# Chem 406: Biophysical Chemistry
# Lecture 7: Structure Determination by X-ray Crystallography

---

**I.** Introduction

**A.** Most of the structures in the Protein Data Bank (PDB) were determined by x-ray diffraction.

1. PDB Statistics
2. Other names for x-ray diffraction
   a. x-ray crystallography
   b. crystallography

**B.** In the past couple of years there has been a growing number of structures, particularly of small proteins and peptides, that have been solved using a combination nuclear magnetic resonance (NMR) and computational chemistry.

1. PDB Statistics

**C.** The structural information obtained from these techniques are the coordinates of the atoms in the molecule.

**Overhead 1 (Print out of a PDB file)**

**II.** Image magnification

**A.** We will look first at techniques that determine structures by producing images; these include light microscopy, electron microscopy and X-ray crystallography; first, though, we need to look into the properties of light.

**B.** Light

1. The wavelike nature of light (electromagnetic radiation)
   a. Light waves are depicted as having oscillating electric and magnetic field components that are at right angles to one another.
   b. These waves are described using the following parameters.

**Overhead 2 and on board ( a light wave)**

i. wavelength ($\lambda$)
ii. frequency ($\nu = c/\lambda$)
iii. Energy ($E = h\nu$)
   1. The constant $h$, is *Planck's constant* and is equal to $6.63 \times 10^{-34}$ Js (Joule-seconds).

2. Light scattering and refraction
   a. When light passes through matter the oscillating electrical field polarizes the atoms present in the matter.
      i. Remember that atoms are made of a nucleus containing positively charged protons and is surrounded by negatively charged electrons.

b. The polarization oscillates in at the same frequency as the electric field of the light
c. This produces an oscillating dipole which acts like a transmitting antenna to re-emit the light at the same frequency as the incident light, however, it is sent out in all directions.
d. This phenomenon is called *light scattering*.
e. What we *see* around us is in effect the result of our eyes and brain analyzing scattered light.
f. When light interacts with matter in this way it slows down, i.e., *it is no longer traveling at the same speed as light in a vacuum*.
g. The degree of slowing is given by the **refractive index**
    i. The refractive index is a unitless parameter of a given material and gives the ratio of the speed of light in a vacuum to its speed in that material
h. When light passes from one material to another, each having a different refractive index, the direction of the light beam can bend.
i. According to Snell's law

$$n_1 \sin(\theta_1) = n_1 \sin(\theta_1)$$

**On Board 3 - Illustrate Snell's Law**

    i. where $n_1$ and $n_2$ are the refractive indices of the two substances and $\theta_1$ and $\theta_2$ are the angles of incidence.
    ii. This property is called *refraction*.
    iii. It can be used to make lenses which can form images with light scattered from an object.

**C.** Light microscopy

1. Lenses are used to produce images, which to the observer, appear larger than they actually are.

**Overhead 4 (a microscope lens)**

2. The resolution of a magnified image is limited by the wavelength of light used
    a. Visible light has wavelengths in the range of 400-700 nm.
    b. As a rule of thumb, images cannot be resolved which are smaller than the wavelength of light used.
    c. Therefore, light microscopy is limited to images which are approximately 200 nm (200 nm = 0.2 μm) or larger
3. Proteins have dimensions of around 10 nm (=100 Å) or less, and the average bond length between atoms in a protein is about 0.1 nm (1 Å)

**Overhead 5 (distance scale)**

a. Often we use the non-SI unit distance called an angstrom (1 nm = 10 Å)

4. To resolve things at an Å resolution requires radiation with a much shorter wavelength than visible light

**Overhead 6 (The electromagnetic spectrum)**

**D.** Electron microscopy

1. Light has both a wave-like and particle nature to it.
2. In 1924 Louis de Broglie reasoned that the same should hold true for matter.

$$\lambda = \frac{h}{p}$$

a. where $h$ is Planck's constant, $\lambda$ is the wavelength, and $p$ is the momentum of the particle.

3. The high energy electron beam produced by a 100 kV electron microscope has a wavelength of 0.004 nm (0.04 Å).

**Overhead 7 (Schematic of electron microscope)**

4. A moving electron produces and associated magnetic field
   a. This allows electron beams to be bent with a magnetic "lens".
5. Constraints of the imaging system reduce the practical resolution to about 0.1 nM (1 Å).
6. Constraints imposed by sample preparation reduce it even further to about 2 nm (20 Å)
7. There is a growing number of structures in the PDB that were determined by electron microscopy
   a. PDB Statistics

**Demonstrate how to do an advance search in the PDB on Experimental Method**

**E.** X-rays

1. X-rays are a form of pure electromagnetic radiation.
2. They have a wavelength of around 1Å.

**Overhead 6 (The electromagnetic spectrum with objects)**

a. This, according to Planck's equation, gives them very high energy and is what makes them so dangerous to living tissue.

**III.** X-ray diffraction (X-ray crystallography)

**A.** Introduction

1. Will not be able to give a detailed explanation of how the technique is applied.
   a. However, I wish to leave you with some appreciation for the power of the technique and the complexities involved with applying it.
2. The power of the technique lies in the fact that it can be used to produce 3-dimensional models of molecules at an *atomic* resolution.

**B.** Requirements

1. For 3-dimensional structures, crystals are required
   a. This is basically a sampling requirement
      i. In order to obtain adequate signal, many copies of the molecule must be viewed at the same time.
      ii. And they must be arranged in a orderly array.

**Overheads 8 (crystal lattices)**

      iii. Crystals represent an array of *unit cells* which repeat to produce a 3-dimensional object.
   b. There is a limited number of different lattice types.
      i. There are 14 *Bravais* lattice types:

| | | |
|---|---|---|
| 1. | P | cubic (a=b=c; $\alpha=\beta=\gamma=90°$) |
| 2. | I | cubic |
| 3. | F | cubic |
| 4. | P | tetragonal (a=b≠c; $\alpha=\beta=\gamma=90°$) |
| 5. | C | tetragonal |
| 6. | P | orthorhombic (a≠b≠c; $\alpha=\beta=\gamma=90°$) |
| 7. | C | orthorhombic |
| 8. | I | orthorhombic |
| 9. | F | orthorhombic |
| 10. | P | hexagonal (a=b≠c; $\alpha=\beta=90°$, $\gamma=120°$) |
| 11. | R | rhombohedral (a=b=c,$\alpha=\beta=\gamma\neq90°$) |
| 12. | P | monoclinic(a≠b≠c; $\alpha=\beta=90°\neq\gamma$) |
| 13. | C | monoclinic |
| 14. | P | triclinic (a≠b≠c; $\alpha\neq\beta\neq\gamma$) |

      ii. where
         1. P = primitive (lattice points at the 8 corners)
         2. I = (extra lattice point at center)

3. C= (two extra lattice point on pair of opposite faces
4. F = (extra lattice points on all faces

c. Biological molecules are themselves asymmetric.
   i. There can be one or more molecules or *asymmetric units per* unit cell.

d. Crystals of biological molecules can be hard to grow

**Overhead 9 (protein crystal lattice from BT)**

   i. molecules tend to be large and flexible
   ii. There are only a limited number of lattice contacts in a crystal (≈50% solvent)

**Overhead 10 (BPTI form I, II & III lattices)**

   iii. Many proteins have multiple crystal forms
   iv. Growing is often and art
      1. Often use hanging drop or vapor diffusion method.

**Overhead 11 (Hanging drop)**

2. For proteins, the amino acid sequence is also needed
   a. With recombinant DNA techniques this is now easier to come by than it used to be.

**C.** Diffraction

1. When electromagnetic radiation passes by an edge between two regions of different density it is bent or *diffracted.*
2. When X-rays are past through a crystal they produce what is called a *diffraction pattern*

**Overhead 12 (diffractometer with diffraction pattern)**

a. Most of the x-rays scattered from the atoms in the crystal are eliminated by destructive interference
   i. For light of longer wavelength this would happen with all but the undeviated light.
      1. This is why the crystals look clear with visible light.

b. When the wavelength of light is shorter that dimensions of the unit cell constructive interference will produce a spots on a piece of film, which is held some distance from the crystal.

**Overhead 13 (Constructive and destructive interference)**

c. This pattern looks like an array of spots
   i. The positions and intensities of the spots must be measured.

d. This pattern can be recorded either on film or using *electronic detectors*
   i. In the past x-ray crystallographers would measure the intensities of tens of thousands to hundreds of thousands of spots produced on photographic film
   ii. Array detectors are now available which can collect and digitize signals from large areas simultaneously
      1. This and can markedly reduce the time needed to collect the data.

3. *Bragg's Law* describes the conditions for forming a spot.

**Overhead 14 (Bragg's Law from BT)**

a. The crystal can be viewed as a collection of parallel planes which intersect the vertices of the unit cells.
   i. The analogy given in Branden and Tooze of driving past a tree farm is a good one.
b. The equation $2d\sin(\theta) = n\lambda$ describes the conditions necessary for obtaining a spot.
   i. where $d$ is the distance of separation of the planes, $\theta$, is the angle of incidence, and $n$ is an integer.

4. The positioning of the spots is determined by the crystal lattice
   a. They can be used to determine the size of the unit cell
5. The intensity of each spot is determined by the structure of the unit cell.

**Overheads 15 (Optical diffraction from CS, Rhodes)**

**IV. Demo - Optical diffraction.**

a. Each spot contains information about the whole unit cell.

**Demo Kevin Cowton's site**

**V.** Fourier Series

**A.** Sine an Cosine waves have three paramenters:

**Overhead 16 (Figure 2.13 from Rhodes.)**

1. Amplitude, $F_o$
2. Wavelength, $h$
3. Phase, $\alpha$

$$f(x) = F_o \cos\left(2\pi\left[hx + \alpha\right]\right) \text{ or}$$

$$f(x) = F_o \sin\left(2\pi\left[hx + \alpha\right]\right)$$

a. A sine wave is just equal to a cosine wave that has a phase shift of 90° .

$$\sin(x) = \cos(x + \pi/2)$$

**B.** More complex periodic functions functions can be represented by a sum of cosine waves with various amplitudes, $F_h$, frequencies, $h$, and phases, $\alpha_h$.

1. Fourier Series

$$\begin{aligned} f(x) &= F_o \cos\left(2\pi\left[0x + \alpha_o\right]\right) \\ &+ F_1 \cos\left(2\pi\left[1x + \alpha_1\right]\right) \\ &+ F_2 \cos\left(2\pi\left[2x + \alpha_2\right]\right) \\ &+ F_3 \cos\left(2\pi\left[3x + \alpha_3\right]\right) \\ &\ldots \\ &+ F_n \cos\left(2\pi\left[nx + \alpha_n\right]\right) \\ &= \sum_{h=0}^{n} F_h \cos\left(2\pi\left[hx + \alpha_h\right]\right) \end{aligned}$$

a. Alternatively, a cosine function of any phase can be represented as a combination of a cosine and sine function, each with a phase of zero:

$$f(x) = \sum_{h=0}^{n}\left(F_{h,\cos}\cos(2\pi hx) + F_{h,\sin}\sin(2\pi hx)\right)$$

b. It is also convenient to represent this as a complex number where $\cos(2\pi hx)$ is the real component and $\sin(2\pi hx)$ is the imaginary component:

$$f(x) = \sum_{h=0}^{n} F_h\left(\cos(2\pi hx) + i\sin(2\pi hx)\right)$$

c. Since $\cos(x) + i\sin(x) = e^{ix}$ (Euler's formula)

$$f(x) = \sum_{h=0}^{n} F_h e^{i2\pi hx}$$

2. Since the structures of crystals are periodic, they too can be represented by a Fourier Series.
   a. The diffraction pattern, which is called the *reciprocal lattice*, is also a periodic function.

**Overhead 18 (Figure 2.5 from Rhodes)**

b. It is called a *reciprocal lattice* because everything is inverted in reciprocal space

i. Things that are far apart in reciprocal space contain information about things that are close together in real space.

**Overheads 19 (Figures 2.15 & 2.16 from Rhodes)**

3. Each spot in the diffraction pattern is called a *structure factor.* Each can be thought of as the sum of the scattering contributions, *f*, from each of the atoms *A, B, C, ...* in the unit cell:

$$F_{hkl} = f_A + f_B + ... + f_{A'} + f_{B'} + ... + f_{F'}$$

**Demo Interactive Structure factor tutorial**

4. Another way to represent the structure factors is to divide the unit cell up into small volume elements, each having an electron density, ρ. Each structure factor is then represented by the sum of the scattering contributions from each of the volume elements in the unit cell:

$$F_{hkl} = f(\rho_1) + f(\rho_2) + ... + f(\rho_m) + ... + f(\rho_n)$$

5. The structure factors can therefore also be represented as a Fourier series:

$$F(hkl) = \sum_x \sum_y \sum_z \rho(xyz) e^{2\pi i(hx+ky+lz)}$$

**VI.** Fourier Transforms

**A.** When one function represents the amplitude for the Fourier series of a second function then the dependent variables for the two functions are reciprocally related.

1. In X-ray diffraction, the dependent variables for the electron density function, *x, y and* z, have dimensions of distance, where as the dependent variables for the structure factor, *h, k* and *l*, have dimensions of frequency, or one over distance.
2. When this situation exists, there exists a mathematical operation called the Fourier Transform, can be used to convert between the two functions.
3. For example, for a one dimensional, discrete function, *F*(*h*), which is represented as the Fourier series of a second function, *f*(*x*):

$$F(h) = \sum_{x=0}^{\infty} f(x) e^{i2\pi hx}$$

a. The Fourier transform is given as

$$f(x) = \sum_{h=0}^{\infty} F(h) e^{-i2\pi hx}$$

4. When this is extended to 3-dimensions

$$F(hkl) = \sum_x \sum_y \sum_z f(xyz) e^{i2\pi(hx+ky+lz)}$$

   a. The Fourier transform is given as

$$f(xyz) = \sum_h \sum_k \sum_l F(hkl) e^{-i2\pi(hx+ky+lz)}$$

5. Substituting $\rho(xyz)$ for $f(xyz)$ we get

$$F(hkl) = \sum_x \sum_y \sum_z \rho(xyz) e^{i2\pi(hx+ky+lz)}$$

   a. The Fourier transfrom is given as

$$\rho(xyz) = \sum_h \sum_k \sum_l F(hkl) e^{-i2\pi(hx+ky+lz)}$$

   b. This provides with a method of solving for the electron density function.

6. Similar Fourier transform pairs are used to perform spectral analyses.
   a. For example, if you record a periodic sound in the time domain (s), such as bell ringing, its Fourier transform will give a spectrum of the various frequency components in the frequency domain (1/s).
   b. This methods of data collection is used in both FT-NMR and FT-IR spectrometers.

**B.** Solving for the electron density

1. As indicated earlier, each structure factor *F(hkl)* is itself a complex number that is given by an amplitude, phase and frequency
   a. Demonstrate this with the Interactive Structure Factor tutorial

**Demo Interactive Structure factor tutorial**

**Overhead 21 (Fig 17.8 from BT, wave parameters)**

2. The frequency is determined by the Miller indices (*h,k,l*), which determine the frequency of the planes cutting through the unit cell.
3. The amplitude of each structure factor $|F(hkl)|$, can be determined from the intensity of each spot $I(hkl)$:

$$|F(hkl)| = \sqrt{I(hkl)}$$

4. What's missing is the phase information
   a. We have taken a black-and-white photo instead of a color one

b. Information about the phase angle, however, is lost. The phase is the offset for the planes designated by given set of Miller indices, so that the planes cut through the regions of highest electron density.

5. All three pieces of information are needed for each spot in order to obtain the structure of the unit cell from the spots

**C.** Solving the phase problem

1. **Multiple isomorphic replacement** (MIR)

a. The protein is labeled in a small number of places with heavy metals, which scatter considerably more than H,C,O,N,P,or S.

i. This is called **isomorphic replacement** when the only difference between the labeled and unlabeled crystals is the presence of the metals

b. The Fourier transform of the intensity differences with and without the metals can be used to map the locations of the metals in the unit cell

i. This basically assumes that the phases are all zero

$$P(xyz) = \sum_h \sum_k \sum_l \left|F(hkl)\right| e^{-i2\pi(hx+ky+lz)}$$

c. The data are displayed as maps of vectors representing distances between the metals in the unit cell .

i. This map is called a **Patterson map**

**Overhead (Figure 18.9 from Brandon & Tooze)**

d. From this map it is possible to determine the amplitude and phase of the contributions from the metals

e. We also know the amplitudes for the protein alone and the protein with the metal

f. With information it is possible to limit the possible phase angles for the protein alone to two possible solutions.

g. By using a minimum of two isomorphic replacements it is possible to determine the right solution.

h. When this information is obtained for two or more isomorphic replacements it is possible to calculate the amplitudes and phases for the spots on the diffraction pattern

i. Using these initial phases can come up with crude model

2. Molecular replacement

a. If you have an existing structure that is similar to the one you are trying to solve for, you can use it to come up with an initial set of phases.

3. c. The phases are used to determine an **electron density map**.

**D.** Model building

**Overhead (CS, electron density map and resolution)**

4. The 3-dimensional structure is obtained by making a model which conforms with the electron density map
5. An amino acid sequence is required to successfully determine a structure
6. The process is subjective and subject to errors
   a. The choice of phase angles is a major source of error
   b. Depending on how far out the data was collected on the diffraction pattern, the resolution is limited
7. At low resolution (5Å) the shape of the molecule can be seen.
   a. For proteins, α-helices can some times be identified.
8. At medium resolution (3Å) the polypeptide can be traced through the molecule.
   a. The general location of side chains can be discerned
9. At high resolution (<2Å) it is possible to distinguish between the different types of amino acids
   a. For example, between leucine and isoleucine.
10. At very high resolution (1Å) it is possible to make out individual atoms.
11. Refinement
    a. For data to greater than 2.5Å resolution, most of the errors can be mitigated by refinement.
    b. The model is changed to minimize the difference between the observed and calculated diffraction intensities.
       i. To do this a hypothetical diffraction map is made from the existing map by Inverse Fourier Transformation.
       ii. R-factor

$$R = \frac{\sum \left| \left| F(hkl)_{obs} \right| - \left| F(hkl)_{calc} \right| \right|}{\sum \left| F(hkl)_{obs} \right|}$$

          1. A value of 0 for perfect agreement
          2. A value of 0.59 for total disagreement.
          3. A value of 0.15-2.0 is obtained for a well-determined structure.
       iii. Thermal factors
          1. Debye-Waller Factors (B) or temperature factors
          2. These are obtained from including isotropic thermal motion into the fitting procedure.
          3. They are often reported in the PDB files
          4. They are a measure of thermal motion.
    c. Simulated annealing
       i. Is a more recent refinement method which uses **molecular dynamics** simulations to search for low energy conformations.