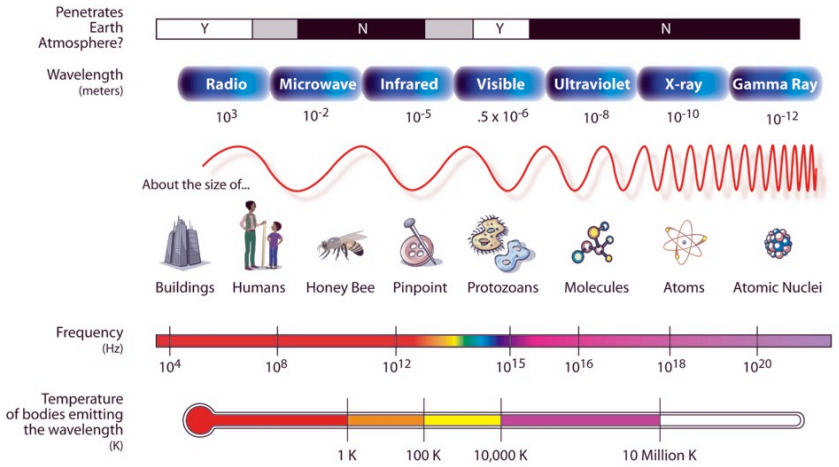
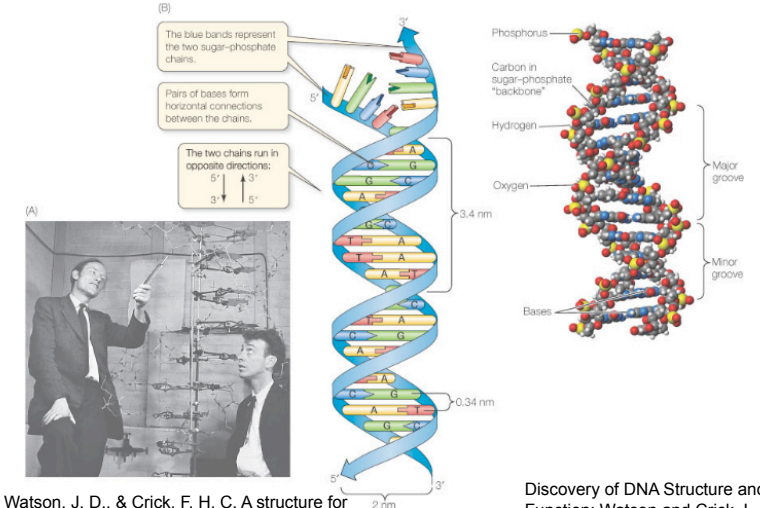
BIOCH 590: Biomacromolecules Spring 2010

Protein X-Ray Crystallography

THE ELECTROMAGNETIC SPECTRUM



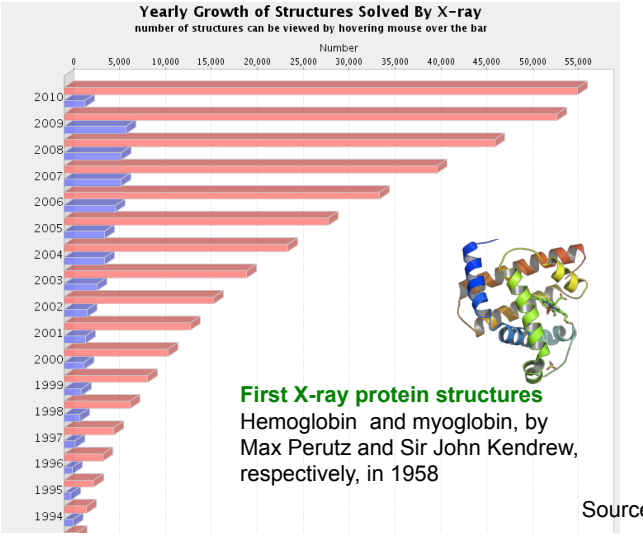
Discovery of DNA Double Helix



Watson, J. D., & Crick, F. H. C. A structure for deoxyribose nucleic acid. Nature 171, 737–738 (1953)

Discovery of DNA Structure and Function: Watson and Crick, L. A. Pray, Nature Education 1 (2008).

Over 85% of Known Protein Structures from X-ray

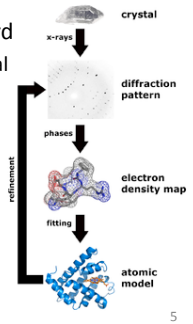


First X-ray protein structures
Hemoglobin and myoglobin, by Max Perutz and Sir John Kendrew, respectively, in 1958

Source: PDB

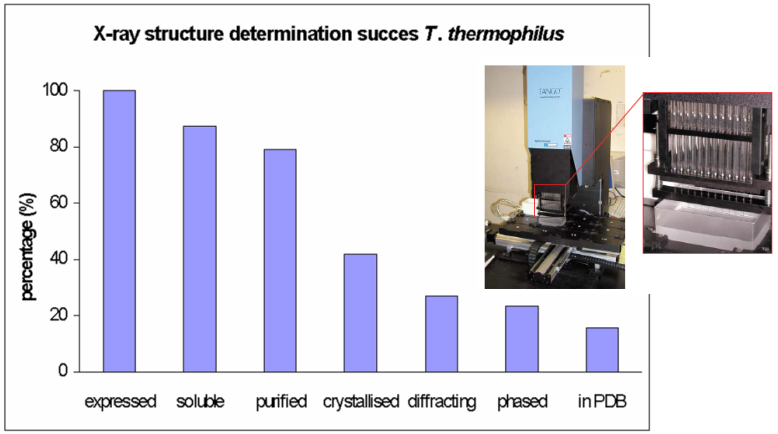
Protein Structure Determination by X-ray Crystallography

- Protein purification: first step of all structural determination efforts
- Protein Crystallization: the most challenging part!
  - Produce well-ordered protein mono-crystals without any inclusion and large enough to diffract X-Ray beam
  - More art than science: largely impossible to predict crystallization conditions
- Crystal mounting: transfer of crystals from solution to detector
  - Proper alignment can improve completeness
- Diffraction data collection: central but somewhat standard
  - Record monochromatic reflections while rotating the crystal
  - Multi-wavelength data collection on synchrotrons
- Data processing: indexing, integration and scaling
- Structure construction: phase problem
  - Molecular replacement; Heavy atom methods
- Structure refinement and analysis:
  - Resolution, R-factor, what about biology!?



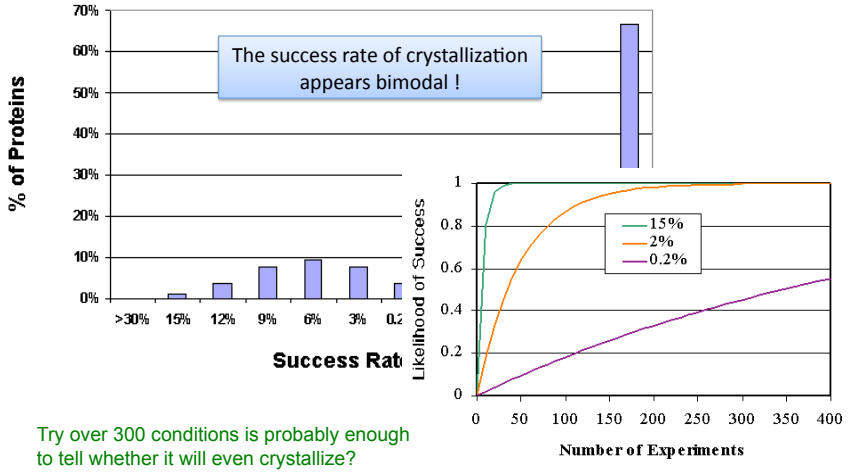
http://proteincrystallography.org/ & Wikipedia

High-throughput protein crystallography: *bottleneck is crystallization*



http://www.bfsc.leidenuniv.nl/teaching

Crystallization space and screening: How much is too much?

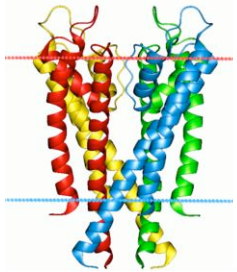


Try over 300 conditions is probably enough to tell whether it will even crystallize?

http://www.ruppweb.org/Xray/tutorial/High_Throughput_EMBL_full.htm

Roderick MacKinnon
Nobel Prize in Chemistry 2003

Growing Crystal Remains Largely an Art



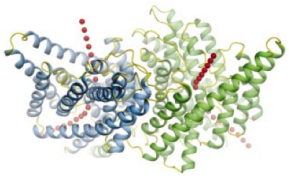
KcsA (PDB: 1r3j)

Nothing automatic about ion-channel structures

Sir— My colleagues and I were shocked to read your News report "Protein chemists favour automatic answers" (ref. 1) in which the chloride ion channel was featured prominently as an example of an important protein structure determined with the help of high-throughput techniques. In the report, Neil Isaacs of Glasgow University is quoted as saying that the chloride ion-channel structure "could not have been done without automation".

In fact, we used no automation or high-throughput methods to solve the chloride-channel structure[2]. Indeed, high-throughput methods have played no part in any of the difficult ion-channel structure determinations completed in my laboratory[3–5]. Our success has rested solely on the intense focus, hard work and thoughtful approach of a small group of scientists intent on solving an important problem in biological chemistry.

I do not wish to join the debate over the wisdom of funding robotic structural biology in the United Kingdom. I do, however, wish to set the record straight concerning a misrepresentation of the science carried out in my own laboratory. The explanation for why we have made

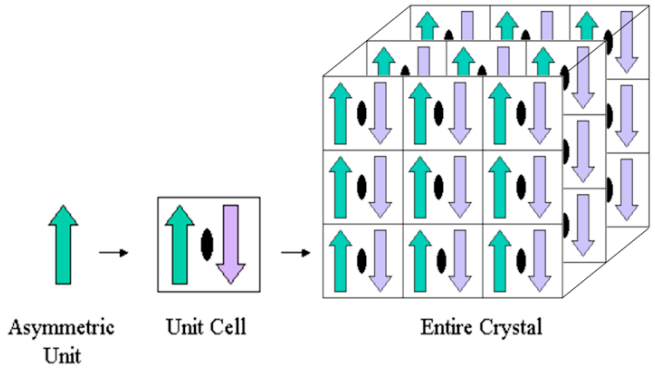


Chloride ion channel: structure solved by traditional science.

MacKinnon, Nature (2002).

Crystal: Asymmetric Unit and Unit Cell

- Crystal: solids with (translational) repeats of a symmetric motif (unit cell)
- An asymmetric unit is the smallest unit of volume that contains all of the structural information and that by application of the symmetry operations can reproduce the unit cell.



Symmetry

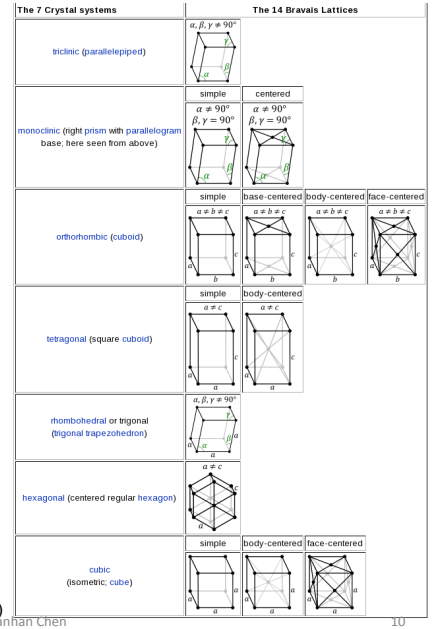
7 crystal Systems
32 point groups (symmetry)
14 lattice types (shape)
=> 230 space groups

International symbol
(Herman-Mauguin symbol):
shortest symbols to general all symmetry elements
e.g., $P3_121$

P: lattice centering type (primitive)
3_1: symmetry of primary axis (trigonal c-axis)
2: symmetry of secondary axis (2 fold)
1: symmetry of tertiary axis

Restriction of chirality:
65 space groups for biomolecules!

See wikipedia for nice overview of crystal symmetry!)



X-ray Diffraction

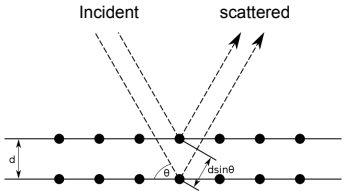
- Determined by electron density (hydrogens largely invisible!)
- Diffraction: maximum amplitudes when all scattered waves are in phase.
- Bragg's Law: (1915 Nobel Prize; WL Bragg was 25 then.)

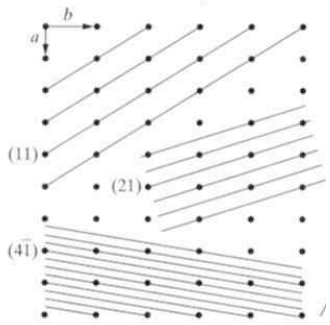
$$2d\sin\theta = n\lambda \rightarrow 2\sin\theta/\lambda = n/d$$

  - Angle of diffraction is quantized
  - Reciprocal relationship: larger spacing leads to smaller diffraction angles
- Miller indices: (*h, k, l*)
  - define the family of lattice planes

$$\frac{2\sin\theta}{\lambda} = \frac{n}{d_{hkl}} = n\sqrt{\frac{h^2}{a^2}+\frac{k^2}{b^2}+\frac{l^2}{c^2}}$$

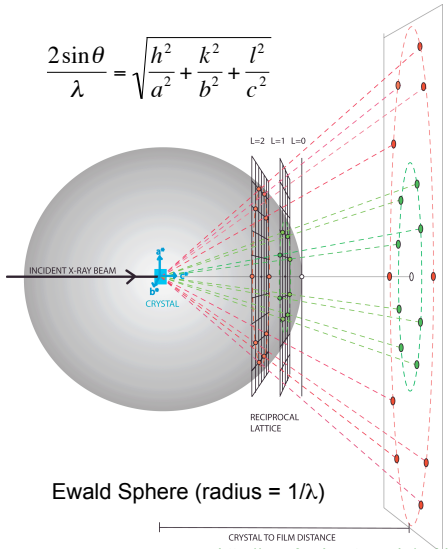
  - von Laue Equation
  - Scattering vector: (*h*/**a**, *k*/**b**, *l*/**c**) (reciprocal space!)





Systematic Rotation to Probe Entire Reciprocal Lattice

$$\frac{2\sin\theta}{\lambda} = \sqrt{\frac{h^2}{a^2}+\frac{k^2}{b^2}+\frac{l^2}{c^2}}$$

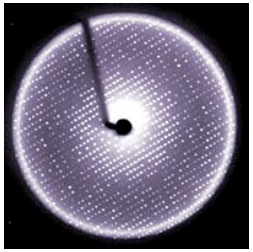


Ewald Sphere (radius = 1/λ)

A single recording of diffraction, i.e., a single reflection, at a fixed crystal orientation only sample a limited number of reciprocal lattice points!

Systematically rotating (a precisely aligned) crystal allows all reciprocal lattice to be sampled (precession imaging).

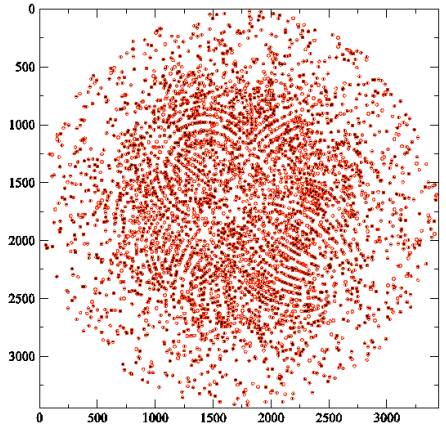
Intensities of the reflections vary dramatically, which is related to the types and positions of atoms within the unit cell.



http://xray0.princeton.edu/~phil/Facility/Guides/XrayDataCollection.html

Diffraction Pattern and Space Group

- Fully automatic?
- Peak picking
- diffraction pattern recognition (systematic absences reveals space group)
- Crystal unit cell morphology (dimensions) determination
- Completeness




```
Indexing 1

Cell lengths (Angstrom):   6.4232     7.3423    15.7870
Cell angles (degree):     95.8000    96.5907   102.5697
…
```

http://renzresearch.com/Precognition/html/chpt2_index.htm

Relating Diffraction Intensities and Atomic Coordinates: Structure Factors

- Multiple atoms within the unit cell diffract at the same angles (Bragg's Law), but with different amplitudes and phases
- Structural factor: sum of scatter waves of all atoms

$$F(h,k,l) = \sum_{j=1}^{N} f_j \, e^{2\pi i(hX_j + kY_j + lZ_j)}$$

f_j: atomic scattering factor
(X_j, Y_j, Z_j): fractional cell coordinates

- Scattered intensities actually observed: $I(h,k,l) = |F(h,k,l)|^2$
  - Information about phase is lost!
- Structural calculation from X-ray diffraction: an inverse problem
  - The forward problem of calculating the diffraction pattern, *I(h,k,l),* from atomic coordinates is trivial
  - The inverse process of calculating atomic structure, (X_j, Y_j, Z_j), is NOT, as the critical phase information is lost (i.e., the famous phase problem).
- Fourier transform between real and reciprocal spaces

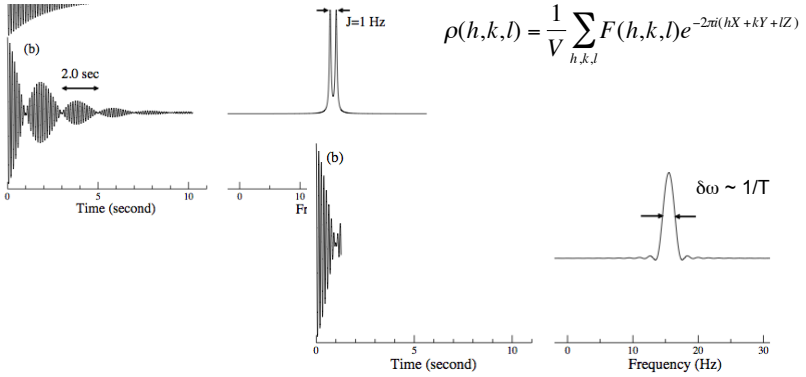
$$\rho(h,k,l) = \frac{1}{V} \sum_{h,k,l} F(h,k,l) e^{-2\pi i(hX + kY + lZ)}$$

ρ: "the electron density"

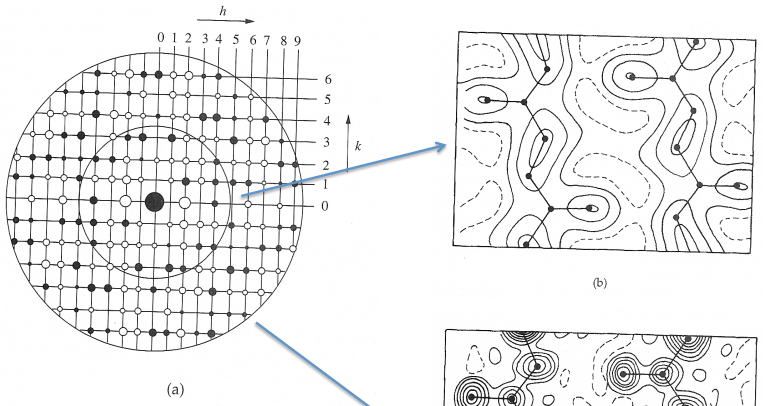
amplitude (known) phase (unknown)

Completeness, Accuracy and Resolution

- Diffractions at larger angles, i.e., higher h, k, l, required for higher resolution (FT uncertainty principle)
- Missing diffractions assumed to be zero by default in the summation: reduced accuracy (and reduced resolution)



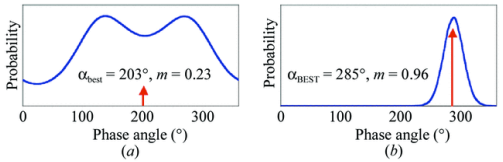
$$\rho(h,k,l) = \frac{1}{V} \sum_{h,k,l} F(h,k,l) e^{-2\pi i(hX + kY + lZ)}$$



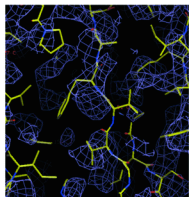
Structural factor data F(h,k): inner circle marks 2A resolution. Including additional data up to the outer circle increases resolution to 1A (Fig 12.16 of Tinoco)

Solving the Phase Problem

- A fundamental problem but with established solutions
- Direct method: small molecules (<1K atoms) with high-resolution data (~1 A)
- Molecular replacement: if initial structural models can be estimated (e.g., through homology modeling), Patterson methods can be used to determined the orientation and thus initial guess of phase.
- Isomorphous replacement: introducing one or more heavy atoms (with large electron densities). Changes in diffraction intensities with and without heavy atoms can be used to obtain initial phase estimation (using either direct or Patterson methods). Often >3 derivatives necessary.
- Anomalous diffraction (MAD and SAD)



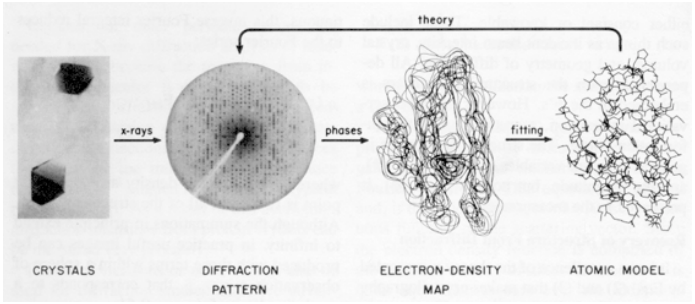
Phase probability for one reflection in a MIR experiment. (a) One derivative. (b) Three derivatives.



A 2.6 Å SIR electron-density map with the final Cα trace of the structure superimposed.

http://journals.iucr.org/d/issues/2003/11/00/ba5050/index.html

Phase/Structure Refinement: R-Factor



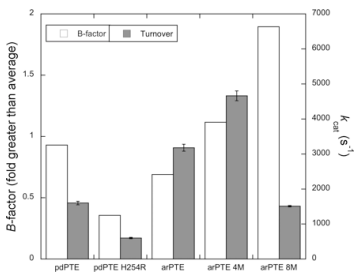
http://www.sci.sdsu.edu/TFrey/Bio750/Bio750X-Ray.html

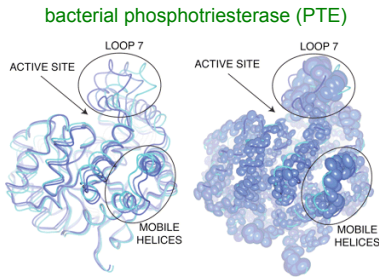
- After initial determination of phase, resulting structural model is used to refine the phases and calculate improved models in an iterative fashion.
- Quality of fitting measured by R-factor:
  - R_{free}: calculated from ~10% of data not already included in refinement.
  - R_{free} ~ resolution / 10 (i.e., ~ 0.2 for 2 A structure)

$$R = \frac{\sum_{\text{all reflections}} |F_o - F_c|}{\sum_{\text{all reflections}} |F_o|}$$

Crystal B-Factors

- Also known as temperature factor (included in PDB files)
- Fit to the data in order to account for the 'blurriness' of the electron density due to thermal motion.
- B-factors of sidechains (5 to 60 Å^2) to be larger and more variable than backbone atoms (5 to 35 Å^2).
- (Unreliable) indication of protein dynamics?



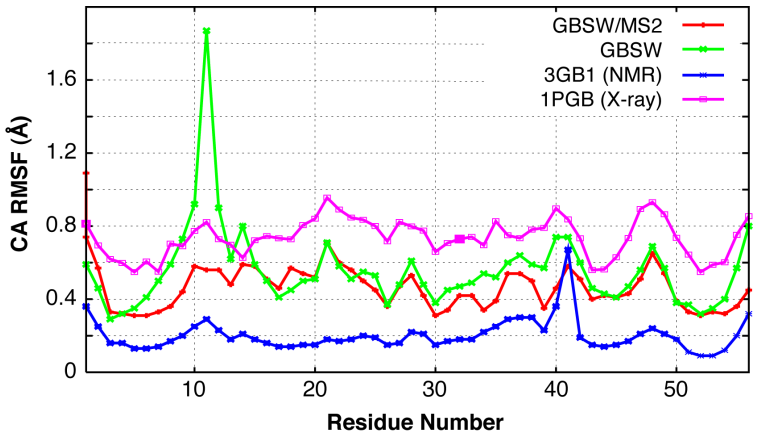


Correlation between loop 7 B factor and kcat in PTE variants

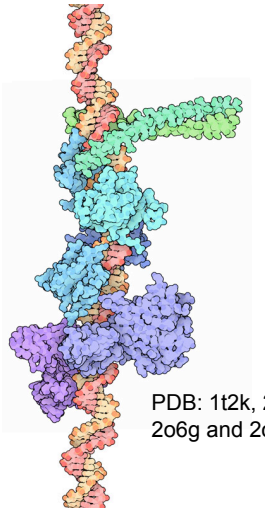
Jackson et al, PNAS (2009)

B-Factor and Conformational Flexibility

B-factor = 8π/3 RMSF^2



PDB
Molecule of the Month:
Enhanceosome



PDB: 1t2k, 2pi0, 2o6g and 2o61

Panne et al, Embo J. (2004)

Table I Data collection (AJ/IRF-3/DNA)

Space group	C2
Molecules in AU	1
Cell parameters (Å)	$a = 186.47$, $b = 65.24$, $c = 83.96$
X-ray source/detector	CHESS A1 (Ithaca, NY)/Quantum 4 CCD
Resolution (Å)	50–3.0
Mosaicity	0.29–0.62[a]
Total reflections	96925/8512[a]
Unique reflections	20151/1958[a]
Multiplicity	4.8/4.3[a]
Completeness (%)	99.6/97.8[a]
R_{symm}[b]	7.5/28.9[a]
I/sigma	19.43/5.78[a]

[a]Highest resolution shell.
[b]$R_{symm} = \sum|Ih - \langle Ih\rangle|/\sum Ih$, where $\langle Ih\rangle$ is the average intensity over symmetry equivalents.

Table II Refinement statistics (AJ/IRF3/DNA)

Resolution (Å)	3.0
R_{free} (test set size/count)	0.297 (5.8%/1178)
R_{work}[a]	0.252
No. of protein atoms	4028
No. of nucleic acid atoms	1265
No. of solvent molecules	9
No. of metal atoms	0
R.m.s.d. bonds (Å)	0.009
R.m.s.d. angles (deg)	1.5
R.m.s.d. dihedral (deg)	20.5
R.m.s.d. improper (deg)	1.82
$\langle B\rangle$ (Å^2) (chain A (IRF-3A), B (IRF-3B), C (DNA), D (DNA), E (c-Jun), H (ATF-2), solvent)	67.3, 97.6, 75.4, 74.3, 121.2, 106.8, 28.6
Ramachandran plot (%) (favorable, additional, generous, disallowed)	83.7/14.3/2.0/0

[a]R_{work} and $R_{free} = \sum||F_o|-|F_c||/\sum|F_o|$, where F_o and F_c are the observed and calculated structure factor amplitudes. R_{free} was calculated with 10% of the reflections not used in refinement.

How about biology?

- Crystal artifacts: crystallization conditions and crystal packing
- What about dynamics?
- Structure is just a beginning …
- NMR and X-ray crystallography is complementary!

NMR	**X-ray crystallography**
short time scale, protein folding	long time scale, static structure
solution, purity	single crystal, purity
< 20kD, domain	any size, domain, complex
functional active site	active or inactive
atomic nuclei, chemical bonds	electron density
resolution limit 2-3.5Å	resolution limit 1.5-3.5Å

Dynamics
Transient interactions