

Biochemistry 503

Michael Wiener (mwiener@virginia.edu, 3-2731, Snyder 360)

Two Lectures in X-ray Crystallography

Outline

1. Justification & introductory remarks
2. Experimental setup
3. Protein crystals – how to obtain them
4. Resolution & Bragg's Law
5. Crystals – what are they?
6. The reciprocal lattice & the Ewald sphere
7. X-ray data to electron density maps
8. The 'fundamental equation' of crystallography
9. Fourier transforms (pictures & examples)
10. Depiction of phases via the Argand diagram
11. Crystallographic phases and how to get them
12. Design of phasing experiments
13. Graphical image of phase determination
14. Crystallographic B-factors
15. Crystallographic Refinement
16. Evaluating a structure
17. (Graphical examples/demonstration)

Goal of lectures (what you should know/extract/learn)

1. Qualitative - description of how to get from x-rays to a structure (i.e., everything not covered by 2 or 3)
2. Semi-quantitative - reciprocal lattice, phasing, evaluation of a structure
3. Quantitative - Bragg's Law, design a phasing experiment, Argand diagram, B-factors

What is X-ray Crystallography ?

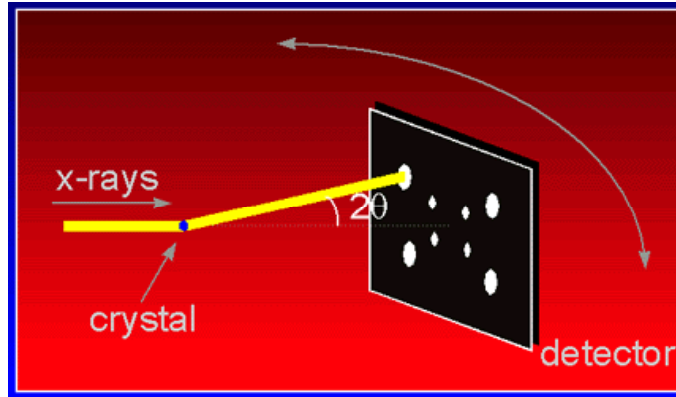
X-ray crystallography is an experimental technique that exploits the fact that X-rays are diffracted by crystals. It is not an imaging technique. X-rays have the proper wavelength (in the Ångström range, $\sim 10^{-10}$ m) to be scattered by the electron cloud of an atom of comparable size. Based on the diffraction pattern obtained from X-ray scattering off the periodic assembly of molecules or atoms in the crystal, the electron density can be reconstructed. Additional phase information must be extracted either from the diffraction data or from supplementing diffraction experiments to complete the reconstruction (the phase problem in crystallography). A model is then progressively built into the experimental electron density, refined against the data and the result is a quite accurate molecular structure.

Why Crystallography ?

The knowledge of accurate molecular structures is a prerequisite for rational drug design and for structure based functional studies to aid the development of effective therapeutic agents and drugs. Crystallography can reliably provide the answer to many structure related questions, from global folds to atomic details of bonding. In contrast to NMR, which is an indirect spectroscopic method, no size limitation exists for the molecule or complex to be studied. The price for the high accuracy of crystallographic structures is that a good crystal must be found, and that only limited information about the molecule's dynamic behavior is available from one single diffraction experiment.

Outline of the experiment

In a macromolecular X-ray diffraction experiment a small protein crystal is placed into an intense X-ray beam and the diffracted X-rays are collected with an area detector (it is advantageous to cool the crystals to low temperatures, primarily to prevent radiation damage). The diffraction pattern consists of reflections of different intensity, and a lot of patterns need to be collected to cover all necessary crystal orientations. After some data processing, we end up with a list of indexed reflections and their intensities.



The diffracted X-rays are scattered by the crystal at a certain angle. The further backwards the x-rays scatter, the higher we say is the resolution of the data set. The extent to which the crystal diffracts determines how fine a detail we can actually distinguish in our final model of the structure. High resolution is thus desirable. Knowing the wavelength and the diffraction angle of a reflection, its resolution d can be easily calculated :

$$d = \frac{1}{2} \left(\frac{\lambda}{\sin \theta} \right)$$

$$\lambda = 2d \sin \theta$$

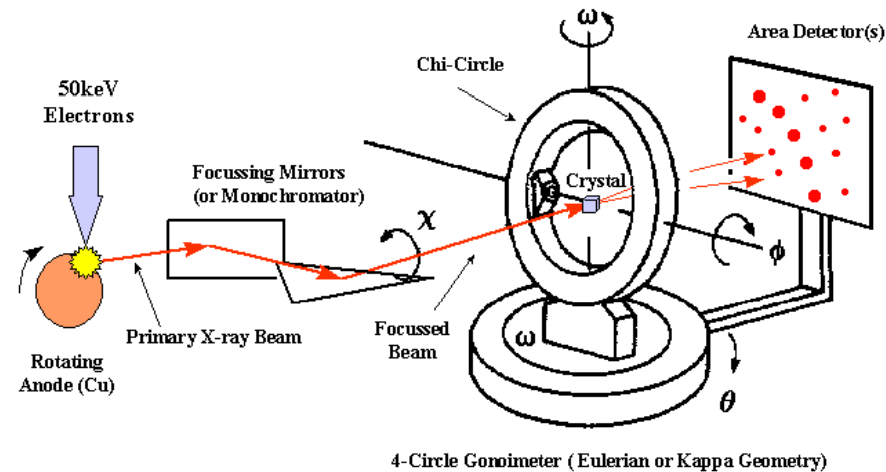
This is just a reformulation of the famous Bragg equation

X-ray Diffraction Equipment

The Experimental Setup

To perform an X-ray diffraction experiment, we need an x-ray source. In most cases a rotating anode generator producing an X-ray beam of a characteristic wavelength is used. Intense, tunable X-ray radiation produced by a Synchrotron provides additional advantages. The primary X-ray beam is monochromated by either crystal monochromators or focusing mirrors. After the beam passes through a helium flushed collimator it passes through the crystal mounted on a pin on a goniometer head. The head is mounted to a goniometer which allows to position the crystal in different orientations

in the beam. The diffracted X-rays are recorded using image plates, Multiwire detectors or CCD cameras.



Flash cooling protein crystals to cryogenic temperatures (~ 100 K) offers many advantages, the most significant of which is the elimination of radiation damage. A part of the X-rays passing through the crystal is scattered in different directions into a detector. The detector delivers an image of the diffraction spots. A large number of these images recorded from different crystal orientations are processed (scaled and merged) into a final list of indexed reflection intensities.

How to grow protein crystals

We all are familiar with crystals from rock collections or small molecules, such as salt or sugar. We usually associate them with properties like hard, durable, and pretty. Unfortunately, only the latter is true for protein crystals.

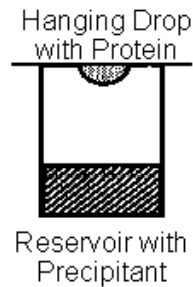
Proteins

Proteins consist of long macromolecule chains made up from 20 different amino acids. The chains can be several hundred residues long and fold into a 3-dimensional structure. It is therefore quite understandable that protein molecules have irregular shapes and are not ideally suited to be stacked into a periodic lattice, i.e., a crystal. Protein crystals are thus very fragile, soft (think of a cube of jelly instead of a brick) and sensitive to all kind of environmental variations. Protein crystals contain on average 50% solvent, mostly in large channels between the stacked molecules on the crystal. The interactions holding the molecules together are usually weak, hydrogen bonds, salt bridges, and hydrophobic interactions, compared to strong covalent or ionic interactions in mineral crystals. This

explains the fragility of the crystals, but allows for the possibility of soaking metal solutions (important for phasing) or even large enzyme substrates or inhibitors, into the crystals.

The Experimental Setup

In order to obtain a crystal, the protein molecules must assemble into a periodic lattice. One starts with a solution of the protein with a fairly high concentration (2-50 mg/ml) and adds reagents that reduce the solubility close to spontaneous precipitation. By slow further concentration, and under conditions suitable for the formation of a few nucleation sites, small crystals **may** start to grow. Often very many conditions have to be tried to succeed. This is usually done by initial screening, followed by a systematic optimization of conditions. Crystals should be at least a tenth of a mm in each direction to be useful for 'in-house' diffraction experiments; smaller crystals can be used at the synchrotron.



Right : The hanging drop technique.

The most common setup to grow protein crystals is by the **hanging drop** technique : A few microliters of protein solution are mixed with an about equal amount of reservoir solution containing the precipitants. A drop of this mixture is put on a glass slide which covers the reservoir. As the protein/precipitant mixture in the drop is less concentrated than the reservoir solution (remember: we mixed the protein solution with the reservoir solution about 1:1), water evaporates from the drop into the reservoir. As a result the concentration of both protein and precipitant in the drop slowly increases, and crystals may form. There is a variety of other techniques available such as sitting drops, dialysis buttons, and gel and microbatch techniques. Robots are useful for automatic screening and optimization of crystallization conditions. The main advantage is the small sample size a crystallization robot can handle reproducibly, but it needs some effort to set it up.

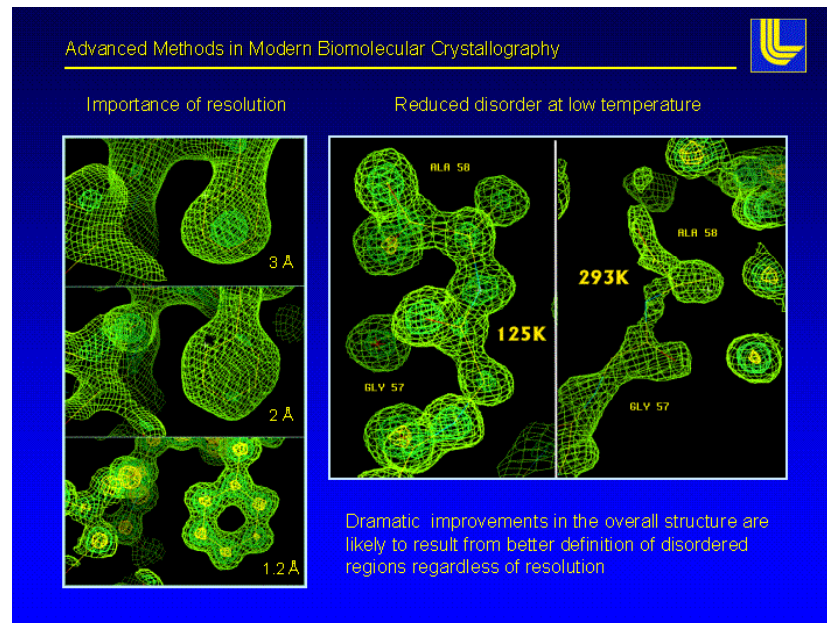
Helpful techniques to screen proteins

The chances for success in crystallization experiments depend strongly on the conformational purity of a protein. A single band on a denaturing SDS gel is a good sign, but other methods such as dynamic or static light scattering, circular dichroism, gel filtration, analytical ultracentrifugation, and native-gel electrophoresis can be useful to

characterize aggregation and dispersity or the folding of your protein sample. A functional assay can also certainly be of utility!

What does high resolution mean?

The following figure shows what a certain resolution, given in Ångström (Å) means for the user of structural models derived from X-ray data. One always has to remember that the cute model you see was built into an experimental electron density. The model may look as good at 3 Å as it does at 1.5 Å, but is it a correct and unique description of reality?



Above: pictures of the electron density at different data set resolution of the same region of a molecule. There is no question that a model of phenylalanine (the 6-ring structure) can be correctly placed into the 1.1 Å data. This still can be done with confidence in the 2 Å case, but at 3 Å we already observe a deviation of the centroid of the ring from the correct model. The left panels shows the same nominal resolution structure at room temperature (293K) and nitrogen cooled (125K). The difference is a striking example for improvements that can be achieved using cryo-techniques. Most protein crystals diffract between 1.8 and 3 Å, a few to very high resolution (the term high resolution is used loosely in macromolecular crystallography, we apply it to data of 1.2 Å or better resolution). The most efficient way to increase resolution (short of trying to grow better crystals) is to cryo-cool the crystals to near liquid nitrogen temperature.

Bragg's Law and Diffraction: How waves reveal the atomic structure of crystals

What is Bragg's Law and Why is it Important?

Bragg's Law refers to the simple equation:

$$n\lambda = 2d \sin\theta \quad (1)$$

derived by the English physicists Sir W.H. Bragg and his son Sir W.L. Bragg in 1913 to explain why the cleavage faces of crystals appear to reflect X-ray beams at certain angles of incidence (theta, θ). The variable d is the distance between atomic layers in a crystal, and the variable lambda λ is the **wavelength** of the incident X-ray beam, n is an integer.

This observation was an example of X-ray **wave interference** (Roentgenstrahlinterferenzen) commonly known as X-ray diffraction (XRD), and was direct evidence for the periodic atomic structure of crystals postulated for several centuries. The Braggs were awarded the Nobel Prize in physics in 1915 for their work in determining crystal structures beginning with NaCl, ZnS and diamond. Although Bragg's law was used to explain the interference pattern of X-rays scattered by crystals, diffraction has been developed to study the structure of all states of matter with any beam, e.g., ions, electrons, neutrons, and protons, with a wavelength similar to the distance between the atomic or molecular structures of interest.

Deriving Bragg's Law

Bragg's Law can easily be derived by considering the conditions necessary to make the phases of the beams coincide when the incident angle equals and reflecting angle. The rays of the incident beam are always in phase and parallel up to the point at which the top beam strikes the top layer at atom z (Fig. 1). The second beam continues to the next layer where it is scattered by atom B. The second beam must travel the extra distance $AB + BC$ if the two beams are to continue traveling adjacent and parallel. This extra distance must be an integral (n) multiple of the wavelength (λ) for the phases of the two beams to be the same:

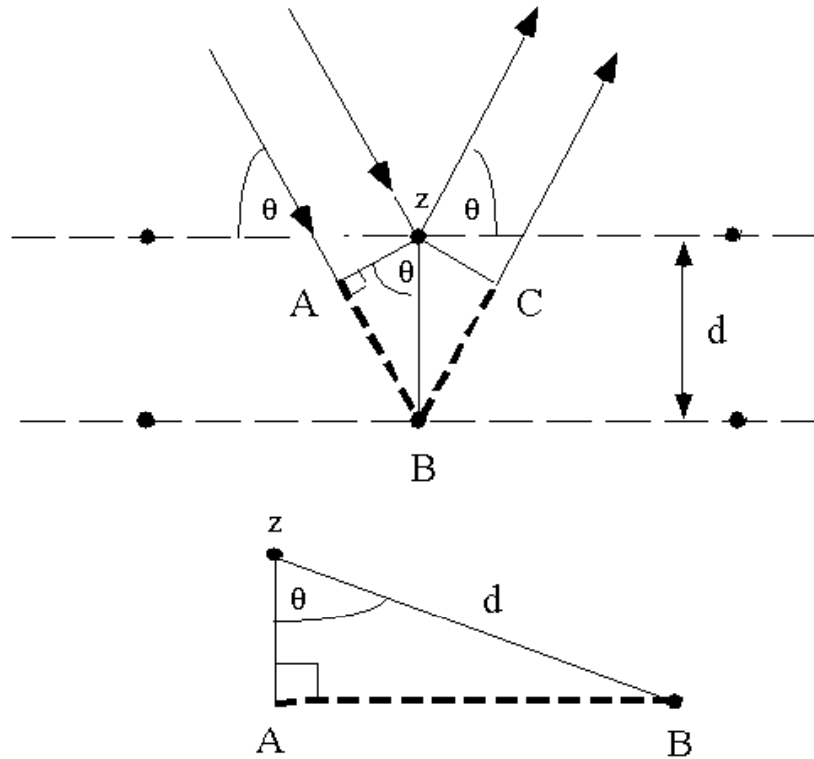


Fig. 1 Deriving Bragg's Law using the reflection geometry and applying trigonometry. The lower beam must travel the extra distance ($AB + BC$) to continue traveling parallel and adjacent to the top beam.

Recognizing d as the hypotenuse of the right triangle Abz , we can use trigonometry to relate d and θ to the distance ($AB + BC$). The distance AB is opposite θ so,

$$AB = d \sin\theta \quad (3).$$

Because $AB = BC$ eq. (2) becomes,

$$n\lambda = 2AB \quad (4)$$

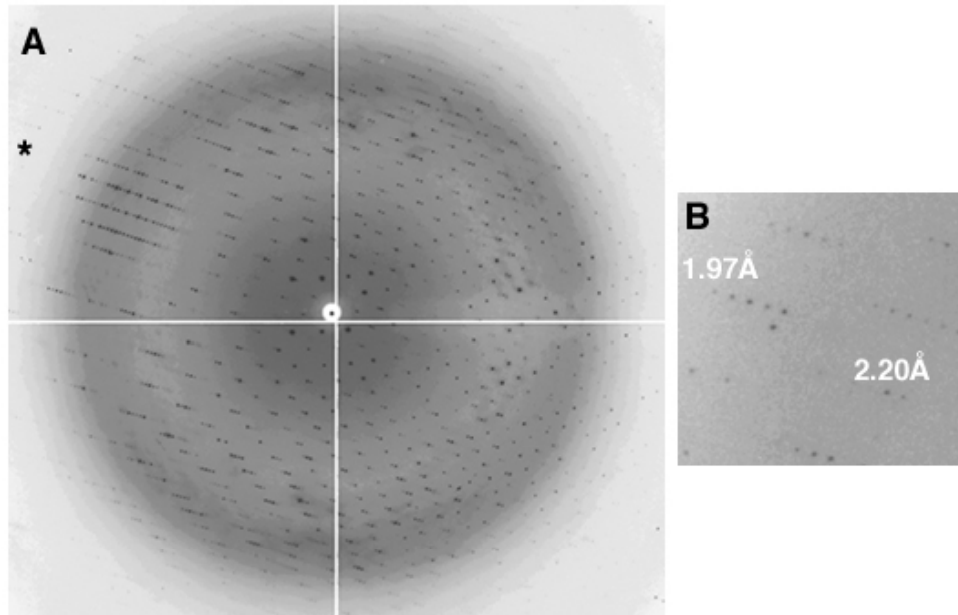
Substituting eq. (3) in eq. (4) we have,

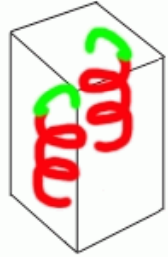
$$n\lambda = 2 d \sin\theta, \quad (1)$$

and Bragg's Law has been derived. The location of the surface does not change the derivation of Bragg's Law.

Experimental Diffraction Patterns

The following figure shows experimental x-ray diffraction patterns of a protein crystal (in this case, a membrane protein crystal) obtained with synchrotron radiation.





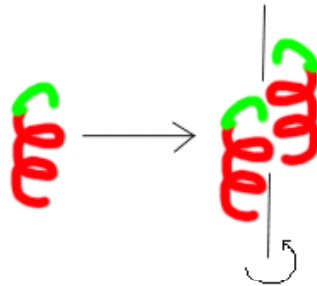
About Crystals, Symmetry and Space Groups

How can proteins be assembled in a periodic lattice ?

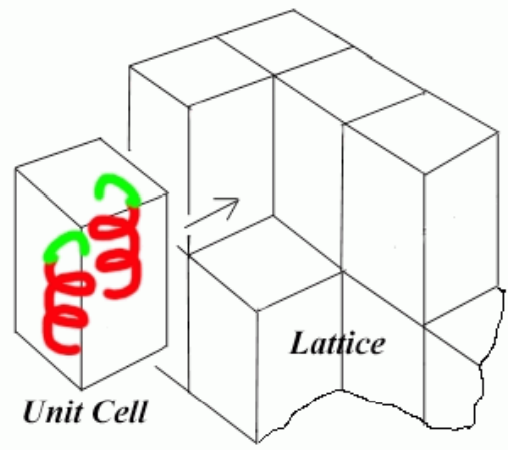
A crystal is a periodic arrangement of a motif in a lattice. The **motif** can be a single atom, a small molecule, a protein or any combination thereof. So here is our protein, RGFP (Red-Green Fluorescent Protein) serving as a structural motif :



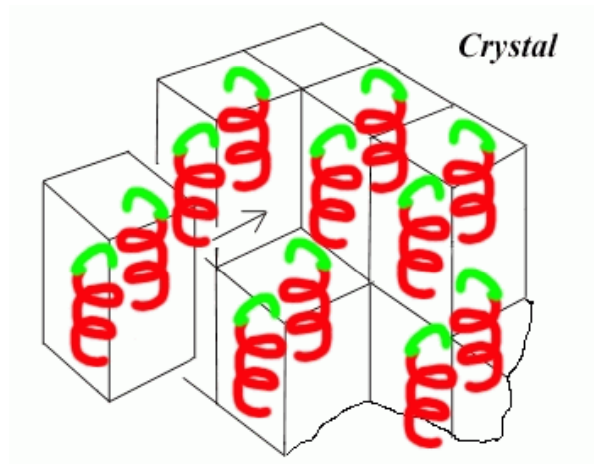
If we just repeat this motif in three dimensions, we have realized the most simple way to form a crystal. Very often the motif, also referred to as the 'asymmetric unit', is subjected to a number of symmetry operations yielding differently oriented copies. Let's just use a 2-fold axis for now :



If there are no additional symmetry operations, we have already created the contents of the unit cell. The crystal is built from the unit cells arranged into a three-dimensional lattice :



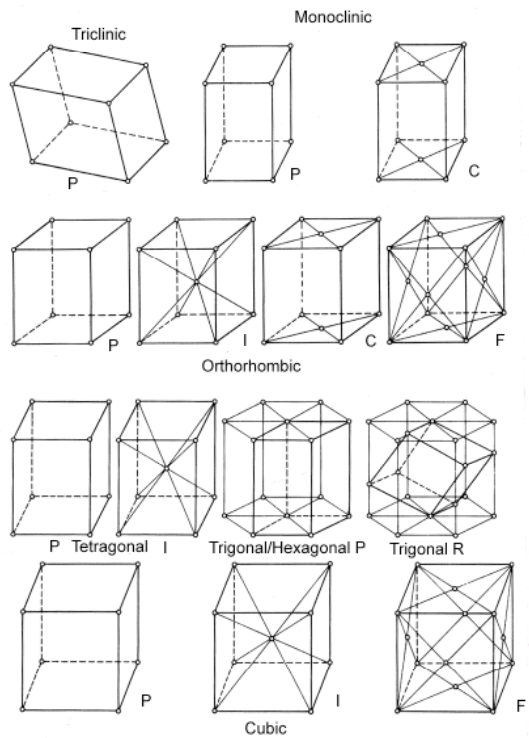
So here is our final crystal of RGFP.

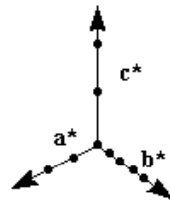


The question arises, how many ways are there to build crystals by combination of symmetry operations and lattice translations ? Infinite ? A few 1000 ? Time to look at this in more detail.

Crystal Systems and Bravais Lattices

Let us now consider in which way we can translate our cell contents in 3 dimensions to obtain a crystal. The translations which are allowed create 14 Bravais lattices which belong to 7 crystal systems. Each system has a primitive cell, and some allow face or body centering, as well as the rhombohedral centering in the trigonal system (the rhombohedral lattice can be derived from a cube by pulling along its space diagonal).

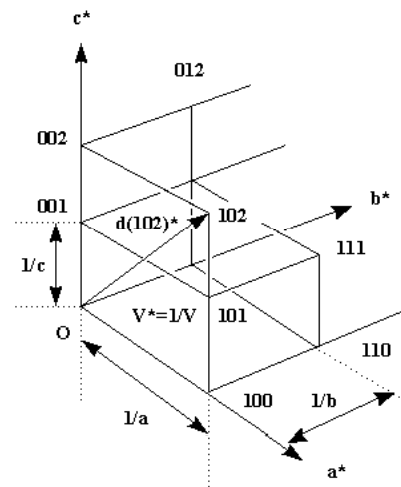




The Reciprocal Lattice

Introduction

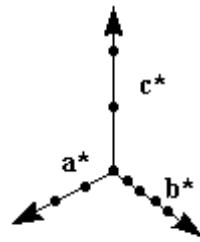
In the introduction to crystal symmetry I have shown that a crystal consists of a periodic arrangement of the unit cell (filled with the motif and its symmetry generated equivalents), into a lattice. In the same fashion we can define the **reciprocal lattice**, whose lattice dimensions are reciprocal to the original cell (and correspond to the reflection *positions*) and whose 'size' (the *intensity* of the reflection) corresponds to the *contents* of the unit cell. The following picture will make this clear.



Each of the lattice points corresponds to the diffraction from a periodic set of specific crystal lattice planes defined by the index triple hkl . The dimensions of the reciprocal lattice are reciprocally related to the real lattice. In the case of the orthorhombic system I have drawn, the relations are simple: $c^* = 1/c$ etc., but in a generic oblique system the relation is more complicated. The length of a reciprocal lattice vector $d(hkl)^*$ (from origin to reciprocal lattice point h,k,l) again corresponds to the reciprocal distance $d(hkl)$ of the crystal lattice planes with this index. In our simple case, for 001 this is just the cell dimension c for $d(001)$ or $1/2 c$ for 002 etc. ($d(001)^* = 1/c$, thus $d=c$).

Resolution revisited

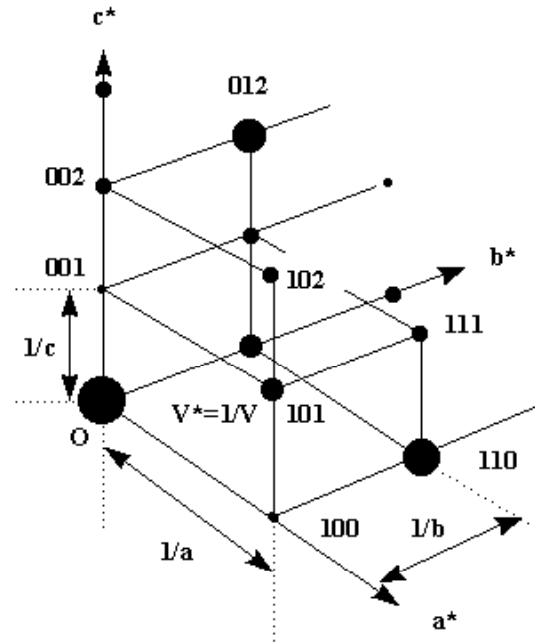
The vector $d(hkl)$ also determines the location of the diffraction spot in the diffraction image. The diffraction *angle* at which we observe the reflection is given by Bragg's formula $\sin(\theta) = \lambda/2d$. The higher the index of a reciprocal lattice point, the larger the diffraction angle will be. We have already seen that the larger the diffraction angle, the higher the *resolution*, i.e. the finer the detail we can observe in the reconstruction of the crystal structure. This can be easily understood now : we need to 'slice' the crystal fine enough (i.e, small $d(hkl)$ = high indices hkl) to have enough information contained in our diffraction pattern to reconstruct details. In the next chapter we will use the reciprocal lattice and the Ewald construction to visualize some important concepts in data collection.



The Ewald Construction

Intensity of diffraction spots

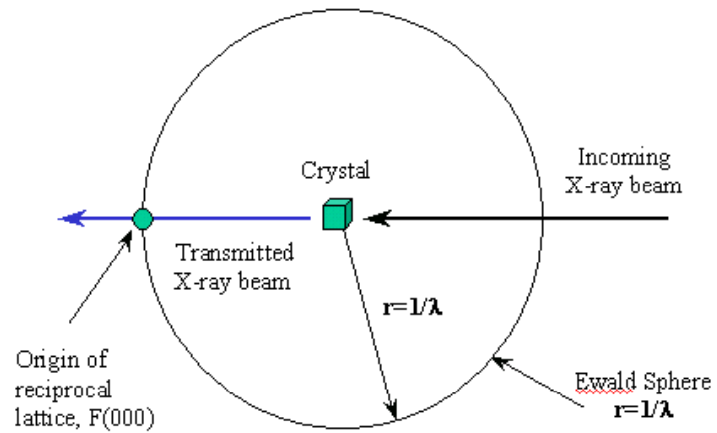
The observed intensity I of the diffraction spots can be thought of as corresponding to the 'size' of the reciprocal lattice point ($I(hkl)$ is proportional to $|F(hkl)|^2$). Clearly, either depends on the contents of the unit cell, and we already suspect that the space group symmetry will thus have some implications on the diffraction pattern symmetry. Before we investigate further, it may be useful to understand how the diffraction pattern can be derived from the reciprocal lattice (RL). Let us look at a RL with spots in it :



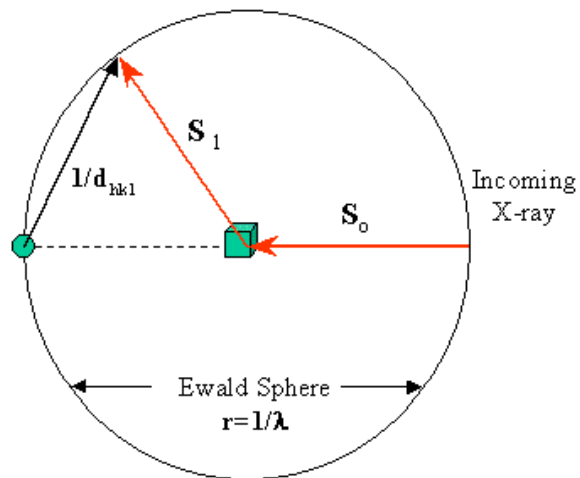
Here we see different magnitudes for the lattice points. The largest spot is at the origin corresponding to $F(000)$, which we know already is the sum of all electrons in the unit cell. The reflection itself is at zero diffraction angle, i.e., in the primary beam path and not observable. Now, where do we expect all the other diffraction spots to appear?

Ewald construction

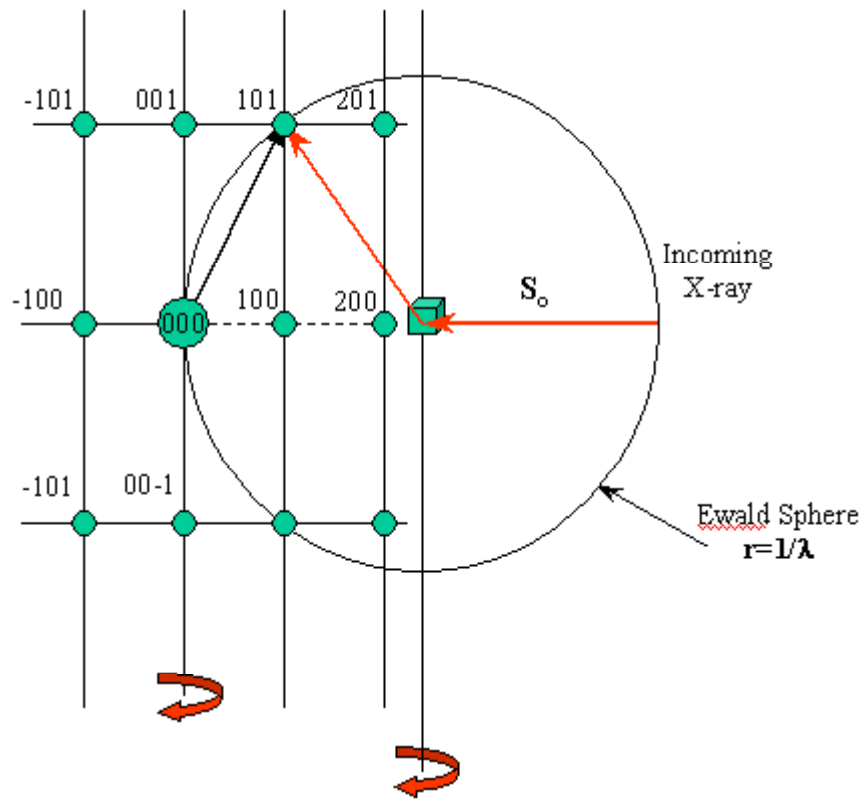
A most useful means to understand the occurrence of diffraction spots is the Ewald construction. Let's begin slowly: We draw a sphere of radius $1/\lambda$, in the center of which we imagine the real crystal. The origin of the reciprocal lattice (RL, see above) lies in the transmitted beam, at the edge of the Ewald sphere.



We know already that diffraction maxima (reflections, diffraction spots) occur only when the 3 Laue equations, or equivalent, the Bragg equation in vector form, are satisfied. This condition occurs whenever a reciprocal lattice point lies exactly on the Ewald sphere.



As you may have assumed already, the chance for this to occur is modest, and we need to rotate the crystal in order to move more reciprocal lattice points through the Ewald sphere. In the following, I have drawn a reciprocal lattice in the origin, and we rotate it along the vertical axis of the drawing. We actually accomplish this by rotating the crystal along the same axis.



Just imagine turning the RL through the Ewald sphere : in the beginning, only (101) and (10-1) give rise to a reflection. After you turned the RL a bit (which actually means turning the crystal around the same axis), the reciprocal lattice point 201 will enter the sphere and create a diffraction spot.

How do we go from x-ray
diffraction data to electron
density maps?

Outline

- Goal of the experiment is to determine $\rho(x,y,z)$, the electron density for all x, y, z in the unit cell.
 - What we measure in the experiment, $|F_{hkl}|$
 - What we still need, ϕ_{hkl} (the phase problem)
- Methods for solving the phase problem
 - Molecular Replacement (MR)
 - Multiple/Single Isomorphous replacement (MIR/SIR)
 - Multiple/Single wavelength Anomalous Diffraction (MAD/SAD)

Goal of the Experiment

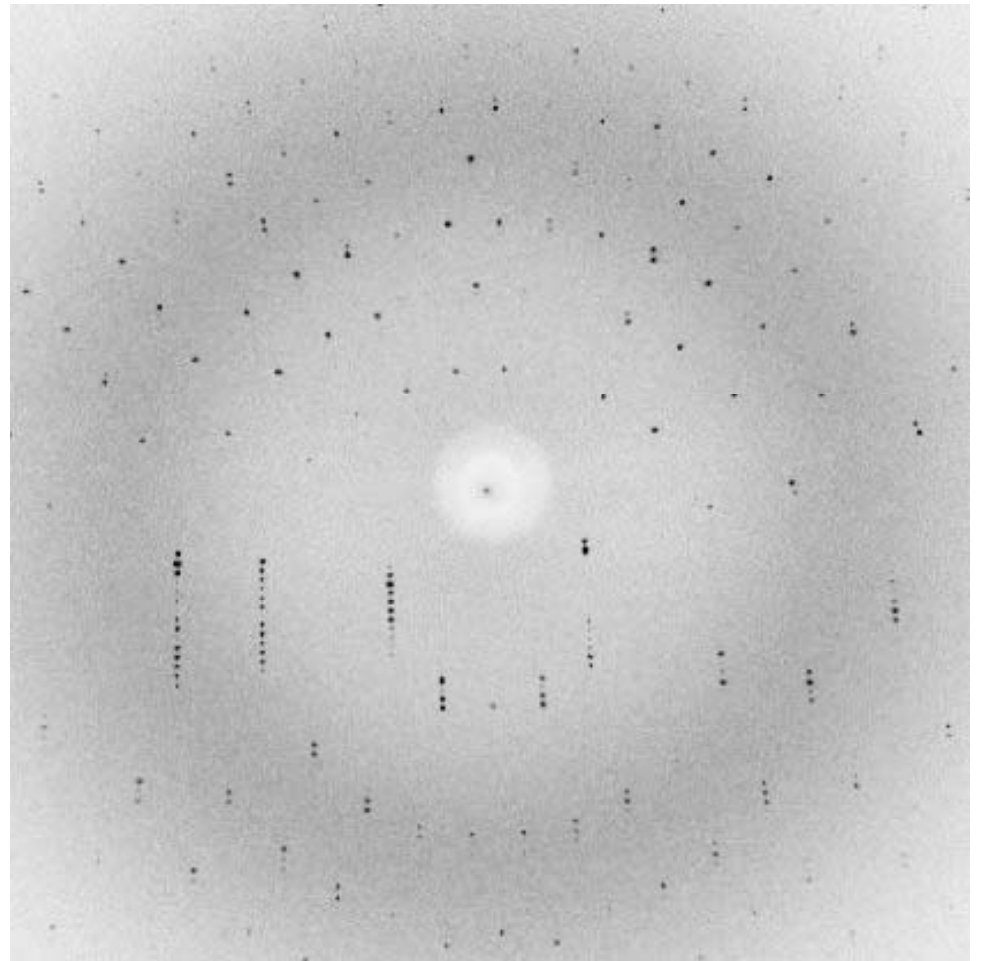
$$\rho(xyz) = \frac{1}{V} \sum_h \sum_k \sum_l |F_{hkl}| \exp[-2\pi \cdot i(hx + ky + lz) + i\varphi_{hkl}]$$

- Need to solve the equation above for all x , y , z in the unit cell
- What do we know?
 - We record the position (the triple index hkl) and intensity, I_{hkl} , of each reflection (spots on the detector)

Diffraction Data

- The measured intensities are proportional to the coefficients of the electron density equation

$$I_{hkl} \propto |F_{hkl}|^2$$



From Diffraction Data to Electron Density

- From the structure factor equation we can see that if we know the contents of the unit cell, we can calculate F_{hkl}
- We are dealing with the inverse problem.
- We have information about F_{hkl} but need to know the contents of the crystal.

From F_{hkl} to $\rho(x,y,z)$

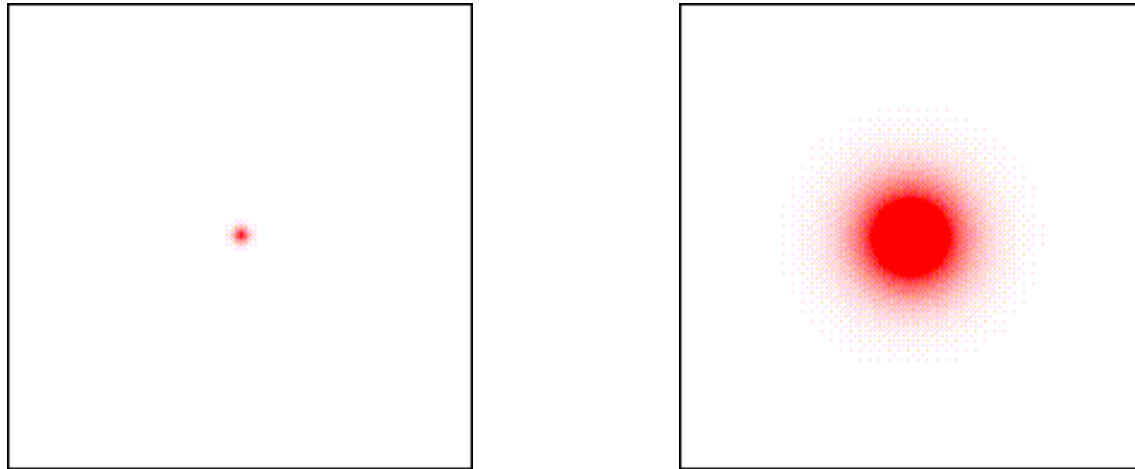
- The structure factor equation is periodic, and is represented a Fourier series.
- Taking the Fourier Transform (FT) of the equation for F_{hkl} we get the necessary equation

$$\rho_{(x,y,z)} = \frac{1}{V} \sum_h \sum_k \sum_l F_{(h,k,l)} \exp[-2\pi \cdot i(hx + ky + lz)]$$

- which describes the electron density in the crystal.
- So the FT of the diffraction data gives us a representation of the contents of the crystal.
 - (The FT of the contents of the crystal gives us the diffraction pattern.)

Fourier Transforms, examples

- An atom, and its Fourier Transform:



- Note the both functions have circular symmetry. The atom is a sharp feature, whereas its transform is a broad smooth function. This illustrates the reciprocal relationship between a function and its Fourier transform.

Fourier Transforms, examples

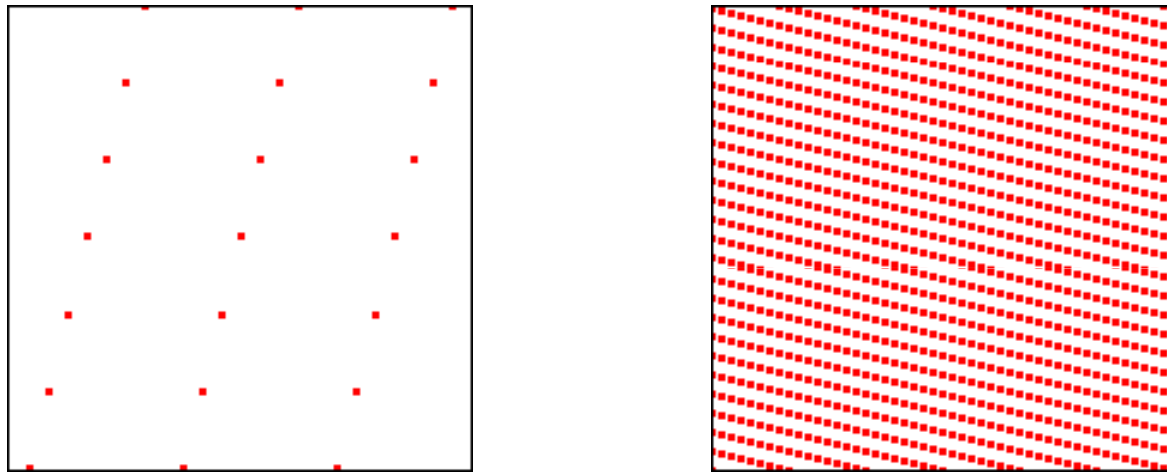
- A molecule and its FT



- The molecule consists of seven atoms. Its transform shows some detail, but the overall shape is still that of the atomic transform. We can consider the molecule as the convolution of the *point atom structure* and the *atomic shape*. Thus its transform is the product of the *point atom transform* and the *atomic transform*.

Fourier Transforms, examples

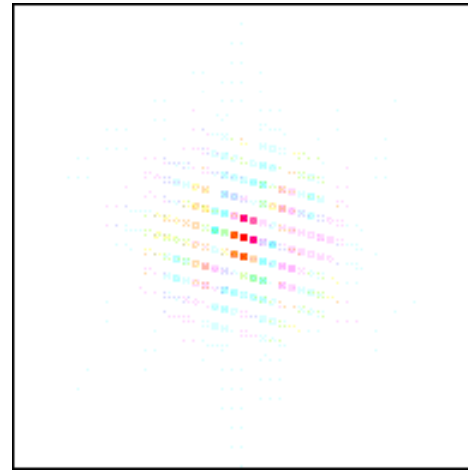
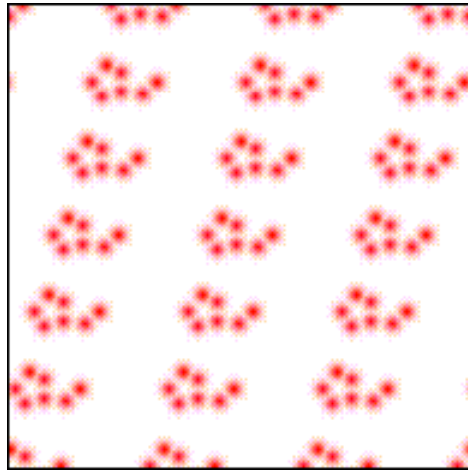
- An lattice, and its Fourier Transform:



- The grid points (delta functions) are exaggerated for clarity. Note that the Fourier transform of a grid is a grid with reciprocal *directions* and *spacings*. This is the origin of the reciprocal lattice.

Fourier Transforms, examples

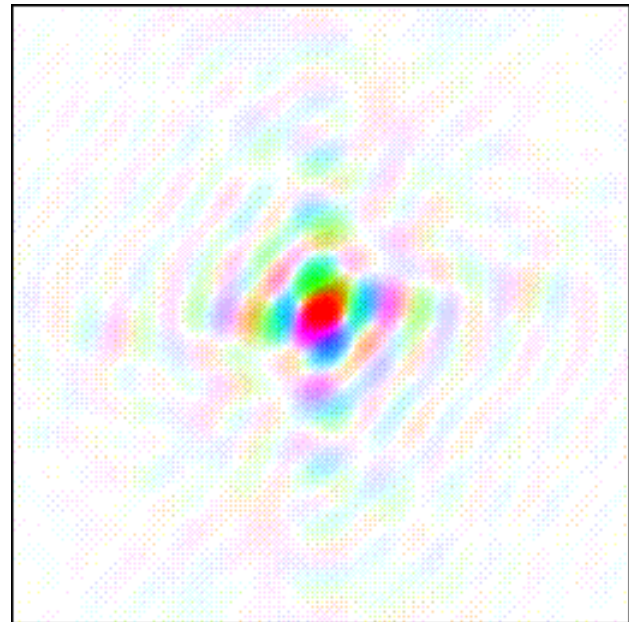
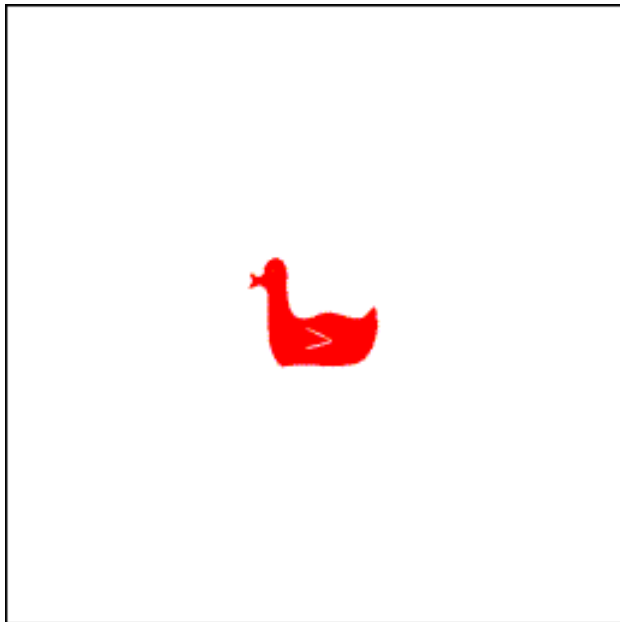
- A crystal, and its Fourier Transform:



- Finally, we build up a crystal by convoluting the *molecule* with the *grid*. The result is a crystal structure. The Fourier transform of the crystal is thus the product of the *molecular transform* and the *reciprocal lattice*. This is the *diffraction pattern*.

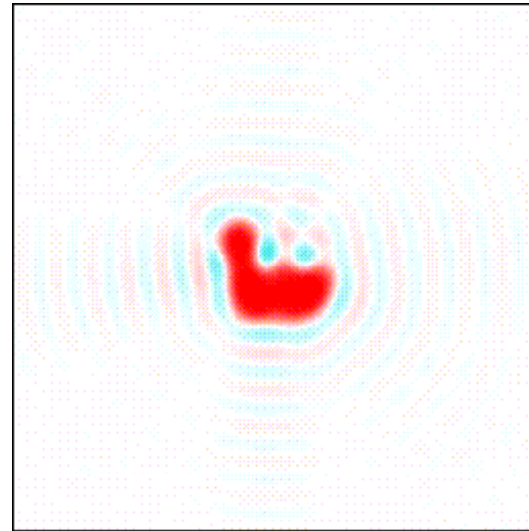
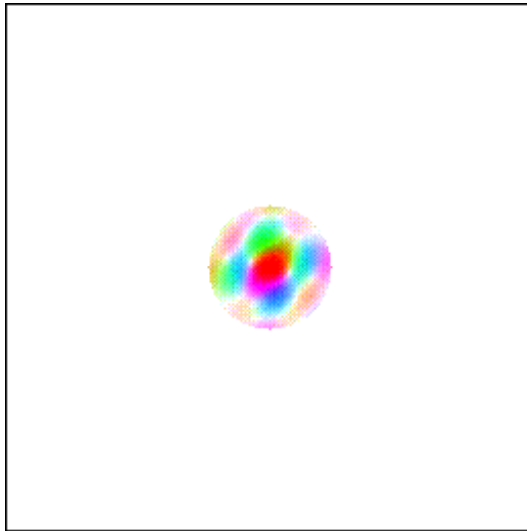
Data Quality Related to Structure

- A duck and its Fourier Transform



Data Quality Related to Structure

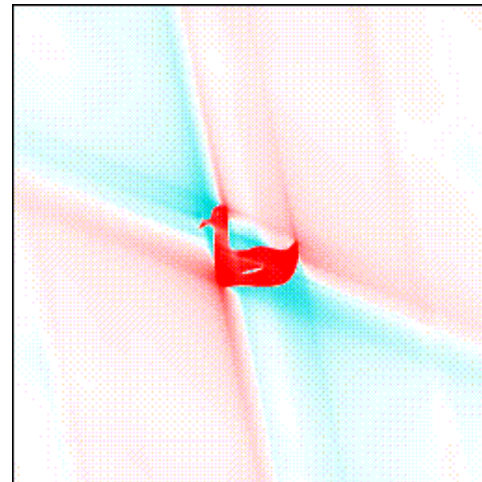
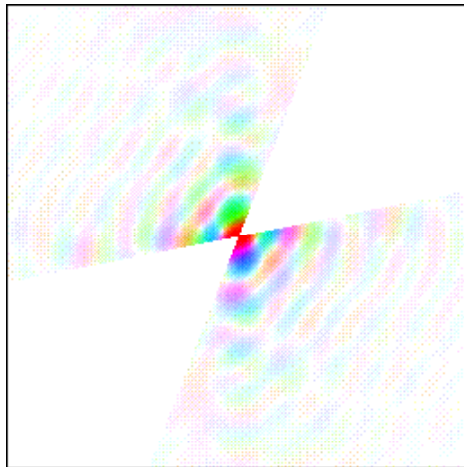
- If we only have the low resolution terms of the diffraction pattern, we only get a low resolution duck:



- **Crystallographic Interpretation:**
- There is considerable loss of detail. At low resolution, your atomic model may reflect more what you expect to see than what is actually there.

Data Quality Related to Structure

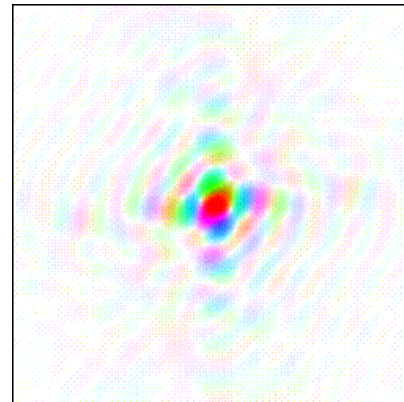
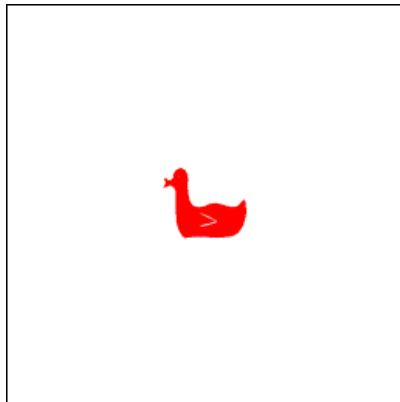
- If a segment of data is missing, features perpendicular to that segment will be blurred.



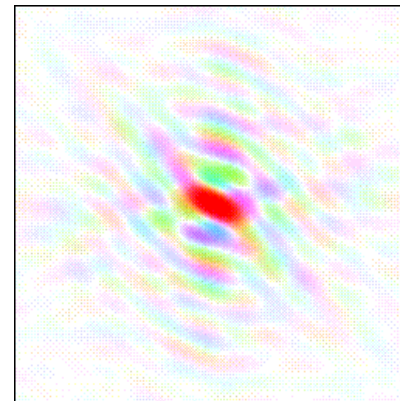
- **Crystallographic Interpretation:**
- Helices parallel to the missing data axis will become cylinders. Beta sheets parallel may merge into a flat blob. Beta sheets perpendicular to the missing data may be very weak. You could get into a lot of trouble with anisotropic temperature factors in this case.

Animal Magic

- Here is our old friend; the Fourier Duck, and his Fourier transform:

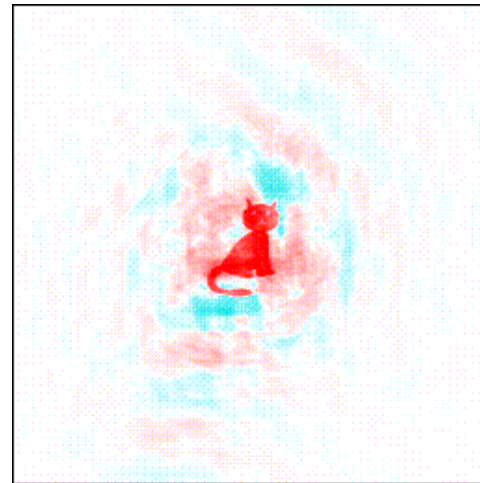
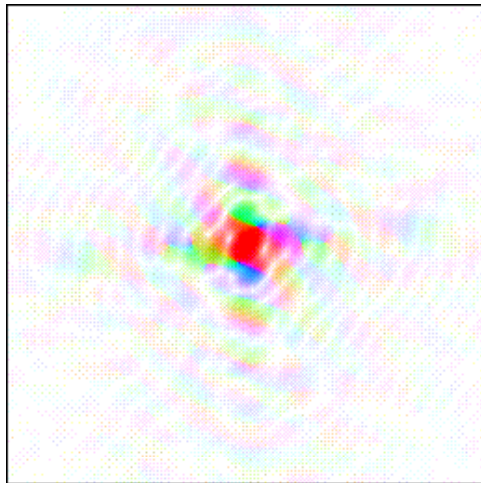


- And here is a new friend; the Fourier Cat and *his* Fourier transform:



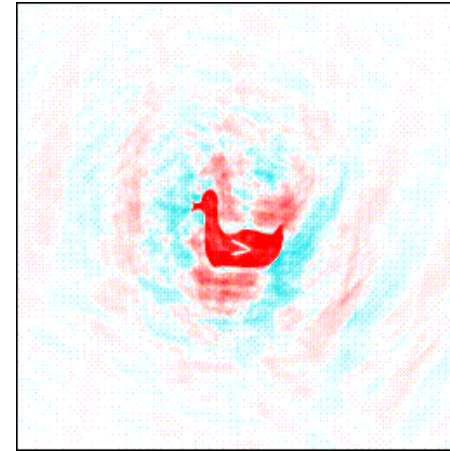
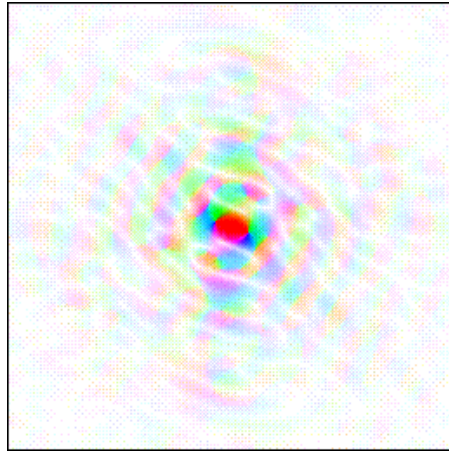
Animal Magic

- Now we will mix them up. Let us combine the the magnitudes from the Duck transform with the phases from the Cat transform. (You can see the brightness from the duck and the colours from the cat). If we then transform the mixture, we get the following:



- We can do the same thing the other way round. Using the magnitudes from the Cat transform and the phases from the Duck transform, we get:

Animal Magic



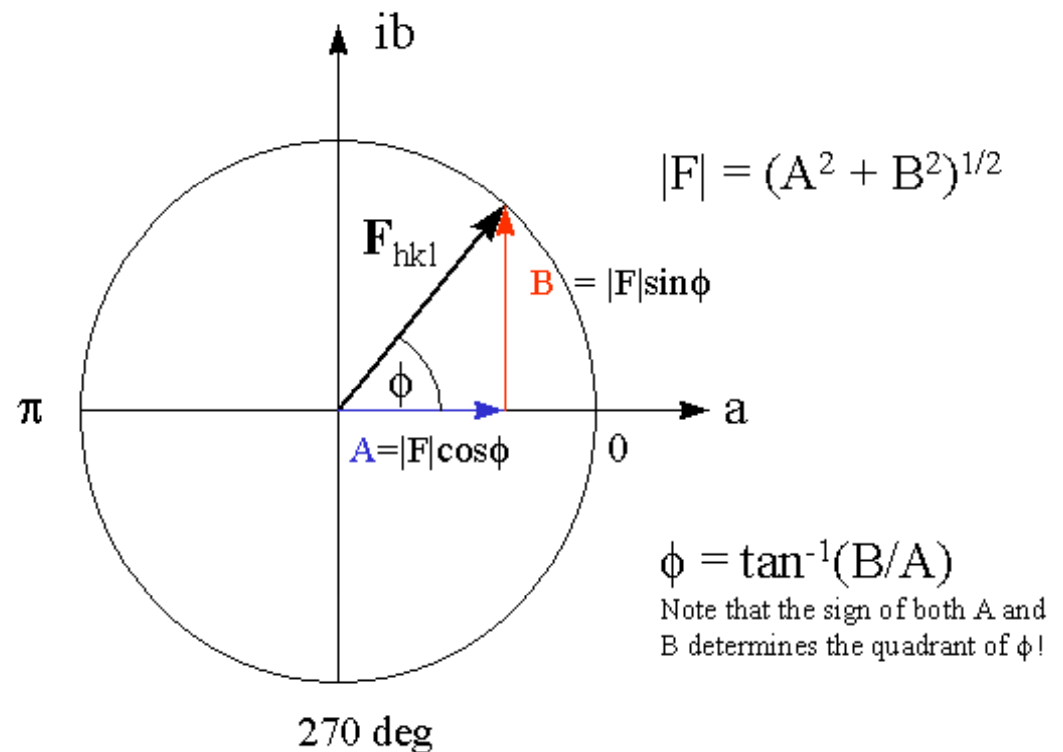
- In each case, the image which contributed the phases is still visible, whereas the image which contributed the magnitudes has gone

Crystallographic Interpretation:

- In X-ray diffraction experiments, we collect only the diffraction magnitudes, and not the phases.
- Unfortunately the phases contain the bulk of the structural information. That is why crystallography is difficult.
- This is also why incorrect phases can cause big problems (e.g. in molecular replacement)

The Phase Problem

- F_{hkl} is complex and can be represented with an Argand diagram.
- $F_{hkl} = A + iB$
- We measured $|F_{hkl}|$ in the experiment but we still need ϕ_{hkl} .



Solving the Phase Problem

- Molecular Replacement (MR)
 - source of initial phases – structure of similar molecule (model)
 - The model is repositioned (replaced) to obtain the best agreement with the x-ray data.
 - Phases are calculated from the model (using the structure factor equation).
 - Calculated phases are combined with the experimental data.

Solving the Phase Problem

- Multiple/Single Isomorphous Replacement (MIR/SIR)
 - Source of phases – intensity differences between data from native and derivative (heavy atom containing) crystals
 - Positions of heavy atoms identified from isomorphous difference Patterson maps
 - Similar to MAD/SAD

Solving the Phase Problem

- Multiple/Single wavelength Anomalous Diffraction (MAD/SAD)
 - Source of phases – anomalous intensity differences
 - Bijvoet differences are between symmetry related reflections at a single wavelength.
 - Dispersive differences are between the same reflection at different wavelengths.
 - Positions of anomalous scatterers identified from anomalous difference Patterson maps.

Summary

$$\rho(xyz) = \frac{1}{V} \sum_h \sum_k \sum_l |F_{hkl}| \exp[-2\pi \cdot i(hx + ky + lz) + i\varphi_{hkl}]$$

- Need to solve the electron density equation which reveals the contents of the crystal.
- From the diffraction data we measure the positions and intensities of the reflections.
- Intensities, I_{hkl} , are proportional to the square of the structure factor magnitudes, $|F_{hkl}|^2$.
- F_{hkl} is the vector sum of the scattering factors of all the atoms in the crystal (f_j).

Summary

- We measure the magnitudes of the F_{hkl} 's, but still need the phases, ϕ_{hkl} .
- Several methods commonly used to obtain phases
 - Molecular Replacement
 - MIR/SIR
 - MAD/SAD

Estimation of the signal from an anomalous scatterer or heavy atom:

$$\text{Signal size} = \text{relative change (\%)} \text{ in } F \cong (2N/N_T)^{1/2} (X/Z_{\text{eff}})$$

N = number of anomalous or heavy-atom scatterers

N_T = total number of nonhydrogen atoms

$$\sim 7.8 * N_{\text{res}} \text{ or } 68 * M \text{ (kD)}$$

Z_{eff} = effective normal scattering at zero angle

$$\sim 6.7 \text{ electrons}$$

$X = Z$ (# of electrons) for isomorphous replacement

= f “ for anomalous diffraction (Bijvoet differences)

What is the corresponding fractional change in intensity (i.e., in the measured diffraction)?

What size signal is required? It should be larger than or at least comparable to the experimental noise of the data.

What is the experimental noise of the data? What are features of a dataset (or of the information on a dataset in the table of a paper) that we should examine?

Excerpt from the summary (logfile) of a data-reduction program (HKL2000, W. Minor & Z. Otwinowski):

```
Summary of reflections intensities and R-factors by shells
R linear = SUM ( ABS(I - <I>)) / SUM (I)
R square = SUM ( (I - <I>) ** 2) / SUM (I ** 2)
Chi**2   = SUM ( (I - <I>) ** 2) / (Error ** 2 * N / (N-1) ) )
In all sums single measurements are excluded
```

Shell limit	Lower Angstrom	Upper Angstrom	Average I	Average error	stat. Chi**2	Norm. Chi**2	Linear R-fac	Square R-fac
25.00	4.30	18163.3	320.7	231.5	2.262	0.047	0.054	
4.30	3.42	12887.9	227.9	173.4	1.760	0.046	0.051	
3.42	2.99	4872.7	122.0	106.4	1.345	0.063	0.065	
2.99	2.71	2043.7	85.4	80.8	0.944	0.088	0.088	
2.71	2.52	1085.5	75.8	74.0	0.718	0.128	0.128	
2.52	2.37	667.6	75.2	74.4	0.546	0.183	0.172	
2.37	2.25	507.3	79.0	78.5	0.503	0.242	0.266	
2.25	2.15	402.6	87.0	86.7	0.421	0.290	0.284	
2.15	2.07	301.7	115.1	114.9	0.385	0.368	0.367	
2.07	2.00	238.3	141.0	140.8	0.338	0.412	0.371	
All reflections		4217.3	134.0	116.7	0.959	0.065	0.056	

Resolution shells: An approximately equal number of reflections are in each shell. Note that equal shells of reciprocal space do NOT correspond to equal shells of real-space.

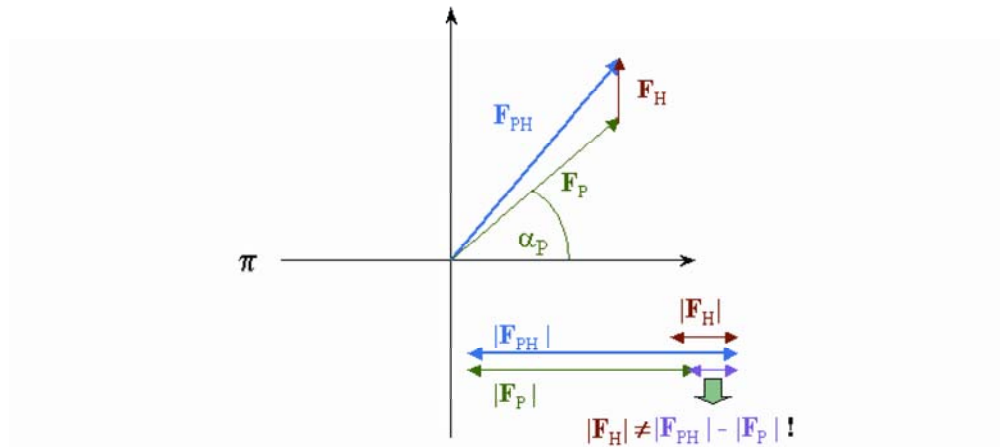
The linear R-factor (often called R_{sym} – note that there are a bunch of R-things used in crystallography and each means something different) is used to estimate the precision with which diffraction data are measured. Because of symmetry, for a given reflection $I(hkl)$ there will be some other $I(h'k'l')$ that are equal to it [or that exact same reflection can be measured again]. Thus, each $I(hkl)$ will belong to some set of reflections that are equal, and deviation of each reflection from the average of the set can be used to measure the ‘noise’ in the data. (This R_{sym} formula should remind you somewhat of the formula for calculating a standard deviation).

Graphical depiction of experimental phase determination

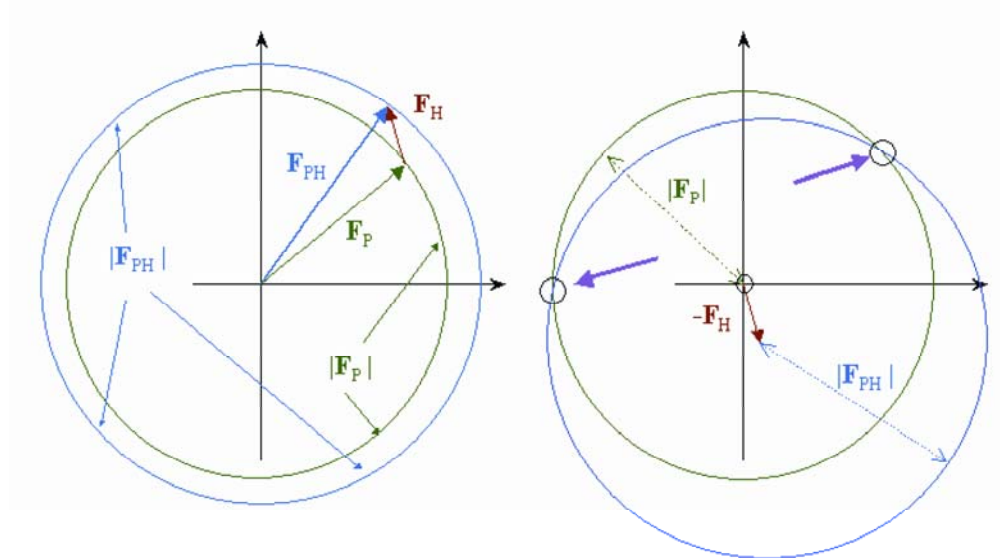
The following discussion is for a heavy-atom derivative used in isomorphous replacement. A similar presentation can be made for anomalous scattering. The fundamental point is that neither a single heavy-atom derivative nor a single-wavelength anomalous diffraction (SAD) dataset is sufficient to determine the experimental phase.

Correlation of native and derivative scattering factors

At this point, you may want to briefly review the introduction to [vector representation of scattering factors](#). The following vector diagram (Harker presentation) illustrates the relationship between native and derivative scattering factors. The objective of a phasing experiment is to derive the unknown phase $\alpha(p)$ of each protein reflection Fp .



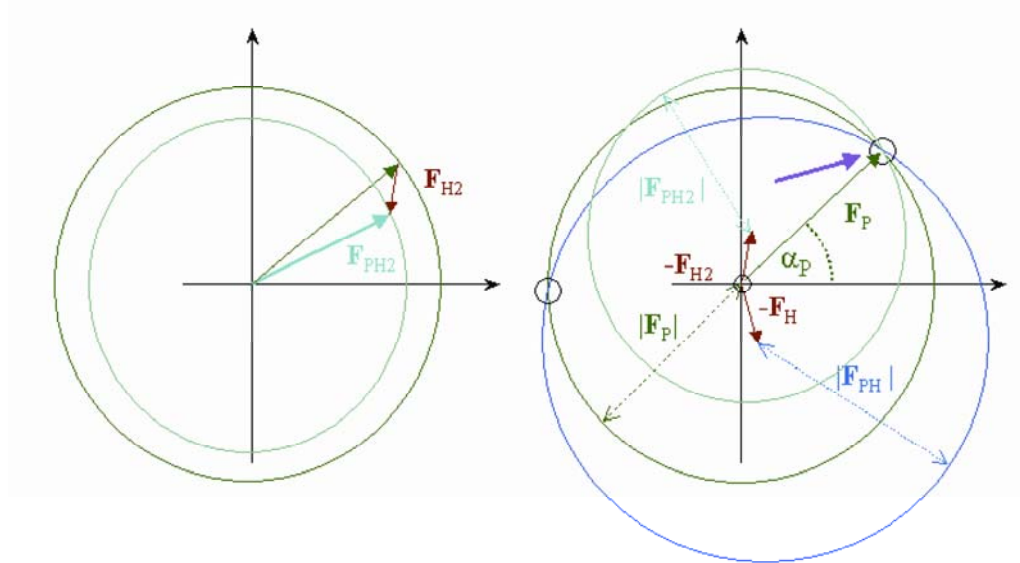
From our the experiment, we know only the magnitudes $|F_{ph}|$ (derivative) and $|F_p|$ (protein) which can be represented in the complex plane as a circle of radius $|F_{ph}|$ and $|F_p|$, respectively. If we **know both the magnitude and the phase of F_h** we can draw both circles offset by vector F_h and obtain 2 solutions for possible phase values for F_p (magenta arrows) :



The phase and magnitude of F_h can be calculated easily if we know the positions of a heavy metal (we will show in the next chapter how to determine these positions). At this point is is clear that the best phase we can obtain from the 2 solutions is the mean in between the 2 possibilities, and the phase error can be quite large. In real cases, F_h is much shorter than my red vector, and the phase error will be quite large, somewhat below 90 degrees. We realize that on average, **SIR (Single Isomorphous Replacement) phases** will be better with larger

contributions from the heavy metal. More about SIR phases later.

In order to eliminate the *phase ambiguity* we can prepare a **second derivative** and repeat the procedure. Provided the heavy atom is not at the same position, we can now obtain a unique solution for $\alpha(p)$, the phase of F_p :



We have now, at least in theory, an exact *solution for the phase angle of F_p* . The theory is based on 2 assumptions : a) ideal isomorphism and b) exact heavy atom positions, neither of which are perfectly met, for practical and experimental reasons in the first case and for theoretical reasons in the second. In our picture it means that the phasing circles may not intersect in exactly in one spot, and another derivative may be necessary to improve the quality of the phases. The method is therefore called **MIR, Multiple Isomorphous Replacement**. More about MIR phases and their quality later.

B-factors

					X	Y	Z	Q	B	
ATOM	1	N	GLY	1	10.875	48.481	-1.836	1.00	12.64	N
ATOM	2	CA	GLY	1	11.311	48.340	-0.462	1.00	10.90	C
ATOM	3	C	GLY	1	10.077	48.258	0.390	1.00	9.05	C
ATOM	4	O	GLY	1	8.955	48.398	-0.115	1.00	7.70	O
ATOM	8	N	TRP	2	10.301	47.947	1.657	1.00	8.60	N
ATOM	9	CA	TRP	2	9.214	47.795	2.594	1.00	8.31	C
ATOM	10	C	TRP	2	8.546	49.132	2.896	1.00	9.37	C
ATOM	11	O	TRP	2	9.100	50.212	2.642	1.00	10.86	O
ATOM	12	CB	TRP	2	9.785	47.131	3.855	1.00	6.63	C
ATOM	13	CG	TRP	2	10.177	45.657	3.665	1.00	7.58	C
ATOM	14	CD1	TRP	2	9.267	44.652	3.862	1.00	7.30	C
ATOM	15	CD2	TRP	2	11.410	45.173	3.291	1.00	7.45	C
ATOM	16	NE1	TRP	2	9.927	43.541	3.622	1.00	7.20	N
ATOM	17	CE2	TRP	2	11.210	43.781	3.295	1.00	6.65	C
ATOM	18	CE3	TRP	2	12.664	45.728	2.986	1.00	7.70	C
ATOM	19	CZ2	TRP	2	12.273	42.937	2.965	1.00	6.90	C
ATOM	20	CZ3	TRP	2	13.724	44.870	2.672	1.00	7.19	C
ATOM	21	CH2	TRP	2	13.524	43.486	2.639	1.00	6.74	C

What is B? If the atoms in a crystal were idealized points, then specification of each atom's positions by x,y,z would be sufficient. However, atoms are not idealized points and do occupy some volume of space. What is the volume of space that an atom occupies? That is related to the degree of average motion it undergoes; for example, you might expect the atoms in amino acids that are in the core of a protein to be very tightly packed and to not move too much, while surface residues may be more mobile. This B-factor is a measure of this 'envelope of motion'. The contents of this envelope or distribution is the number of electrons in that atom, but the distribution can be very sharp (little or no motion) or broad (more motion). If a particular part of a protein is very mobile and has no long-time-averaged average center-of-mass position, then it will not possess long-range order through the crystal. Therefore, that portion of the protein will not contribute to the diffraction pattern, and will not be determined in the structure. Note: the best that you can

achieve in a crystal structure determination is an accurate image of the average structure of the contents of the asu. By using a simple harmonic oscillator model (atoms as balls on springs), B can be used to estimate the average rms displacement $\langle u \rangle^2$

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

B has units of \AA^2

So, some dynamics information can be obtained from a crystal structure.

Crystallographic Refinement

Crystallographic refinement can be formulated as a chemically-restrained non-linear optimization problem. The goal is to optimize the simultaneous agreement of an atomic model with observed data and known chemical information.

$$E = E_{\text{chem}} + w_{\text{data}} E_{\text{data}}$$

where:

E_{chem} comprises empirical information about chemical interactions between atoms in the model. It is a function of all atomic positions and includes information about both covalent and non-bonded interactions.

E_{data} describes the difference between observed and calculated data.

w_{data} is a weight chosen to balance the gradients arising from the two terms.

Note that refinement is generally NOT an unrestrained fit solely to experimental data. Stereochemical and other 'outside' information is used as well. A well-refined model will have good stereochemistry and good agreement with the observed amplitude data.

The crystallographic R-factor is used.

$$R = \frac{\sum_{hkl} \left| |F_{obs}(hkl)| - k |F_{calc}(hkl)| \right|}{\sum_{hkl} |F_{obs}(hkl)|}$$

[In some cases (SAD or MAD) the $\phi_{\text{calc,exp}}$ can be so accurate that they are also fit to with ϕ_{calc} .]

Various minimization methods are used to minimize an error function such as the one above, which will lead to a reduction in R over the course of the refinement. However, use of R alone can lead to problems. Specifically, determination of the proper number of parameters to use in refinement is nontrivial. For 'simple' functional fitting, one can count the datapoints and count the number of free parameters in the fitting function and easily determine if the procedure is over- or under-determined. However, because the atoms are not

clear what parameters are statistically and physically valid to use. (For instance, x,y,z of atoms, but what about group or individual B-factors? Use of ordered water molecules?) More parameters in a fitting function (in this case the function is the macromolecular model) will fit the data better and will lower R.

What to do?

The free-R factor (R_{free}) is always utilized in modern macromolecular crystallographic refinement. R_{free} is an example of the statistical method of cross-validation. The idea is simple. Before any refinement is performed, a fraction of the F_{obs} dataset (typically ~ 5%) is selected randomly and omitted from refinement; this is the test set. That is, the refinement calculation is seeking to minimize some error function that includes minimizing the differences between F_{calc} and F_{obs} . So, of course one expects R to decrease because the minimization is trying to reduce those differences. However, in addition to calculating R for this 'working' set, an R-factor (R_{free}) is calculated for the 'test set'. This test set has not been used to refine against, so if R_{free} drops then the improvement in the model is statistically justified and has been shown to be so by this cross-validation technique. In a refined structure, R_{free} is typically 2-6% higher than R with R having a value between 15 and 25%. These precise values are dependent upon resolution, data quality, etc.

Evaluation of a structure

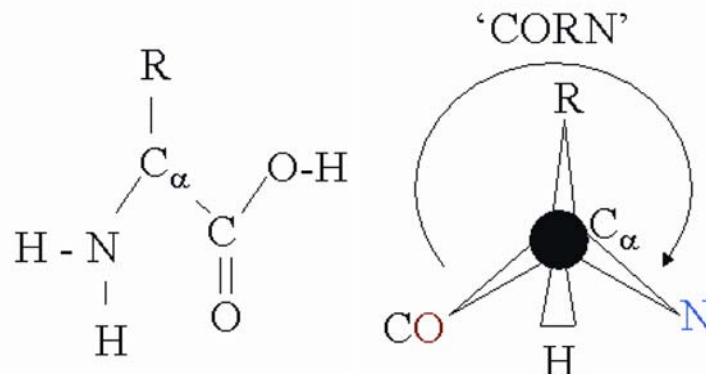
rms deviations of bond-angle and bond-length. These are usually reported, and should typically be 2-2.5 degrees or less

(or up to 3-3.5 degrees for structures of lower resolution). Bond-length rms should be less than about 0.02Å (up to ~ 0.03Å for lower-resolution). The header of the pdb file (especially of more recent depositions) will indicate other issues such as close contacts, etc. There are also programs that can be run on a pdb file to check for features that may be peculiar. Also, the chemical environment of sidechains should make sense.

What do I mean by that?

Amino acids

The basic structure of an α -amino acid is quite simple. R denotes any one of the 20 possible side chains (see table below). We notice that the C_{α} -atom has 4 different ligands (the H is omitted in the drawing) and is thus [chiral](#). An easy trick to remember the correct L-form is the CORN-rule: when the C_{α} -atom is viewed with the H in front, the residues read "CO-R-N" in a clockwise direction.

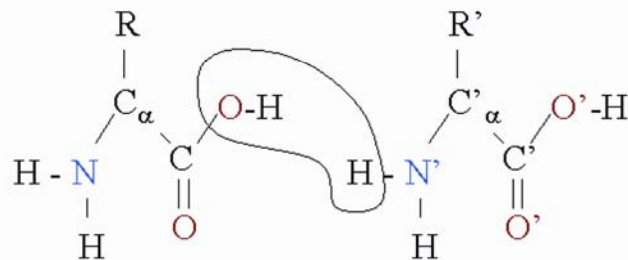


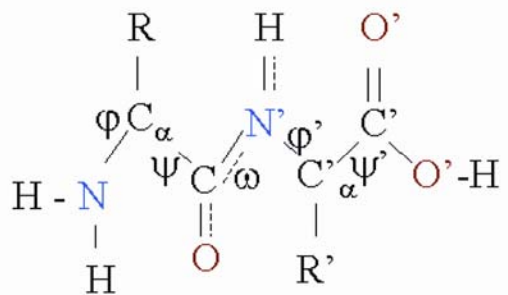
The different side chains R determine the chemical properties of the amino acid or residue (the residue is the amino acid side chain plus the peptide backbone, see below).

Name (Residue)	3-letter code	Single code	Relative abundance (%) E.C.	MW	pK	VdW volume(Å ³)	Charged, Polar, Hydrophobic
Alanine	ALA	A	13.0	71		67	H
Arginine	ARG	R	5.3	157	12.5	148	C+
Asparagine	ASN	N	9.9	114		96	P
Aspartate	ASP	D	9.9	114	3.9	91	C-
Cysteine	CYS	C	1.8	103		86	P
Glutamate	GLU	E	10.8	128	4.3	109	C-
Glutamine	GLN	Q	10.8	128		114	P
Glycine	GLY	G	7.8	57		48	-
Histidine	HIS	H	0.7	137	6.0	118	P,C+
Isoleucine	ILE	I	4.4	113		124	H
Leucine	LEU	L	7.8	113		124	H
Lysine	LYS	K	7.0	129	10.5	135	C+
Methionine	MET	M	3.8	131		124	H
Phenylalanine	PHE	F	3.3	147		135	H
Proline	PRO	P	4.6	97		90	H
Serine	SER	S	6.0	87		73	P
Threonine	THR	T	4.6	101		93	P
Tryptophan	TRP	W	1.0	186		163	P
Tyrosine	TYR	Y	2.2	163	10.1	141	P
Valine	VAL	V	6.0	99		105	H

The polypeptide chain

Two amino acids are combined in a condensation reaction. Notice that the peptide bond is in fact planar due to the delocalization of the electrons. The sequence of the different amino acids is considered the **primary structure** of the peptide or protein. Counting of residues always starts at the N-terminal end (NH₂-group).





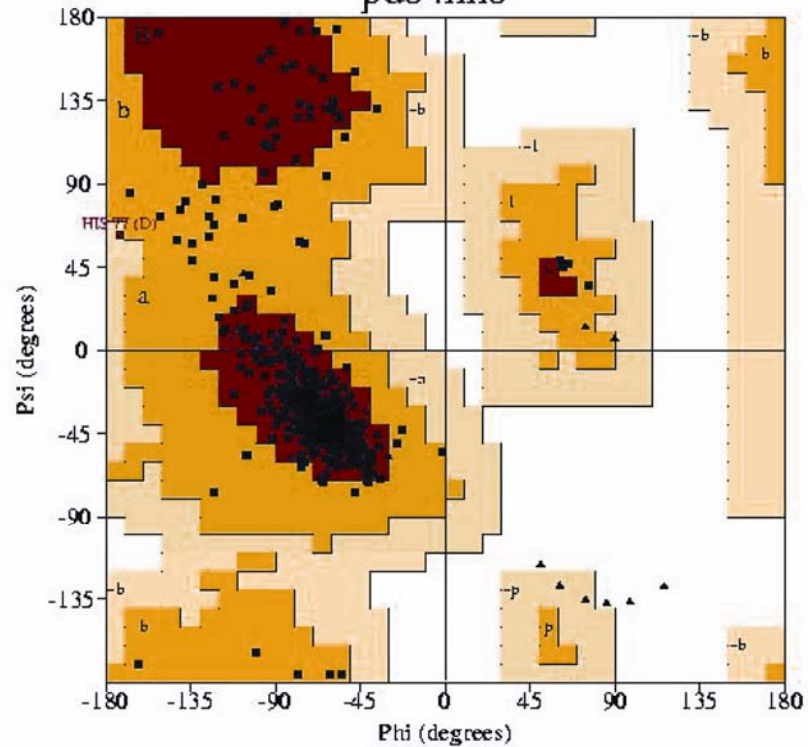
In contrast to the rather rigid peptide bond angle ω (always close to 180 deg) , the bond angles phi ϕ and psi ψ can have a certain range of possible values. They are restrained by geometry to allowed ranges typical for particular secondary structure elements, and represented in a Ramachandran plot (discussed below). A few important [bond lengths \[1\]](#) are given in the table below.

Peptide bond	Average length	Single Bond	Average length	Hydrogen Bond	Average (± 0.3)
C α - C	1.53 (Å)	C - C	1.54 (Å)	O-H ... O-H	2.8 (Å)
C - N	1.33 (Å)	C - N	1.48 (Å)	N-H ... O=C	2.9 (Å)
N - C α	1.46 (Å)	C - O	1.43 (Å)	O-H ... O=C	2.8 (Å)

PROCHECK

Ramachandran Plot

pdb4hhb



Plot statistics

Residues in most favoured regions [A,B,L]	444	89.2%
Residues in additional allowed regions [a,b,l,p]	53	10.6%
Residues in generously allowed regions [-a,-b,-l,-p]	1	0.2%
Residues in disallowed regions	0	0.0%
Number of non-glycine and non-proline residues	498	100.0%
Number of end-residues (excl. Gly and Pro)	8	
Number of glycine residues (shown as triangles)	40	
Number of proline residues	28	
Total number of residues	574	

Based on an analysis of 115 structures of resolution of at least 2.0 Angstroms and R-factor no greater than 20%, a good quality model would be expected to have over 90% in the most favoured regions.

Table 1. The crystal systems and related data for a chiral molecule

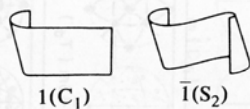
System	Necessary cell parameters	Bravais lattice	Class	Number	Available space groups	Multiplicity
Triclinic	$a, b, c, \alpha, \beta, \gamma$	P	1	1	P1	1
Monoclinic	a, b, c, β ($\alpha = \gamma = 90^\circ$)	P C	2	3–4 5	P2, P2 ₁ C2	2 4
Orthorhombic	a, b, c ($\alpha = \beta = \gamma = 90^\circ$)	P C F I	222	16–19 20–21 22 23–24	P222, P222 ₁ , P2 ₁ 2 ₁ 2 ₁ , P2 ₁ 2 ₁ 2 C222, C222 ₁ F222 I222, I2 ₁ 2 ₁ 2 ₁	4 8 16 8
Tetragonal	$a (=b), c$ ($\alpha = \beta = \gamma = 90^\circ$)	P I P	4	75–78 79–80 89–96	P4, P4 ₁ , P4 ₂ , P4 ₃ I4, I4 ₁ P422, P4 ₂ 2, P4 ₁ 22, P4 ₁ 2 ₁ 2, P4 ₂ 22, P4 ₂ 2 ₁ 2, P4 ₃ 22, P4 ₃ 2 ₁ 2 I422, I4 ₁ 22	4 8 8 8
Trigonal	$a (=b), c, \gamma = 120^\circ$ ($\alpha = \beta = 90^\circ$)	P R P P	3	143–145 146 149, 151, 153 150, 152, 154	P3, P3 ₁ , P3 ₂ R3 P312, P3 ₁ 12, P3 ₂ 12 P321, P3 ₁ 21, P3 ₂ 21	3 3 6 6
Hexagonal	$a (=b=c), \alpha = \beta = \gamma \neq 90^\circ$ $a (=b), c, \gamma = 120^\circ$ ($\alpha = \beta = 90^\circ$)	R P	6	155 168–173 177–182	R32 P6, P6 ₁ , P6 ₂ , P6 ₃ , P6 ₄ , P6 ₅ P622, P6 ₂ 22, P6 ₃ 22, P6 ₃ 22, P6 ₄ 22, P6 ₅ 22	6 6 12
Cubic	$a (=b=c)$ ($\alpha = \beta = \gamma = 90^\circ$)	P F I I P F I	23	195, 198 196 197 207–8, 212–3 209–210 211, 214	P23, P2 ₁ 3 F23 I23 P432, P4 ₃ 32, P4 ₃ 32, P4 ₁ 32 F432, F4 ₃ 32 I432, I4 ₁ 32	12 48 24 24 96 48

The lattice types are P (= primitive), C (= C-face centred), F (= all faces centred), and I (= body centred). Alternative lattice types may occasionally be chosen. The symbols under Class refer to the rotational symmetry axes which are a characteristic of it. The Herman–Mauguin nomenclature for space groups gives the lattice type first, then the symmetry elements in an order which depends upon the crystal system. Refer to *International tables for X-ray crystallography*, Volume A, for a fuller explanation of these symbols. Number refers to the number in International Tables. Multiplicity gives the number of copies of the asymmetric unit in the unit cell.

Table 3. Conditions affecting possible reflections

Element	Symbol	Reflection observed for	Notes
Primitive lattice	P		
Lattice centred on the C face	C	hkl with $h + k$ even	The C face is contained by a and b
Face centred lattice	F	hkl with h, k , and l all odd or all even	
Body-centred lattice	I	hkl with $h + k + l$ even	
Rhombohedral lattice	R	$-h + k + l = 3n$	' $= 3n$ ' means divisible by 3
Twofold screw axis $\parallel c$	2 ₁	00 l with l even	For an axis along a , the row is $h00$
Threefold screw axes $\parallel c$	3 ₁ , 3 ₂	00 l with $l = 3n$	The two possible 3-fold axes have the same pitch but opposite hands
Fourfold screw axes $\parallel c$	4 ₁ , 4 ₃ 4 ₂	00 l with $l = 4n$ 00 l with l even	cf. the 2-fold screw axis
Sixfold screw axes $\parallel c$	6 ₁ , 6 ₅ 6 ₂ , 6 ₄ 6 ₃	00 l with $l = 6n$ 00 l with $l = 3n$ 00 l with l even	cf. the 3-fold screw axes cf. the 2-fold screw axis

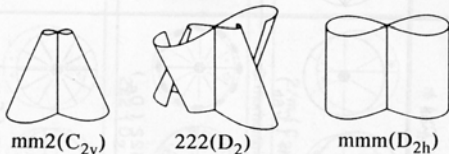
Triclinic



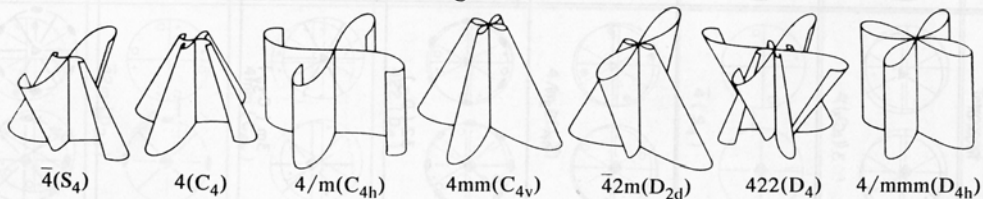
Monoclinic



Orthorhombic



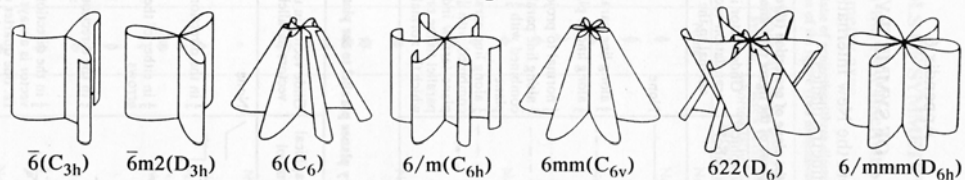
Tetragonal



Trigonal



Hexagonal



Cubic

