pdb file with a text editor (gedit or kate)

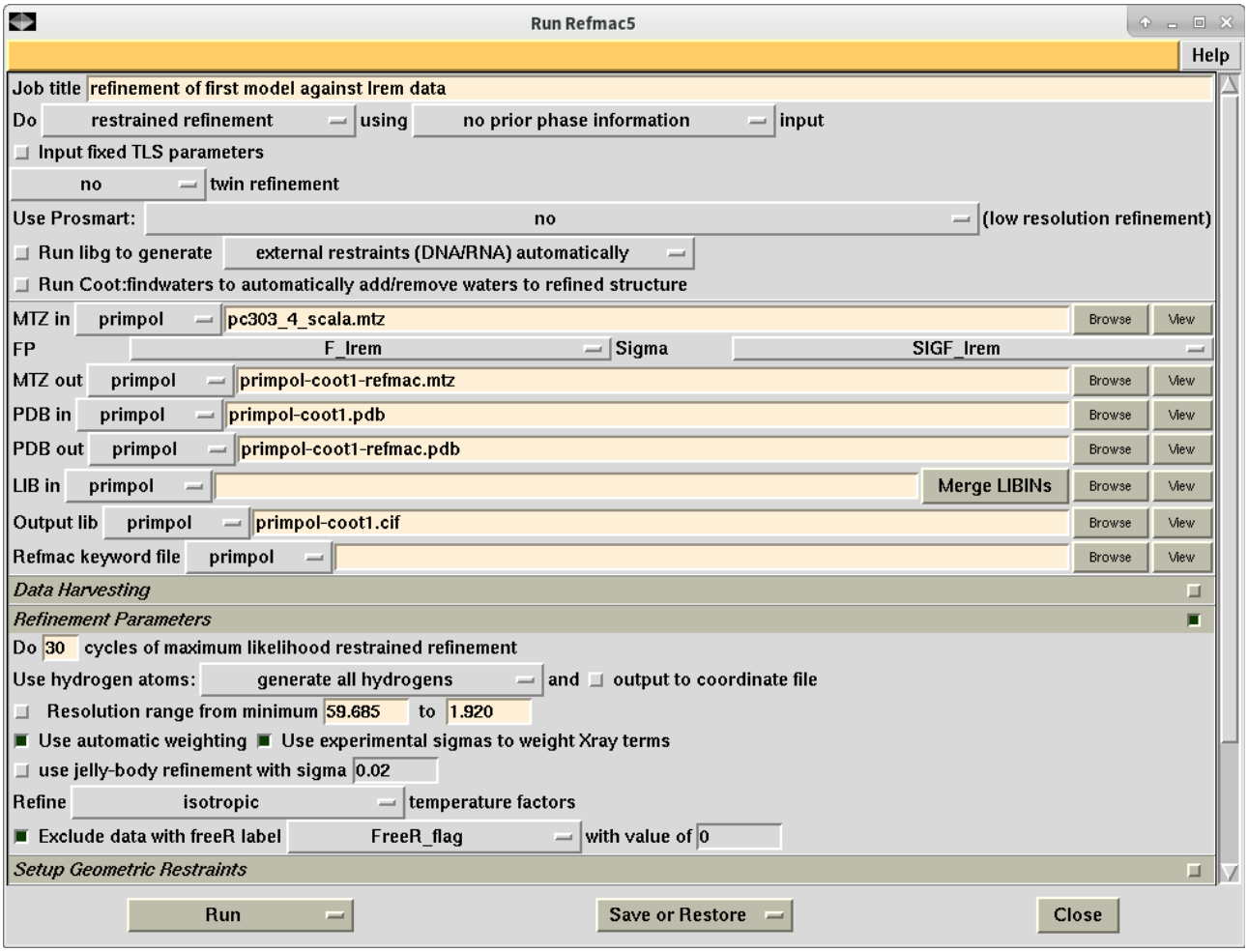
The amino acid sequence of the N-terminal domain is given below. Please, note: the sequence does not contain the methionine mutations!

Amino acid sequence of pRN1 40-255

```
  40 S SERIRYAKWF LEHGFNIIPI DPESKKPVLK EWQKYSHEMP    80
  81   SDEEKQRFLK MIEEGYNYAI PGGQKGLVIL DFESKEKLKA   120
 121   WIGESALEEL CRKTLCTNTV HGGIHIYVLS NDIPPHKINP   160
 161   LFEENGKGII DLQSYNSYVL GLGSCVNHLH CTTDKCPWKE   200
 201   QNYTTCYTLY NELKEISKVD LKSLLRFLAE KGKRLGITLS   240
 241   KTAKEWLEGK KEEED                             255
```

11.8 Refinement with Refmac5

After model building, you refine your first model against the low energy remote data set with Refmac5. In the CCP4 GUI, choose the module "Refinement" and the task "Run Refmac5". Fill in the names of the low-energy remote data set and of your first model, and choose 30 cycles of maximum likelihood refinement. Hint: rename the mtz output file similar to the pdb output file.



Check the R-factors of the refinement run and compare the starting and final values. What do you expect? What do you observe?

Inspect the refined model and the corresponding electron density maps in Coot by opening your refined model ("File" → "Open Coordinates") and the electron density maps after refinement ("File" → "Auto Open MTZ"). Don't forget to display the symmetry equivalent atoms!

Are the electron density maps after refinement clearer than the initial electron density map? Do the electron density maps after refinement show where the model could be further improved?

For regions, where you can see how to improve your model, compare the "2mFo-DFc" and "mFo-DFc" electron density maps after refinement with maps using Fourier coefficients "FP" & "PHIC" and "FC" & "PHIC" ("File → Open MTZ", choose the MTZ file after refinement, select the appropriate Fourier coefficients). Which electron density map shows best how to improve your model, which map shows it worst (model bias)?

Make some nice pictures of example regions for the protocol.

Complete any missing residues in your model and correct any errors in the part that you have built. If the protein model is (more or less) complete, you may add waters with "Calculate" → "Other Modelling Tools" → "Find Waters" → choose the difference density map and set the threshold to 3.0*rmsd.

Iterate over model building and refinement until you reach R-factors in the low 20% range.

11.9 Validation with Coot

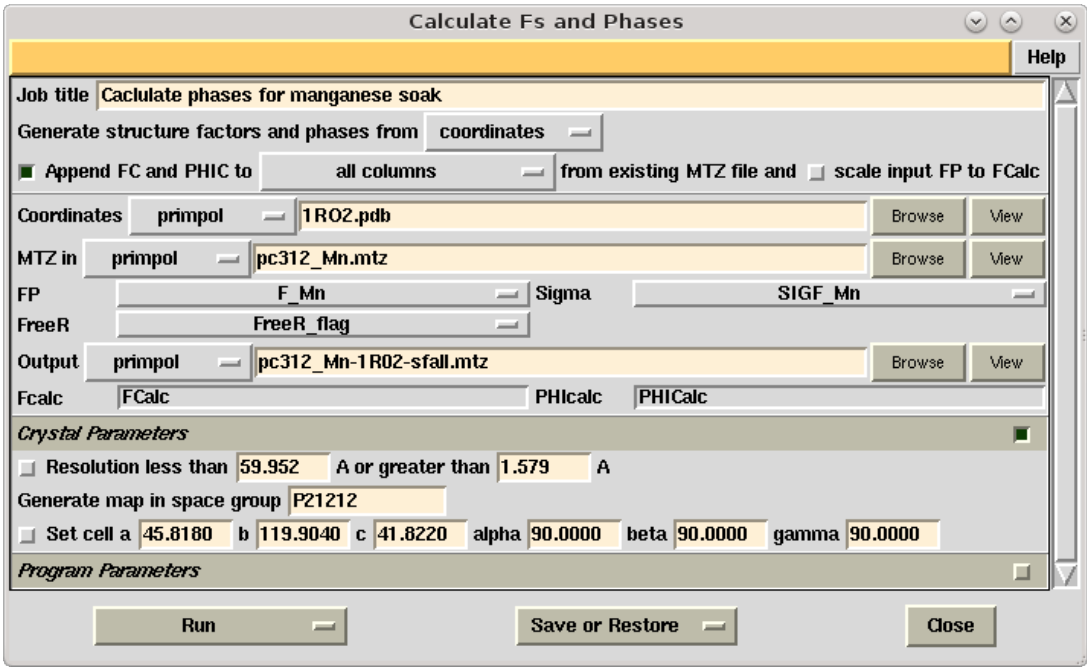
Coot has some very useful tools for validating the structure, which can be found under "Validate". Check your model for Ramachandran outliers, incorrect chiral volumes, rotamers. You may also try the other validation tools.

11.10 Anomalous difference density map

You will use the anomalous signal of manganese after a manganese ion soak from a processed synchrotron data set to identify the active site ion.

Calculate phases from the refined prim/pol coordinates `"1RO2.pdb"` after a manganese ion soak with the module "Reflection Data Utilities" and task "Calculate Fs & Phases". Append the

calculated amplitudes, "FC", and phases, "PHIC", to all columns after the manganese ion soak of the MTZ file `"pc312_Mn.mtz"`.



Load 1RO2.pdb and the output mtz file into Coot and calculate an anomalous difference density map with: "File" → "Open MTZ", activate the "Expert Mode", and choose the amplitudes "DANO_Mn" and phases "PHIC". Try to locate the active site manganes in the anomalous difference density map. What is the height of the anomalous difference density peak at the managanese site?

Make a picture of the anomalous difference density map and the manganese at the active site. Do you see any other anomalous difference density map peaks? If yes, which atoms show anomalous signals?

11.11 Table and figures for publication

Table for publication

Prepare a table with data processing and refinement statistics as for a publication, similar to table 2 in the Lipps et al. paper (but without native and manganese soak data sets).

Figures for publication

Use the computer graphics program PyMOL (http://www.pymol.org) for making figures as for a publication. PyMOL can be invoked from a terminal with the command `"pymol &"`. Make pictures that show interesting aspects of your structure, like overall fold, charge distribution, active site manganese, etc. You can use the figures in the Lipps et al. paper as a rough guide.

12 Appendix: Getting started with COOT

1 Running COOT

The program COOT is relatively new and already the most widely used graphics program for model building and real space refinement in the community of macromolecular crystallographers. COOT binaries are available for Linux, MacOS X, SGI, and Windows. Compared to other model building programs like O and Main, it is very easy to use and can interact with the CCP4 program suite. More information can be found at the COOT website, http://lmb.bioch.ox.ac.uk/coot/.

In a typical model building session, electron density maps are used to build a model of a protein such that the model atoms explain both the electron density and chemical knowledge as good as possible. Initial electron density for structure determination can be calculated from experimental phases or from phases calculated on the basis of the structure of a similar protein (molecular replacement).

To start COOT, type the following command in the terminal window:

```
% coot &
```

You will see the COOT graphical user interface (GUI) appear. You are now ready to use COOT. If you started COOT already before from the same directory, the program will ask you if you wish to continue your old session.

2 Reading in coordinates

Coordinates of proteins or nucleic acids are usually stored as PDB files and can be downloaded for all published structures from http://www.pdb.org.

In the main menu-bar, click on "File" > "Open Coordinates"

[COOT displays a Coordinates File Selection window]

Select your PDB-file of interest (XXX.pdb) from the "Files" list. Click "OK".

[COOT displays the coordinates in the Graphics Window]

Alternatively, if you know the PDB-identifier of a published structure (e.g. 1mbn.pdb), COOT allows you to directly download this file from the PDB *via* internet by clicking on:

"File" > "Get PDB using Accession Code"

Once loaded, you can drag, rotate, and translate the molecule with the mouse (for more details, see "Keyboard/Mouse Commands in COOT ").

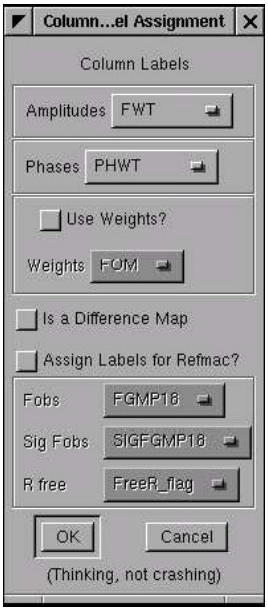
3 Load & display maps

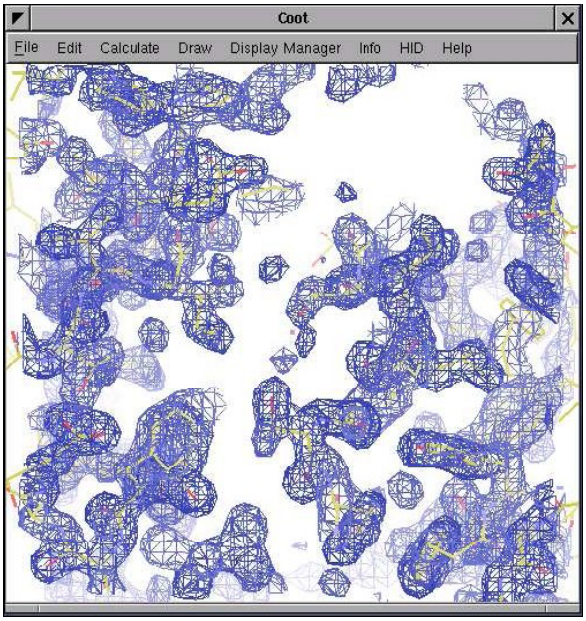
We can load electron density maps that have been generated from experimental data after phasing and density modification. Visual inspection of electron density maps with programs like COOT is an efficient way of determining the success of phasing approaches.

Refinement programs store their data (labelled lists of structure factor amplitudes and phases) in a so-called "mtz" file (XXX.mtz). COOT can use mtz-files directly to calculate electron density maps on-the-fly. From the COOT menu-bar, click on "File" > "Auto Open MTZ..." menu item

[COOT displays a Dataset File Selection window].

Select an MTZ file and open it.





If you choose instead "Open MTZ, cif or phs...", Coot displays a File Selection window:

Select a filename[COOT displays a Dataset Column Label Selection window,]

You will have to select column labels for the amplitudes and phases (defaults are "FWT" and "PHWT").

Press "OK" in the Column Label Window

You will see an electron density map in the main window.

4 Loading other file types...

COOT is also able to read additional formats of coordinate, and map files (in the "File" menu). It can also load user-specific libraries and coordinate files of small molecules (SMILES).

5 Deleting coordinates and maps from COOT

Sometimes, you wish to remove a previously loaded file from COOT. To unload (delete) a coordinate or map file, do the following:

In the main menu bar, click on "File" > "Close molecule/map..."

Choose the file you like to remove and click "Delete"

6 Controlling the view

6.1 Adjust virtual trackball

By default, Coot has a "virtual trackball" to relate the motion of the molecule to the motion of the mouse. If you don't like this, you might want to try the following:

In the Coot main menu-bar:

Select "HID" > "Virtual Trackball" > "Flat"

(Use the "Spherical Surface" option to turn it back to how it is by default)

6.2 Displaying or hiding coordinates and maps

To hide or show maps and coordinates, click in the main menu bar on "Display Manager". A new window will appear, in which all loaded files are displayed. Clicking the "Display" button beneath each filename, turns visibility on or off. Additional features of map and coordinate display can be changed by clicking on the respective button "Properties" in the same window.

6.3 Change background colour

The background colour can be changed by clicking in the main menu bar on

"Edit" > "Background Colour" (Black or White)

6.4 Change map parameters

To change the map colour, click in the main menu bar on

"Edit" > "Map Colour"

To change the radius (size) and quality of a displayed map, click on

"Edit" > "Map Parameters".

Note: Displaying electron density maps, requires much computing power. Increasing the Map Radius to high values usually results in slow-down of the screen display.

Depending on computer power, "Safe values" are usually between 15 and 40 Å.

6.5 Change the clipping (slab)

From the Coot menu-bar, select "Draw" > "Clipping. . . " and move the slider.

[Coot displays a Clipping window]

Click "OK" in the Clipping window

Alternatively, you can use "D" and "F" on the keyboard,

or Control Right-mouse up/down (Control Right-mouse left/right does z-translation).

6.6 Re-contour the map

Scroll your scroll-wheel forwards one click

Scroll your scroll-wheel forwards and backwards and see the contour level changing.

Alternatively, you can use "+" and "-" on the keyboard.

Note that the "Scroll" button in the "Display Manager" selects which map is affected.

7 Model building and refinement

The functions described below are designed to build an atomic model into existing electron density and to perform different types of refinement on the model as well as on the electron density.

Most functions described below can be found in a window opened by clicking on: "Calculate" > "Model/Fit/Refine" and "Calculate" > "Other Modelling Tools".

Note: In order for the external program Refmac to work, you must have loaded the respective map with "Open MTZ, cif or phs...", with the button "Assign Labels for Refmac?" pressed and labels assigned correctly.

7.1 A useful visual guide: the map skeleton

A so-called map-skeleton displays stretches of connected electron density as stick skeletons. Particularly for noisy initial maps, this visualization can be very helpful in identifying regions of continuous electron density and possible polypeptide chains.

Click on "Calculate" > "Map Skeleton", turn skeleton mode "On".

You can click on "Colour and Prune", to delete possible side chain points, so that your skeleton looks cleaner. This procedure creates a skeleton (points and connection) in the map.

7.2 Cα-baton mode for model building

The Baton mode is a tool that allows fast building of protein-main chains. To open the Baton mode: Put the crosshair in the main graphics window on a skeleton point where a continuous polypeptide chain can be seen in the map.

Click "Calculate" > "Other Modelling Tools" > "Ca Baton Mode".

This builds a Cα-atom at the crosshair and connects that to another point 3.8 Å away (distance of Cα-Cα). You can change where the 2nd Cα is by choosing "Try Another".

You can also lengthen/shorten the baton with the appropriate options. When you're satisfied with the baton, click "Accept". If you make a mistake, click "Undo".
Keep building the Cα frame until there is no clear direction on where to add the next Cα atom.
In order to convert the Cα trace into a polypeptide mainchain, click on "Calculate" > "Other Modelling Tools..." > "Ca Zone > Mainchain". Click on the range of the Cα trace (e.g. residue #2 and #10) that you want to convert into mainchain atoms. COOT will process this request (this takes a while) and produces a mainchain configuration - if possible. Sometimes, the connection you make violates the allowed geometry of mainchain atoms. In this case, nothing will be built.

7.3 Add additional N- or C-terminal residue

This creates a new residue at the C- or N- terminus and tries to fit it into the electron density map. Φ/Ψ angle pairs are selected at random based on the Ramachandran plot probability (for a generic residue). It is possible that a wrong position will be selected for the terminal residue and if so, you can reject this fit and try again with Fit Terminal Residue. Each of the trial positions are scored according to their fit to the map and the best one selected. It is probably a good idea to run "Refine Zone" on these new residues.

7.4 Mutating residues

7.4.1 Single mutations

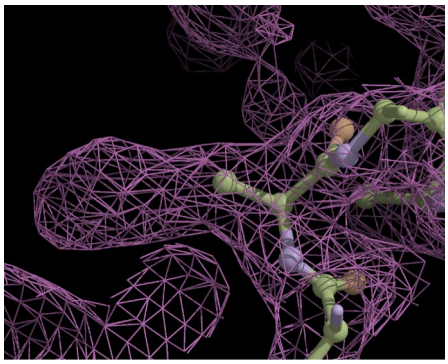
Mutations can be done on a 1-by-1 basis. After selecting "Mutate..." from the "Model/Fit/Refine" dialogue, click on an atom in the graphics window. A "Residue Type" window will now appear. Select the new residue type you wish and the residue in the graphics is updated to the new residue type. Initially, the most probable rotamer is chosen.

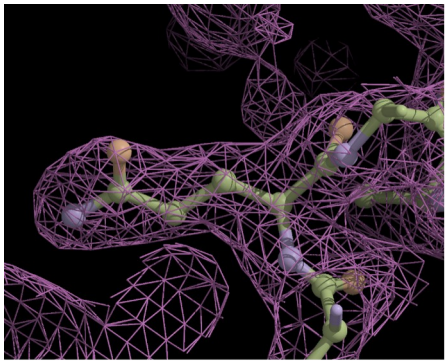
7.4.2 Multiple mutations

This dialogue can be found under "Calculate" → "Mutate Residue Range". A residue range can be assigned a sequence and optionally fitted to the map. This is useful for converting a poly-ALA model to the correct sequence.

7.4.3 Mutate and auto-fit side chains

The function combines "Mutation" and "Auto Fit Rotamer" and is the easiest way to make a mutation and then fit to the map.
Example mutation of Ala to Gln:



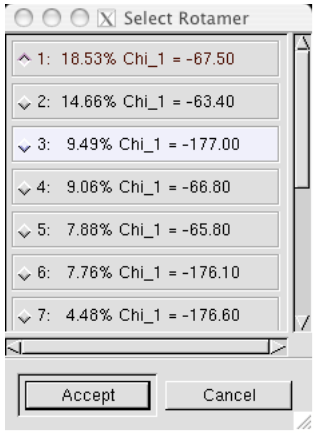


7.5 Add OXT atom to the C-terminal residue

At the C-terminus of a chain of amino-acid residues, there is a "modification" so that the C-O becomes a carboxyl, i.e. an extra terminal)oxygen (OXT) needs to be added. This atom is added so that it is in the plane of the Cα, C and O atoms of the residue.

7.6 Rotate side chains into the "correct" position

7.6.1 Rotamers



At the early stage of model building, side chains of amino acids often point to the wrong direction outside the electron density. To fix this, in the "Model/Fit/Refine..." window, click on "Rotamers". In the graphics window, (left-mouse) click on an atom. COOT now displays the "Select Rotamer" window.

Choose the Rotamer that most closely puts the atoms into the density. You can toggle through the rotamers with the "," and "." keys. Click "Accept" if you are satisfied with a rotamer. You can undo this operation by clicking on the "Undo" button.

Note: It is often necessary to improve the fit of the side chain to the electron density map. This can be done with *Real Space Refinement* (see below paragraph "Real Space Refinement" in this section).

7.6.2 Auto fit rotamer

You can also try to let COOT automatically fit a residue into the electron density map using "Auto Fit Rotamer" (in the "Model/Fit/Refine" window)

7.6.3 Edit Chi angles

This function allows you to manually rotate each bond angle of a side chain. This

function is normally only useful for large, non-rigid amino acids, for which good electron density is available but none of the above described functions provide sufficient overlap of the model and electron density.

In the "Model/Fit/Refine" window click on "Edit Chi Angles". Then click on the side chain to modify. A new window appears that has buttons for all angles. Click on one button, press the left mouse button in the main graphic window and move the mouse. The bond angle changes and is displayed at the top of the graphics window. This function should only be used with care!

7.7 Refine whole regions of a molecule

7.7.1 Geometry refinement (regularize)

"Regularize Zone" in COOT uses ideal geometry parameters to improve the stereochemistry of the model.

Select "Calculate" > "Model/Fit/Refine" > "Regularize" and click on two atoms. You can also click on the same atom twice if you only want to regularize one residue. COOT then regularizes the residue range. At the end, COOT displays the intermediate atoms in white and also displays a dialogue, in which you can accept or reject this regularization.

7.7.2 Real-space refinement

The use of "Real Space Refinement" is similar to Regularization but with the addition of using a map. The map used to refine the structure is set by using the "Refine/Regularize Control" dialogue. If you have read/created only one map into COOT, then that map will be used (there is no need to set it explicitly).

Click on "Calculate" > "Model/Fit/Refine" > "Real Space Refine Zone"

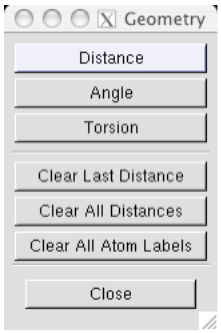
For both "Regularize Zone" and "Refine Zone" one is able to use a single click to refine a residue range. Pressing A on the keyboard while selecting an atom in a residue will automatically create a residue range with that residue in the middle. By default the zone is extended one residue either size of the central residue. This can be changed to 2 either side using (set-refine-auto-range-step 2). To prevent the unintentional refinement of a large number of residues, there is a limit of 20 residues.

7.8 Maximum-likelihood refinement with Refmac

COOT also allows you to interact with external refinement programs like Refmac. Use

the "Run Refmac...." button to select the dataset and the coordinates on which you would like to run Refmac5.

7.9 Measure angles/distances



In order to measure atomic distances or angles in atomic models, click on "Measure" > "Distances & Angles". In the new window "Geometry" click on "Distance". Then use the mouse to click on the two atoms, for which you wish to measure their distance.

7.10 Validation

There are several ways to analyse structural problems and several of them are available in COOT. Open the main menu bar "Validate", where you find a large collection of useful tools to detect problematic or incomplete regions in your built model.

7.11 Keyboard/Mouse Commands in Coot

Keyboard

Rotation

Q	Rotate + X Axis
W	Rotate - X Axis
E	Rotate + Y Axis
R	Rotate - Y Axis
T	Rotate + Z Axis
Y	Rotate - Z Axis

Translation

Keypad 3	Push View (+Z translation)
Keypad .	Pull View (-Z translation)

Clip

D	Slim clip
F	Fatten clip

Contouring

Use + or - to change the contour level

Undo

Ctrl-Z	Undo last modification
U	Undo last move/navigation

Previous/Next Residue

“Space”	Next Residue
“Shift” “Space”	Previous Residue

Previous/Next Rotamer

When in “Rotamer ” mode, these keyboard short-cuts are available:
“.” Next Rotamer
“,” Previous Rotamer

Mouse

(L-, M- R-Mouse means
Left-, Middle, Right-Mouse Button)

L-mouse Drag	Rotate view
Ctrl L-Mouse Drag	Translates view
Shift L-Mouse Click	Label Atom
R-Mouse Drag	Zoom in and out
Shift R-Mouse Drag	Change clipping and Translate in Screen Z up+right/down+left shifts in z, up+left/down+right changes the slab
Ctrl Shift R-Mouse Drag	Rotate View about Screen Z
M-mouse Click	Centre on atom
Scroll-wheel Forward	Increase contour level
Scroll-wheel Backward	Reduce contour level