Macromolecular Structure Determination

Part II: Symmetry, Space Groups and Data Integration

Tim Grüne

Dept. of Structural Chemistry, University of Göttingen

September 2010

http://shelx.uni-ac.gwdg.de

tg@shelx.uni-ac.gwdg.de
So Far, So Good . . .

Crystals produce a “regular” pattern of spots, the **diffraction pattern**, when held into X-rays.

With *some effort*\(^a\) these spots can be turned into a beautiful model of the molecule inside the crystal.

The first step is **data integration**, *i.e.* the determination of spot locations (which corresponds to the unit cell parameters by means of the Laue conditions) and their intensities.

\(^a\)How — that is what this lecture is all about . . .
Symmetry and Space Groups
Crystallography and Symmetry

Historically crystallographers described the appearance of minerals and their regularities. E.g. Nicolaus Steno formulated the law of constant angles in 1669, long before the advent of X-rays.

1801 René-Just Haüy describe the symmetry of crystals (after group theory had been developed).

1850 Auguste Bravais describes the 14 different Bravais lattices.

1890/1891 Arthur Moritz Schönflies and Jewgraf Stepanowitch derive the 230 possible space groups.

1912 Max von Laue, Walter Friedrich, and Paul Knipping carry out the first diffraction experiment and show the wave nature of X-rays and the lattice structure of crystals.
The Use of Symmetry

Historically it was certainly a matter of curiosity to realise that crystals obey certain rules of repetition and regularity (that’s what symmetry is about).

In principal one could solve a structure without taking symmetry into account.

There are two important advantages of taking symmetry into account:

1. Improvement of data quality by increasing the accuracy of the measurement
2. Reduction of work. E.g. ignoring a 4-fold symmetry one would have to refine four molecules which are basically identical.

The aim of this section: Understanding both aspects.
Symmetry in Nature

Symmetry is a *mathematical concept* with its *origin* in nature:

- Butterflies: mirror plane
- Flower with 5-fold rotational symmetry

In arts and also sociologically symmetry is often associated with beauty.
Symmetry in Molecules

Benzene: 6-fold rotational symmetry, mirror planes

single macromolecules (Protein, DNA, RNA) are never symmetric.
Symmetric Arrangements

Any object, symmetric or not, can be \textit{arranged} in a symmetric way.

Five ribosomes arranged with a 5-fold rotation axis.

Note that the ribosome cannot be arranged to have a mirror plane, because it consists of \textit{chiral} compounds.
Symmetry Operations

Loosely speaking, a symmetry operation is a movement that leaves (at least *the appearance* of) the object unchanged.

There are three **basic types** of symmetry operations:

<table>
<thead>
<tr>
<th>Rotation</th>
<th>Mirror Plane</th>
<th>Inversion</th>
</tr>
</thead>
<tbody>
<tr>
<td>We speak of an <em>n-fold</em> symmetry (axis) when <em>the movement</em> is a rotation about (\frac{360^\circ}{n}). E.g., the angle between one ribosome and the next on the previous slide is (\frac{360^\circ}{5} = 72^\circ).</td>
<td>![Mirror Plane Image]</td>
<td>![Inversion Image]</td>
</tr>
</tbody>
</table>
Combination of Symmetry

One can combine symmetry operations. This often generates additional symmetries:

- Mirror the butterfly to create a second one.
- Rotate both butterflies by 180° now there are four butterflies.
- The whole composition contains a new mirror plane, generated by the combination of the first mirror plane and the rotation.
Combination of Symmetry

One can combine symmetry operations. This often generates additional symmetries:

- Mirror the butterfly to create a second one.
Combination of Symmetry

One can combine symmetry operations. This often generates additional symmetries:

- Mirror the butterfly to create a second one.
- Rotate both butterflies by 180° - now there are four butterflies.
Combination of Symmetry

One can combine symmetry operations. This often generates additional symmetries:

- Mirror the butterfly to create a second one.
- Rotate both by $180^\circ$ - now there are four.
- The whole composition contains a new mirror plane, generated by the combination of the first mirror plane and the rotation.
Screw Axes

One special type of symmetry elements are **Screw Axes**. They are combinations of a rotation by $\frac{360^\circ}{n}$ with a translation along the unit cell axis by $(\frac{k}{n})$ of the axis length. We speak of an $n_k$-fold screw axis.

The figure shows an example of a $4_1$ screw axis:

A rotation by $1/4 \cdot 360^\circ$, i.e. $90^\circ$, is combined with a translation of $1/4$ of the length of the unit cell axis along which the screw axis runs. After four such screws, one comes to a point in the next unit cell which is the starting point translated by the cell axis.

We are going to meet screw axes again when we deal with space group determination.
Symmetry in Crystals

There seems to be an infinite number of possible combinations of the symmetry operations.

In crystallography, however, the possible number is restricted: the symmetry must cooperate with the crystal lattice, and this imposes some restrictions:

- Start with an arbitrary unit cell
Symmetry in Crystals

There seems to be an infinite number of possible combinations of the symmetry operations.

In crystallography, however, the possible number is restricted: the symmetry must cooperate with the crystal lattice, and this imposes some restrictions:

- Start with an arbitrary unit cell
- apply it $90^\circ$ rotation (4-fold rotation axis)
Symmetry in Crystals

There seems to be an infinite number of possible combinations of the symmetry operations. In crystallography, however, the possible number is restricted: the symmetry must cooperate with the crystal lattice, and this imposes some restrictions:

- Start with an arbitrary unit cell
- apply it $90^\circ$ rotation (4-fold rotation axis)
- the gap between the two unit cell cannot be filled by this unit cell. But crystals are not allowed to have gaps.
Possible Symmetries

Because of the restriction of the symmetry operations to match with the lattice, the only possible symmetry operations available for crystals are:

- rotations (only 2-, 3-, 4- and 6-fold axes)
- mirrors and inversion centres (only small molecules!)

and their combinations.

Tim Grüne
Space Groups and Naming Conventions

There are 230 different possibilities for symmetric arrangements within a lattice. They are called the space groups.

There are two different notations for space groups:

1. Herrmann-Mauguin notation, \( e.g. \ P1, I4_{1}32, F\bar{4}3c \). The first letter describes the lattice type (primitive, face centred, . . . ), the rest the symmetries per axis.

2. Schönflies notation, \( e.g. \ C_{1}^{1}, O^{8}, T_{d}^{5} \), which is derived from the mathematical group names.

This course uses the Herrmann-Mauguin notation (if at all . . . ).
Symmetry of Macromolecules

Because macromolecules are chiral, a macromolecule cannot crystallise with a space group which contains an inversion centre or a mirror plane.

This leaves “only” 65 chiral space groups in macromolecular crystallography.

Interestingly, macromolecules tend to crystallise in a high symmetry space group (with many possible symmetry operations), whereas small molecules tend to crystallise in a low symmetry space group.
All spacegroups with their properties (e.g. symmetry operators) are listed in the *International Tables for X-Ray Crystallography*. 

**P222**

- **No. 16**
- **Orthorhombic**
- **P222**

### Generators selected

- (1); (1.0.0); (0.1.0); (0.0.1); (5); (3)

### Positions

- Origin at 0, 0, 0
- Along [010] p 2 mm
- a' = cb' = a

### Minimal non-isomorphic subgroups of nearest index

- (5); [P222] [a.0.b.0] [a.b.0] [a.0.b]

- Minimal non-isomorphic supergroup

- [P222] [a.b.c.0] [a.0.b.c] [a.b.0.c] [a.0.b.c]

### Minimal non-isomorphic supergroups

- [P222] [a.0.b.0] [a.b.0] [a.0.b]
Choosing the Unit Cell

- An artificial crystal from the ribosome.

By convention the unit cell is chosen as small as possible but should also reflect the symmetry of the lattice. In this example, the $90^\circ$ angles make the 2-fold axis (and the two mirror planes) more apparent.
Choosing the Unit Cell

- An artificial crystal from the ribosome.
- It has 2-fold symmetry about the marked axes (not 4-fold!)
Choosing the Unit Cell

- An artificial crystal from the ribosome.
- It has 2-fold symmetry about the marked axes (not 4-fold!)
- One possible unit cell
Choosing the Unit Cell

- An artificial crystal from the ribosome.
- It has 2-fold symmetry about the marked axes (not 4-fold!)
- One possible unit cell
- Another possible unit cell that shows the symmetry.

By convention the unit cell is chosen as small as possible but should also reflect the symmetry of the lattice. In this example, the $90^\circ$ angles make the 2-fold axis (and the two mirror planes) more apparent.
Asymmetric Unit

The *unit cell* is the smallest volume required to build up the whole crystal using *only* translation.

The *asymmetric unit* is the *smallest volume* we need to know in order to reconstruct the whole crystal using *both* translation and the *symmetry operators* of the crystal.

We only need to find the atoms inside the asymmetric unit in order to describe the molecule, all other atoms can be found by symmetry operations.
Crystal Systems

The unit cell parameters $a, b, c, \alpha, \beta, \gamma$ can be classified according to their degree and type of regularity. One speaks of the seven crystal systems, and there are seven of them:

- **Orthorhombic**: $a \neq b \neq c$, $\alpha = \beta = \gamma = 90^\circ$
- **Cubic**: $a = b = c$, $\alpha = \beta = \gamma = 90^\circ$
- **Tetragonal**: $a = b$, $\alpha = \beta = \gamma = 90^\circ$
- **Hexagonal**: $a = b$, $\alpha = \gamma = 90^\circ$, $\beta = 120^\circ$
- **Trigonal**: $a = b = c$, $\alpha = \beta = 90^\circ$, $\gamma = 120^\circ$
- **Monoclinic**: $a \neq b \neq c$, $\alpha = \gamma = 90^\circ$, $\beta \neq 90^\circ$
- **Triclinic**: $a \neq b \neq c$, $\alpha \neq \beta \neq \gamma$
Bravais Lattices

The restriction based on “Choosing the Unit Cell” and the seven “Crystal Systems”, i.e. the combination of crystal symmetry with lattice types, leads to the 14 Bravais Lattices.
Bravais Lattices - Key

The “dots” in the previous presentation represent special positions, i.e. they mark locations of symmetry operators. There do not need to be atoms at these positions.

These are the meanings of the abbreviations of the Bravais lattices:

<table>
<thead>
<tr>
<th>Choice of unit cell</th>
</tr>
</thead>
<tbody>
<tr>
<td>P  primitive</td>
</tr>
<tr>
<td>F  face-centred</td>
</tr>
<tr>
<td>C  C-centred</td>
</tr>
<tr>
<td>R  rhombohedral</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Crystal System</th>
</tr>
</thead>
<tbody>
<tr>
<td>a  triclinic</td>
</tr>
<tr>
<td>o  orthorhombic</td>
</tr>
<tr>
<td>h  hexagonal / trigonal</td>
</tr>
<tr>
<td>m  monoclinic</td>
</tr>
<tr>
<td>t  tetragonal</td>
</tr>
<tr>
<td>c  cubic</td>
</tr>
</tbody>
</table>

Symmetry and X-ray Diffraction

The symmetry of the crystal can be observed on the diffraction pattern:

Diffraction image of Lysozyme (nearly) oriented along its 4-fold axis: Especially at the centre the symmetry becomes visible (Z. Dauter).

The symmetry of the reflections is imposed on the data and used to correct for systematic errors during data collection and hence to improve the data quality.
Further Reading: Symmetry and Space Groups


The Ewald Sphere Construction
The Ewald Sphere Construction

Before we continue with “Data Collection”, we have to introduce the reciprocal lattice and the Ewald Sphere.

While the Laue conditions are merely helpful for programmers, the Ewald Sphere is extremely educational and a powerful tool to understand an X-ray diffraction experiment.

The Ewald sphere “lives” in reciprocal space.
The Reciprocal Lattice — Orthorhombic Case

For an orthorhombic lattice, i.e., all three angles \( \alpha = \beta = \gamma = 90^\circ \), the term *reciprocal* lattice is fairly understandable:

- \( |\vec{a}^*| = \frac{1}{|\vec{a}|} \)
- \( |\vec{b}^*| = \frac{1}{|\vec{b}|} \)
- \( |\vec{c}^*| = \frac{1}{|\vec{c}|} \)

\[ |\vec{b}| = 1 \]

“direct or real space”

“reciprocal space”
The Reciprocal Lattice: Formal Definition

In general, the vectors $\vec{a}^*$, $\vec{b}^*$, $\vec{c}^*$, which span the reciprocal space, are mathematically defined as:

- $\vec{a}^* = \frac{\vec{b} \times \vec{c}}{V}$, i.e. $\vec{a}^* \perp \text{plane}(\vec{b}, \vec{c})$
- $\vec{b}^* = \frac{\vec{c} \times \vec{a}}{V}$, i.e. $\vec{b}^* \perp \text{plane}(\vec{c}, \vec{a})$
- $\vec{c}^* = \frac{\vec{a} \times \vec{b}}{V}$, i.e. $\vec{c}^* \perp \text{plane}(\vec{a}, \vec{b})$

The volume $V$ of the unit cell and the volume $V^*$ of the reciprocal unit cell (the box spanned by $\vec{a}^*$, $\vec{b}^*$, $\vec{c}^*$) always fulfil $V = \frac{1}{V^*}$.

A long “real space vector” corresponds to a short “reciprocal vector”. Does this ring a bell?
The Reciprocal Lattice

The reciprocal lattice are all the points that can be described as

\[ h\vec{a}^* + k\vec{b}^* + l\vec{c}^* \]

with integers \( h, k, l \).

These integers \( h, k, l \) — again — turn out to be the Miller indices of a reflection \((h, k, l)\).
Ewald Sphere Construction

The crystal rotates about the origin of the reciprocal lattice.
Ewald Sphere Construction

Draw a sphere with radius $1/\lambda$ that touches the lattice origin. The sphere centre lies aligned with the X-ray source.

This sphere is the **Ewald Sphere**.
Ewald Sphere Construction

The scattering vector \( \vec{S} \) points from the origin to the lattice point. Exactly those lattice points on the surface of the Ewald sphere fulfil the Laue conditions. They are the recordable reflections.
Ewald Sphere Construction

Some of these spots hit the detector.
Ewald Sphere Construction

Crystal rotation = Lattice rotation = New spots

(Rot. axis perpendicular to slide)
Use of the Ewald Sphere

The Ewald Sphere construction allows to understand the diffraction patterns we observe during data collection. The so-called lunes - the reflection spots arranged in a circular pattern - are the intersection of the lattice points with the surface of the sphere.

By the way the reciprocal lattice is constructed from the unit cell vectors, the reciprocal lattice has the same point symmetry as the direct lattice. This is way the diffraction pattern show the (point) symmetry of the crystal.

The point symmetry is the crystal's symmetry without any translational parts, because the Ewald sphere always stays attached to the (0, 0, 0) lattice point (by construction).
Data Integration
Goal of Data Collection

From a X-ray diffraction experiment we learn the intensities of a large number of reflections*. Every reflection is identified with its Miller index, and the measurement results in a long list of intensities $I(hkl)$. Typically for a macromolecule a dataset contains 10,000-1,000,000 reflections.

Target of data collection and data integration is to determine the intensities of as many reflections as correctly as possible.

Why?

* and an error estimate of the intensities and the unit cell parameters
Goal of Data Collection

Before we can create a model of the molecule(s) inside the crystal we have to determine the electron density map $\rho(x, y, z)$.

The intensity $I(hkl)$ of a reflection can be calculated from the electron density map as

$$I(hkl) = \text{const} \cdot \left| \int_{\text{unit cell}} \rho(x, y, z) e^{2\pi i(hx + ky + lz)} \right|^2$$

We are, though, in the opposite situation: we can measure many of the $I(hkl)$ and want to calculate $\rho(x, y, z)$. Therefore we would have to invert the above equation.
Intensity to Density

The actual inversion of the equation on the previous slide is mostly the topic of phasing, which will be dealt with later.

For now, bear in mind that the more reflections $I(hkl)$ we can measure, the more accurately we can calculate the electron density map $\rho(x, y, z)$, which allows us to build a more accurate model.
What we want to collect

- As many reflections as possible.
- In reciprocal space this means: make as many lattice points as possible transverse the Ewald Sphere.
- This is achieved by rotating the crystal.

Standard setups, e.g. at a synchrotron allows to rotate the crystal around one axis. More sophisticated machines allow to rotate the crystal around more than one axis: one can reach a better completeness of the data.
Caveat to the Ewald Sphere

The Ewald sphere construction shows the reciprocal lattice. One can rotate the crystal which also rotates the reciprocal lattice and hence allows to imagine how and which reflections can be collected.

Bear in mind: Translating (shifiting) the crystal does not move reflections through the Ewald sphere: The Ewald sphere always stays attached to the reflection (000).

Therefore the diffraction pattern only shows the symmetry of the point group of the crystal and not its full symmetry.
How Data are Collected: Frames

Our detector is planar, only two-dimensional. The reflections we want to collect are distributed in three-dimensional space.

If one would rotate the crystal for $360^\circ$ and record everything on the detector, one would not know when each reflection was recorded.

Data are collected as slices, or frames.
Frames

Diffraction images are like computer tomography at a hospital: Many slices are taken from the tissue (brain, leg, etc.) from which the three-dimensional object can be reconstructed.
In X-ray crystallography the same is achieved by rotating the crystal by a small angle while the detector detects the signal. Typically the angle for each image (its frame width) ranges between $0.1^\circ$ to $2^\circ$. One data set consists of a hundred to several thousand images.
Optimal Framewidth

In general the data become better the finer each slice. However, it takes 1800 images to collect a crystal rotation of 180° with a frame width of 0.1°, ten times more than with 1° slices. This also increases the radiation dose the crystal is exposed to and therefore the risk of radiation damage.

Even though data are routinely collected at 100 K, every crystal suffers from radiation damage: the X-rays produce free radicals that in turn break bonds and thus destroy the crystal (lattice).

On average, synchrotron data is collected with 0.5° – 1° frame width; on inhouse sources one often collects with ≈ 0.2°, because the less intense beam causes less radiation damage.
Integration Programs

Popular and less popular programs for data processing (=data integration) include

- XDS
- Mosflm
- Eval
- automar
- HKL2000
- Saint
- d*trek

None of these programs is superior to the others, and it is often worth trying at least two in order to get the best integrated data set.
Determination of the Spot Intensities

1. Cell/Orientation
2. (prelim.) Spacegroup
3. Integration
   * Background
   * Spot area
   * Summation
4. Corrections
5. Spacegroup
6. Scaling
1. **Cell/ Orientation**

1. The **scattering vector** $\vec{S}$ and the **scattering angle** $\theta$ for each reflection $(hkl)$ are “macroscopic” quantities: They can be calculated from
   
   (a) the spot position on the detector 
   
   (b) the distance between crystal and detector

2. The **Laue Conditions** and **Bragg’s Law** relate them to the unit cell parameters $\vec{a}, \vec{b}, \vec{c}$

3. There are enough reflections on 1-2 images to determine the unit cell and its orientation. This step of determining the unit cell dimensions and orientation is called **indexing**, because it is equivalent to assigning to each reflection its Miller index.
2. Spacegroup

The spacegroup that best matches the unit cell dimensions and has high symmetry (many symmetry elements) is chosen:

<table>
<thead>
<tr>
<th>Bravais</th>
<th>Score</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>alpha</th>
<th>beta</th>
<th>gamma</th>
</tr>
</thead>
<tbody>
<tr>
<td>* 31</td>
<td>aP</td>
<td>0.0</td>
<td>92.3</td>
<td>92.4</td>
<td>127.9</td>
<td>90.0</td>
<td>90.0</td>
</tr>
<tr>
<td>* 44</td>
<td>aP</td>
<td>0.0</td>
<td>92.3</td>
<td>92.4</td>
<td>127.9</td>
<td>90.0</td>
<td>90.0</td>
</tr>
<tr>
<td>* 39</td>
<td>mC</td>
<td>0.0</td>
<td>160.0</td>
<td>92.3</td>
<td>127.9</td>
<td>90.0</td>
<td>90.0</td>
</tr>
<tr>
<td>* 10</td>
<td>mC</td>
<td>0.3</td>
<td>160.0</td>
<td>92.3</td>
<td>127.9</td>
<td>90.0</td>
<td>90.0</td>
</tr>
<tr>
<td>* 34</td>
<td>mP</td>
<td>0.5</td>
<td>92.3</td>
<td>127.9</td>
<td>92.4</td>
<td>90.0</td>
<td>120.0</td>
</tr>
<tr>
<td>* 29</td>
<td>mC</td>
<td>0.5</td>
<td>92.3</td>
<td>160.0</td>
<td>127.9</td>
<td>90.0</td>
<td>90.0</td>
</tr>
<tr>
<td>* 38</td>
<td>oC</td>
<td>0.5</td>
<td>92.3</td>
<td>160.0</td>
<td>127.9</td>
<td>90.0</td>
<td>90.0</td>
</tr>
<tr>
<td>* 13</td>
<td>oC</td>
<td>0.8</td>
<td>92.4</td>
<td>160.0</td>
<td>127.9</td>
<td>90.0</td>
<td>90.0</td>
</tr>
<tr>
<td>* 14</td>
<td>mC</td>
<td>0.8</td>
<td>92.4</td>
<td>160.0</td>
<td>127.9</td>
<td>90.0</td>
<td>90.0</td>
</tr>
<tr>
<td>* 12</td>
<td>hP</td>
<td>0.8</td>
<td>92.3</td>
<td>92.4</td>
<td>127.9</td>
<td>90.0</td>
<td>90.0</td>
</tr>
<tr>
<td>35</td>
<td>mP</td>
<td>250.0</td>
<td>92.4</td>
<td>92.3</td>
<td>127.9</td>
<td>90.0</td>
<td>90.0</td>
</tr>
</tbody>
</table>

XDS example output for $P6_122$

Actually, only the Laue Group is of interest during integration. The Laue group is similar to, but not identical to the point group which was mentioned above.
3. Integration

Magnified spot on detector. To measure its intensity:

- estimate the average background (grey)
- estimate the spot area
- count the pixel values of the spot
- subtract the background

Correctly estimating the background and spot area are the crucial parts, especially for weak reflections.
3.1 2D- and 3D-spots

Spots have a certain *volume* and appear on more than one frame.

Some integration programs, *e.g.* Mosflm and HKL2000, treat each frame separately and write the *fraction* of each spot per frame to the output file. They leave it to a separate *scaling program* to put the fraction together. These are 2D-integration programs.

Other programs like XDS and Saint integrate over all frames that contribute to a reflection and only write out the final total intensity per spot. These are 3D-integration programs.

In my experience, fine-slicing (thin frame width) has a greater impact on the data quality with 3D- than with 2D-integration programs.
4. Corrections

The integration step basically consists of counting the pixel values and subtracting the background. Once all measurable reflections are processed, certain corrections must be applied:

- technical corrections like Lorentz- and polarisation-correction
- improved estimate of unit cell dimensions using all data
- improvement of experimental parameters like crystal-to-detector distance, distortions of detector, . . .

It is often worth repeating the whole integration process with the improved parameters.
5. Spacegroup Determination

With all reflections processed and the settings of the experiment (unit cell dimensions, detector distance, . . . ) improved and refined, the spacegroup determination can now be done more accurately than before.

Especially, spacegroups with screw axes show so called extinctions.

E.g. in spacegroup $P2_1$, the reflections (001), (003), (005), . . . are mathematically zero, because the screw axis leads to systematic destructive interference for these reflections. This is the only way to distinguish between $P2_1$ and $P2$. 
6. Scaling

Scaling is a second type of correction. It takes into account that

- the crystal is not spherical: the volume of irradiated crystal changes with crystal orientation (larger volume = higher intensities)
- radiation damage leads to reduction in the scattering power of the crystal
- CCD detectors are made of several “chips”. Each chip may react slightly differently to the impact to X-rays.

Scaling adjusts the data as much as possible as though it came from a perfect crystal measured with a perfect instrument, because this is what the subsequent steps (refinement, building) assume.
6.1 Symmetry Related Reflections

Every symmetry operation can be expressed by a matrix multiplication and a vector addition (translation).

*E.g.* the space group $P4_1$ has the symmetry operator

$$\begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \frac{1}{4} \end{pmatrix}$$

This means, that the reflections* 

$$\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$$

and 

$$\begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} = \begin{pmatrix} -2 \\ 1 \\ 3 \end{pmatrix}$$

should (mathematically) have identical intensities.

*because the Ewald sphere is attached to the (000) reflection, there is no translational part in reciprocal space.
Result of Integration: the \textit{hkl}-file

At the end of the integration step, all hundreds or thousands of images are reduced to the reflections they contain. We end up with a reflection file containing a list of Miller indices each with its intensity and the error estimate:

\begin{verbatim}
2 2 0 10.9258 0.81100
3 0 0 0.86624 0.53398
0 3 0 0.09921 0.79861
1 3 0 59.3246 3.54304
3 1 0 68.3514 3.82527
-1 3 0 53.2978 3.36226
2 3 0 39.5588 2.47039
\end{verbatim}

(Example of a Thaumatin data set in space group \textit{P4}_1\textit{2}1\textit{2}, maximum resolution 1.6 Å, 283,862 reflections in total.)
Resolution of the Data Set

The intensity of the reflections fades as we move towards the edge of the detector (i.e., as we increase the scattering angle $\theta$). There is a maximal angle to which a crystal diffract. This is the resolution limit of the crystal.

- For each reflection we know the angle $\theta$ it forms with the normal between detector and crystal.
- From Bragg’s Law $\lambda = 2d \sin \theta$ we can calculate $d$, the resolution of the reflection.
- The smallest distance to which reasonable data can be measured is called the resolution of the dataset.
Reasonable Data: Determination of the Resolution

There is a problem with the resolution of a dataset:

The integration program does not really distinguish between background and reflections: it calculates the location of the reflections (from the Laue conditions), sums up the pixels in that area and substracts the background.

The crystallographer has to decide about the resolution cut-off.

A good guide for the resolution cut-off is where the average signal divided by its error, $\frac{I}{\sigma_I}$, drops below 2.0.
Example Statistics from the program xprep

<table>
<thead>
<tr>
<th>Resolution</th>
<th>#Data</th>
<th>#Theory</th>
<th>%Complete</th>
<th>Redundancy</th>
<th>Mean I</th>
<th>Mean I/s</th>
<th>R(int)</th>
<th>Rsigma</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inf - 2.15</td>
<td>634</td>
<td>1500</td>
<td>42.3</td>
<td>0.42</td>
<td>261.5</td>
<td>7.61</td>
<td>0.2371</td>
<td>0.1225</td>
</tr>
<tr>
<td>2.15 - 1.84</td>
<td>634</td>
<td>856</td>
<td>74.1</td>
<td>0.75</td>
<td>261.3</td>
<td>7.20</td>
<td>0.1009</td>
<td>0.1267</td>
</tr>
<tr>
<td>1.84 - 1.66</td>
<td>652</td>
<td>914</td>
<td>71.3</td>
<td>0.73</td>
<td>154.7</td>
<td>7.17</td>
<td>0.0548</td>
<td>0.1268</td>
</tr>
<tr>
<td>1.66 - 1.52</td>
<td>678</td>
<td>936</td>
<td>72.4</td>
<td>0.73</td>
<td>94.5</td>
<td>6.70</td>
<td>0.1284</td>
<td>0.1311</td>
</tr>
<tr>
<td>1.52 - 1.42</td>
<td>698</td>
<td>1008</td>
<td>69.2</td>
<td>0.71</td>
<td>79.3</td>
<td>6.62</td>
<td>0.0693</td>
<td>0.1350</td>
</tr>
<tr>
<td>1.42 - 1.34</td>
<td>670</td>
<td>976</td>
<td>68.6</td>
<td>0.71</td>
<td>60.2</td>
<td>5.88</td>
<td>0.1076</td>
<td>0.1449</td>
</tr>
<tr>
<td>1.34 - 1.28</td>
<td>638</td>
<td>904</td>
<td>70.6</td>
<td>0.73</td>
<td>49.6</td>
<td>5.33</td>
<td>0.1229</td>
<td>0.1570</td>
</tr>
<tr>
<td>1.28 - 1.22</td>
<td>726</td>
<td>1102</td>
<td>65.9</td>
<td>0.67</td>
<td>47.9</td>
<td>5.30</td>
<td>0.1469</td>
<td>0.1622</td>
</tr>
<tr>
<td>1.22 - 1.17</td>
<td>691</td>
<td>1132</td>
<td>61.0</td>
<td>0.64</td>
<td>46.6</td>
<td>5.07</td>
<td>0.1338</td>
<td>0.1656</td>
</tr>
<tr>
<td>1.17 - 1.13</td>
<td>656</td>
<td>1038</td>
<td>63.2</td>
<td>0.66</td>
<td>44.8</td>
<td>5.02</td>
<td>0.1779</td>
<td>0.1702</td>
</tr>
<tr>
<td>1.13 - 1.09</td>
<td>703</td>
<td>1164</td>
<td>60.4</td>
<td>0.64</td>
<td>34.7</td>
<td>4.30</td>
<td>0.1747</td>
<td>0.1942</td>
</tr>
<tr>
<td>1.09 - 1.05</td>
<td>818</td>
<td>1402</td>
<td>58.3</td>
<td>0.62</td>
<td>23.7</td>
<td>3.65</td>
<td>0.2062</td>
<td>0.2451</td>
</tr>
<tr>
<td>1.05 - 1.02</td>
<td>641</td>
<td>1198</td>
<td>53.5</td>
<td>0.58</td>
<td>19.8</td>
<td>3.15</td>
<td>0.1933</td>
<td>0.2888</td>
</tr>
<tr>
<td>1.02 - 0.99</td>
<td>726</td>
<td>1380</td>
<td>52.6</td>
<td>0.57</td>
<td>13.9</td>
<td>2.38</td>
<td>0.2356</td>
<td>0.3843</td>
</tr>
<tr>
<td>0.99 - 0.96</td>
<td>791</td>
<td>1460</td>
<td>54.2</td>
<td>0.59</td>
<td>11.7</td>
<td>2.15</td>
<td>0.2367</td>
<td>0.4385</td>
</tr>
<tr>
<td>0.96 - 0.93</td>
<td>826</td>
<td>1696</td>
<td>48.7</td>
<td>0.54</td>
<td>9.8</td>
<td>1.82</td>
<td>0.3466</td>
<td>0.5415</td>
</tr>
<tr>
<td>0.93 - 0.90</td>
<td>738</td>
<td>1998</td>
<td>36.9</td>
<td>0.41</td>
<td>13.6</td>
<td>1.83</td>
<td>0.3125</td>
<td>0.4487</td>
</tr>
<tr>
<td>0.90 - 0.85</td>
<td>569</td>
<td>3392</td>
<td>16.8</td>
<td>0.19</td>
<td>6.3</td>
<td>1.11</td>
<td>0.3375</td>
<td>0.8445</td>
</tr>
<tr>
<td>0.94 - 0.85</td>
<td>1872</td>
<td>6570</td>
<td>28.5</td>
<td>0.32</td>
<td>10.2</td>
<td>1.60</td>
<td>0.3442</td>
<td>0.5529</td>
</tr>
<tr>
<td>Inf - 0.85</td>
<td>12489</td>
<td>24056</td>
<td>51.9</td>
<td>0.55</td>
<td>65.5</td>
<td>4.50</td>
<td>0.1662</td>
<td>0.1587</td>
</tr>
</tbody>
</table>

The third last column suggests to cut the resolution at 0.95 Å.
Further Reading: Data Integration

Summary and Outlook

So far we ended up with a long list of reflections, i.e., with one Miller index for each reflection together with its intensity and error estimate.

This does not suffice to determine the electron density, which we need in order to start building a model of the molecule.

We still require the phases for each reflection.

This is the topic of tomorrow’s lecture.
Part III: Phasing
Phasing

The equation

$$I(hkl) = \text{const} \cdot \left| \int_{\text{unitcell}} \rho(x, y, z) e^{2\pi i (hx + ky + lz)} \right|^2$$

joins X-ray crystallography with chemistry because it shows how the (measured) reflection spots $I(hkl)$ are connected to the electron density $\rho(x, y, z)$ in the crystal.

Unfortunately, this equation reads the wrong way: We want to calculate the electron density from the intensities, because the electron density is needed in order to construct an atomic model for the molecules.

The inversion of the equation is the content of the section phasing.
The Structure Factor

The reflections are the result of small waves from the electrons in the crystal. This notion leads (after some calculations . . . ) to the concept of the structure factor $F(hkl)$. It is a complex number and the builds a nearly two-way bridge between intensities and density:

\[ I(hkl) = \text{const} \times |F(hkl)|^2 \]
\[ F(hkl) = \text{const} \times \int \rho(x, y, z) e^{2\pi i(hx+ky+lz)} \]

The latter equation can be inverted*:

\[ \rho(x, y, z) = \text{const} \times \sum_{h,k,l} F(hkl) e^{-2\pi i(hx+ky+lz)} \]

*the constants in all equations can be derived and do not pose a problem
The Phase Problem

Unfortunately, the structure factor $F(hkl)$ is a complex number. As such it consists of an amplitude $|F(hkl)|$ and a phase $\phi(hkl)$ and can be written as $F(hkl) = |F(hkl)|e^{i\phi(hkl)}$.

The square root of the intensity delivers the structure factor amplitude $|F(hkl)|$.

The phase angle $\phi(hkl)$ cannot be measured directly. This fact is called the phase problem of crystallography.

Without knowing the phases that belong to each reflection, we cannot proceed.
Illustrating the Phase Problem

The fact that the phases do not show up in the diffraction pattern is comparable to drawing a three-dimensional object:

Which side of the cube is the front side?
Illustrating the Phase Problem

The fact that the phases do not show up in the diffraction pattern is comparable to drawing a three-dimensional object:

Which side of the cube is the front side? We cannot decide without further information.
Important Notice

It is (computationally) straightforward to calculate/predict the reflections from a model (which is the final representative of the electron density \( \rho(x, y, z) \)).

Because the phases are only determined indirectly, and because the reflections are the actual experimental result, one usually compares the calculated intensities with the experimental intensities in the subsequent steps (phasing, model building, validation).

A good match between calculated and measured amplitudes indicates we have e.g. a good model or good phases. This is not fool-proof, though, which is why there is a part Validation in this lecture.
Limits of Phasing

All phasing methods provide only an estimate of the phases, and once found the phases must be further improved to get closer to the real phases. Finding this initial phase estimate is phasing.

It is not be obvious, even to more experienced crystallographers, that the improvement of these phases is the role of model building and refinement.
Overview of Phasing Methods

The most common methods to solve the phase problem are:

- Molecular Replacement
- Isomorphous Replacement
- Anomalous Dispersion
Molecular Replacement
Structural Similarity

Proteins are alike! Proteins consist mainly of helices and beta sheets. Although the possible sequences of amino acids are nearly endless, the variations in tertiary structure is rather limited.

Proteins with homologous sequences are considered to share a similar tertiary structure, too. An identity of only 30% can be sufficient for structural similarity so that Molecular Replacement works.

But (there is always a “but”) sometimes, even 100% sequence similarity is not enough to find a solution by molecular replacement.
Molecular Replacement - Flow Chart

The steps of Molecular Replacement are:

1. Find a similar structure - e.g. by sequence comparison against all known structures in the Protein Database\textsuperscript{a}. 30\% sequence similarity is considered the minimum.
2. Correctly place this search prototype in the unit cell
3. Combine the phases $\phi(hkl)$ calculated from the placed prototype with the structure factor amplitudes $|F(hkl)|$ derived from the measured intensities $I(hkl)$.

Step (2) is the tricky step.

\textsuperscript{a}The PDB can be used free of charge and can be accessed e.g. at www.pdb.org or www.pdbe.org
Why does this work?

When the search prototype is sufficiently similar and if it is correctly placed within the unit cell, the calculated phases are close enough to the real phases to get an interpretable map.
Molecular Replacement Programs

**Phaser** probably *the* program of choice for molecular replacement.

1. Easy to use
2. tolerant of *clashes*
3. offers choice of spacegroups in ambiguous cases
4. fast (approx. 30 minutes for an average structure)

**MrBump** model search and preparation based on sequence

There are also the programs **Amore, Molrep, EPMR**, but I have no or little experience with these programs.
MR: Model Preparation

Every difference between the search prototype and the molecule inside the crystal reduces the chance for a good molecular replacement solution.

There is some advice as to how to prepare the search model before carrying out the search:

- Remove solvent and ligands
- Remove flexible parts, mostly loop regions.
- Split the molecule into domains and search one after the other
- Try several copies per asymmetric unit (oligomeric proteins)
MR: Example Model Preparation

Molecules like this one with several domains tend to be flexible and might crystallise in slightly different ways.

- Separate into three domains: blue, green + half of red linker helix, yellow + half of red linker helix
- remove loop region in yellow domain
- remove disconnected (disordered) helix in blue domain

PDB-ID: 1OFC
MR: Number of Molecules

Macromolecules often crystallise as oligomers. For molecular replacement this can become an obstacle since the search program must now how many copies of the molecule it should be looking for.

← Realistic packing for a small protein in a large unit cell (spacegroup $I4_122$)

← Molecule crystallised as heptamer (7-mer) ⇒ 112 molecules in unit cell.

← It could easily be one molecule more or less without sterical clashes.
Estimating Number of Molecules: the Solvent Content

The previous unit cell shows large “white” areas. These are filled with solvent molecules (water, salt), which are disordered and therefore cannot be seen in the crystal structure.

--- relative frequency ---
4.17 -
3.85 --
3.33 -----
2.94 ------------
2.78 ----------------
2.50 ---------------------------------------
2.38 --------------------------------------
2.27 --------------------------------------
2.08 --------------------------------------
1.92 ----------------
1.79 ---
1.67 -
1.61 * (COMPOSITION*1)
1.56 -

Macromolecules crystallise typically with 30-70 % solvent content, centred around 50%. With these statistics the example on the left says that it is very unlikely that there are 13 molecules in the asymmetric unit (but not impossible).

Phaser log-file with 13 copies

According to statistics there are probably 9 copies of the molecule in the asymmetric unit (in this case there were 7).
Model Bias - the Main Risk

Molecular replacement uses “foreign” phases in order to calculate the electron density map. One hopes the these phases are close enough to the real phases that the electron density is correct or at least close enough to allow for improvements.

Thought Experiment (Kevin Cowtan, http://www.ysbl.york.ac.uk/~cowtan/fourier/fourier.html) :

The phase \( \phi \) of the duck determines the shape.
A Simple Test for Model Bias

The risk of model bias is particularly high at medium or low resolution.

To check whether an molecular replacement solution is correct or just a random solution, do the following

1. Before running the MR program, remove part of the search model, e.g. half an $\alpha$-helix.
2. Carry out the MR
3. Look at the resulting map: If there is electron density for the removed part, the solution is certainly correct. If not: You are most likely (but not certainly) looking at a false solution
Example Test

Some residues of the helix in this search model were removed before MR.
The resulting map does not show any signs of density for these residues.
Therefore, this is most likely a false solution.