X-ray Crystallography

2-lecture Introduction

Agenda

Goals

- Not to Propagate crystallographers…
- Intelligent reader / user

http://xtal.ohsu.edu/teaching/conj668/X-ray%20Crystallography.pdf

Topics

- Why crystals, X-rays?
- Crystal growth
- Diffraction – why, how?
- Phase problem – solving.
- Density → Atomic model
  - Refinement
  - Accuracy

Why Structures? Why Crystallography?

- Coat-hanger – frame hypotheses
  - Basic biochemistry; rational design…
  - Nature / Science / Cell
- Structural database – ~~109~~ 119,000 (May 20~~15~~16):
  - ~~106,000~~ 116,453 crystal structures
  - ~~11,400~~ 11,817 NMR
  - ~~1050~~ 1514 EM

Bibliography / Resources

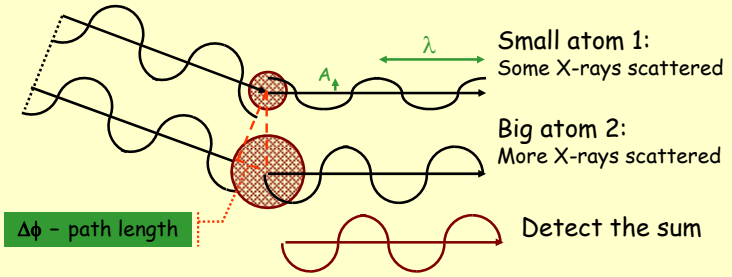
- McPherson, A. (2009) (solid introduction)
  - Introduction to Macromolecular Crystallography, Wiley-Liss, Hoboken, NJ; 2nd Ed., ISBN 978-0470185902 ($77.50)
- Drenth, J. (2010) (more technical)
  - Principles of protein x-ray crystallography. New York, Springer, 3rd Ed., ISBN 978-1441922106 ($84.95)
- Rupp, B. (2010) (comprehensive text; reference)
  - Biomolecular Crystallography, Garland, New York, ISBN 978-0815340812 ($106)
- Rhodes, G. (2010) (less technical)
  - Crystallography made Crystal Clear, Academic Press, 3rd Ed., ISBN 978-0125870733 ($64)

Chapter 2

X-rays & their interactions with Crystalline materials

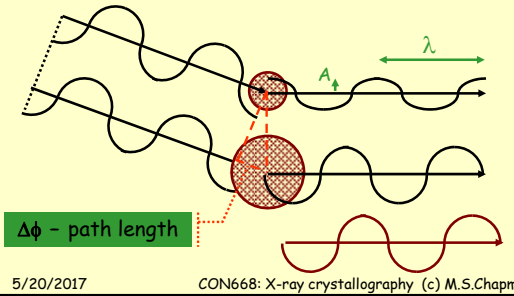
Why X-rays?

- X-rays ~ Electro-magnetic radiation ~ Light
- Scattered in all directions by atoms.
- Intensity (direction) = sum (interference) of scattering.
- Sum depends on size & path length ~ position of each atom, i.e. structure!



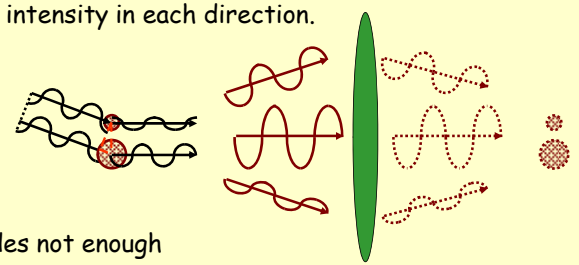
Why X-rays? - 2

- Sum most sensitive to structure when
  - Path length difference = O(wavelength)
- (Visible light:
  - Much longer wavelength
  - Insensitive to atomic-level structure.)



Diffraction - Crystallography in a nutshell

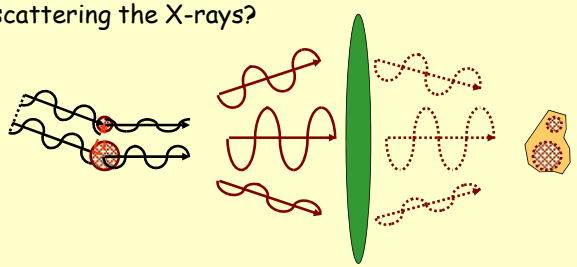
- No X-ray lenses
- Computationally mimic, by summing scattered waves.
  - (Fourier transform)
- Measure intensity in each direction.



- Amplitudes not enough
- Phases – synchronization of waves
  - How they line up, how far peaks lag behind each other.
  - Can't be measured directly – "*Phase Problem*" Challenge.

Image *Electron* density not atomic structure

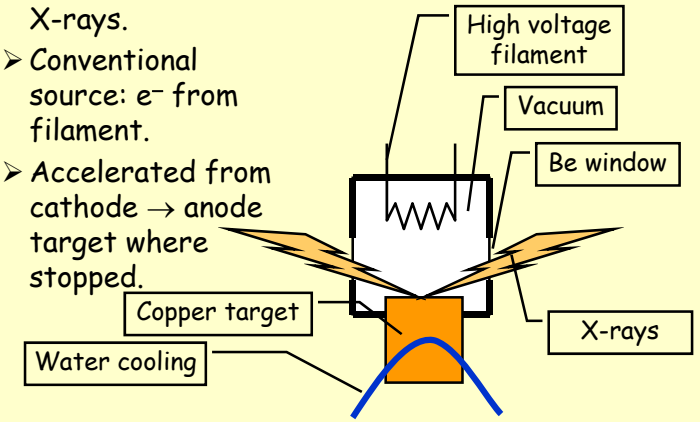
- What is scattering the X-rays?



- Atoms
  - Not nuclei, but electron clouds
- Image electron density – infer nuclear positions
  - Exptl. error in density can → difficult interpretation.

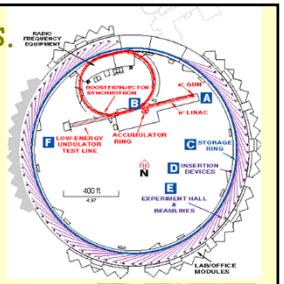
Conventional sources of radiation

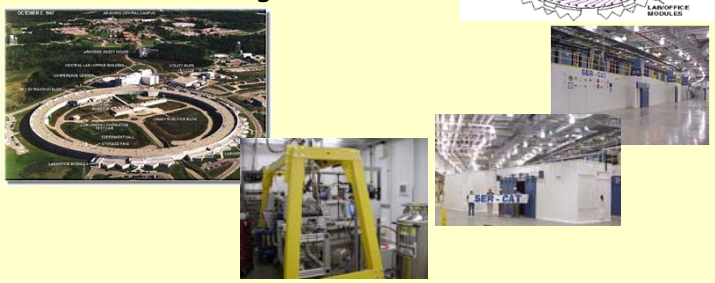
- e^- acceleration → X-rays.
- Conventional source: e^- from filament.
- Accelerated from cathode → anode target where stopped.



Synchrotrons.

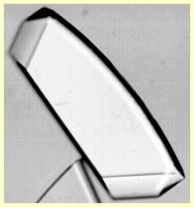
- e^- or e^+ traveling round circle (> 50m).
- Expensive shared facilities.
- High intensity
- Variable wavelength → Phases

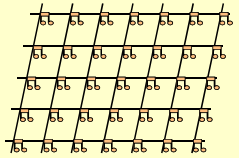


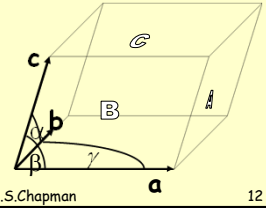


Why crystals?

- Electrons scatter x-rays photons inefficiently (1 in 10^{16})
- Dataset from one molecule ~ 100 trillion yrs
- Solutions – average of all orientations
- Crystals are arrays of ~ 10^{15} molecules with same orientation.

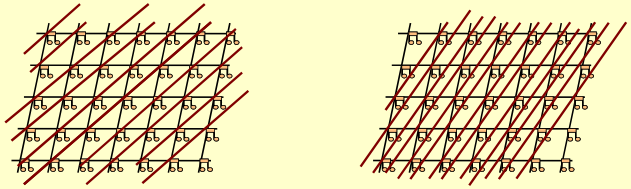






Lattice Planes – rationale for diffraction directions

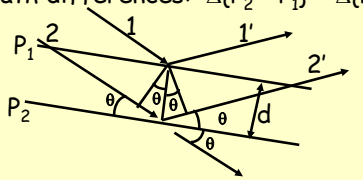
- Plane (Line) through multiple grid points.
- Parallel planes (lines) through every grid point.



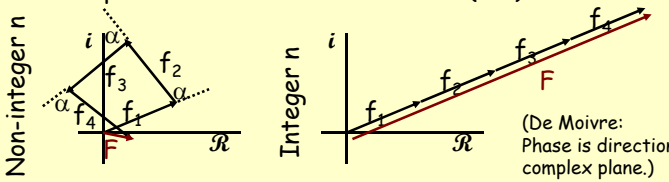
- Infinitely many ways – what's the point?
- Bragg: diffraction in directions of "reflections" from these imaginary planes

Bragg's Law

- Consider || planes P_1, P_2, ... P_j, P_{j+1}, ... P_N.
  - Path differences: $\Delta\{P_2 - P_1\} = \Delta\{P_{j+1} - P_j\} = 2d \sin\theta$



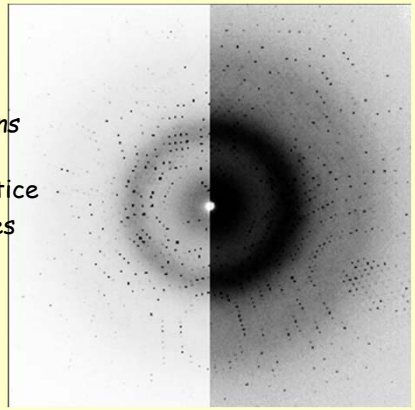
- Σ_{planes} scatter much larger when "in phase".
  - ... path difference = $2d\sin\theta = n\lambda$; (n=1)



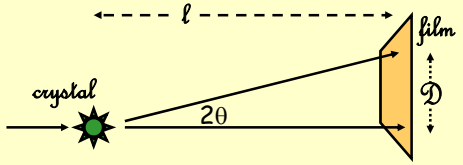
(De Moivre: Phase is direction in complex plane.)

Implications of Bragg's Law

- Particular directions, diffraction strong
  - Elsewhere, ~ zero
  - *spots* a.k.a. *reflections*
- Spots positions → geometry of crystal lattice
- Intensities → amplitudes of scattered waves
  - "Sum" (Fourier transform) → electron density



Bragg's Law → Resolution



- Let D_{max} be distance of furthest spot from direct beam.
- Let d_{min} be its interplanar spacing.
- $d_{min} = \lambda/(2\sin\theta_{max}) = \lambda/2\sin\{\frac{1}{2}\tan^{-1}(D_{max}/\ell)\}$
- d_{min} is *de facto* resolution limit.
  - Smallest spacing between objects that can be resolved
- Note d_{min} refection at θ_{max}, i.e. farthest from beam.

Chapter 3: Crystallization

Empirical w/ some rationale.

What's important to crystal quality?

1. Purity
2. Purity
3. Purity

   97 – 99+ % purity – no other bands on gels.
4. Beyond purity - Homogeneity
   - Post-translational modification
     - Phosphorylation; glycosylation, cleavage…
   - Conformation

2 steps of Crystallization:

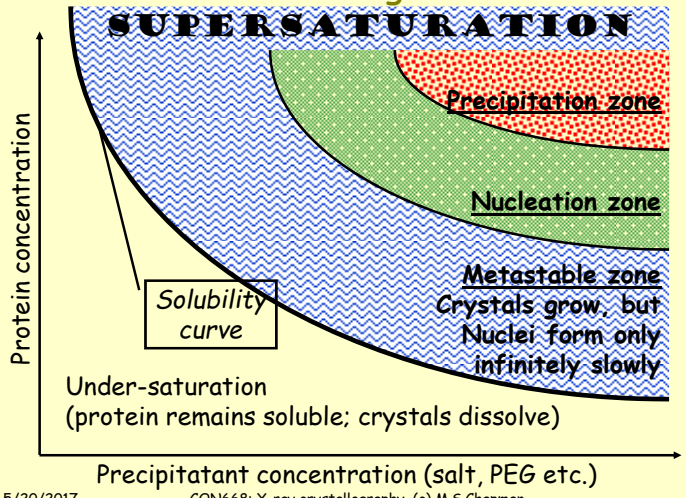
2: Growth.
   Thermodynamically *favorable*
   Each added molecule makes many contacts
   Slow down – want few imperfections
   Lowest concentration possible

1: Nucleation - first aggregation.
   Thermodynamically *unfavorable*
   Only one contact when 1st 2 molecules collide
   High concentration → favorable free energy
   Too high, many nuclei → many small crystals

Crystallization from supersatured solutions

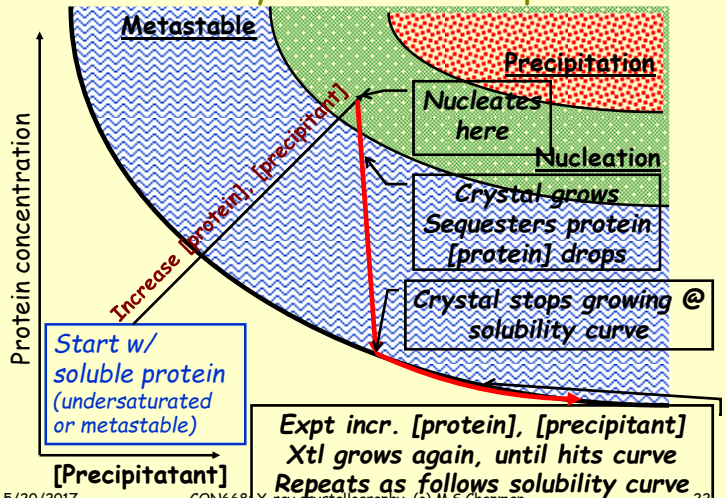
- Supersaturation: concentration > solubility
  - If at equilibrium → solid
  - But not at equilibrium
- Crystallization methods:
  - Start w/ supersaturated solution
  - Controlled equilibration
  - Solution → Solid phase.
- Solid: 3D crystal, liquid crystal, precipitate…
  - Precipitate is solid that is not ordered.
  - Crystals: need controlled equilibration.

Phase diagrams



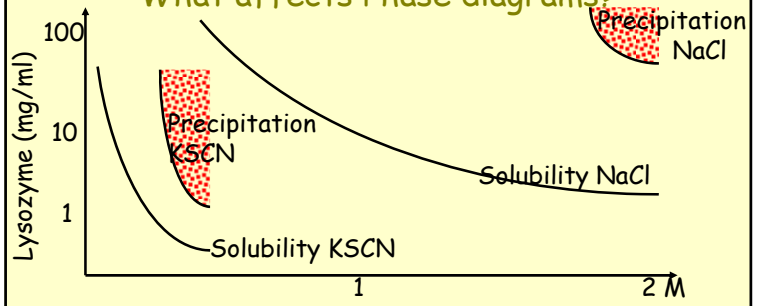
5/20/2017 CON668: X-ray crystallography (c) M.S.Chapman 21

Course of Crystallization Experiment



5/20/2017 CON668: X-ray crystallography (c) M.S.Chapman 22

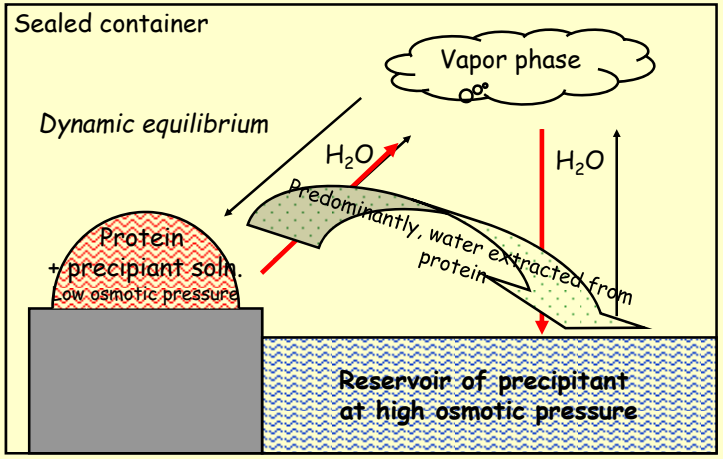
What affects Phase diagrams?



- Type of precipitant is most critical
  - Type of ion affects solubility – try many
- Many other variables – pH, Temperature, Additives…
- Phase diagrams usually unknown
  - Lots of combinations to test empirically

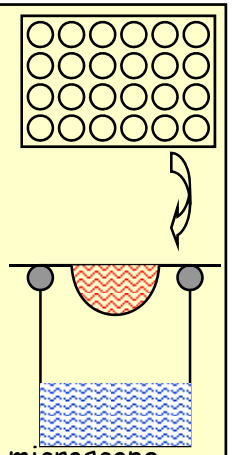
5/20/2017 CON668: X-ray crystallography (c) M.S.Chapman 23

Principles of Vapor Diffusion



5/20/2017 CON668: X-ray crystallography (c) M.S.Chapman 24

Hanging drops – most popular



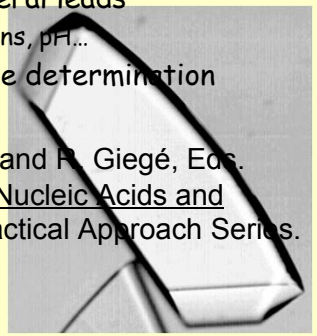
- 24-well culture plate
  - Test many conditions
- Microscope cover slip used as cap
  - Sealed on w/ vacuum grease
- Protein drop hangs from coverslip
  - 20 μL down to 75 nL
- Advantages
  - Small scale
  - Approaches equilibrium slowly
  - Crystals seen thro' cover-slip w/ microscope

Crystallization depends on…

1. Purity
2. Type of precipitant
3. Concentration of precipitant
4. pH
5. Protein concentration
6. Temperature
7. Ionic strength
8. Additives at low concentration
   1. Ions, esp. divalent
   2. Ligands, coenzymes
   3. Detergents (membrane proteins)
   4. Organic co-precipitants

Concluding comments on crystallization

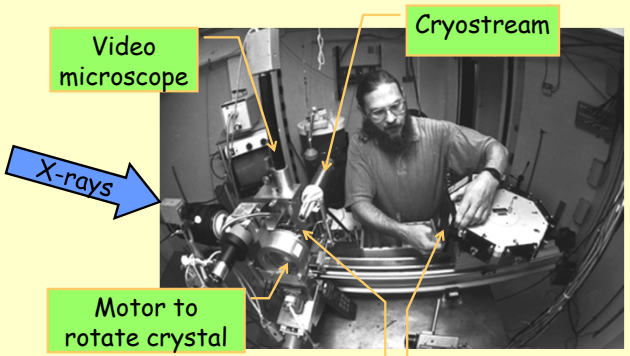
- Many things to screen
  - Thousands of combinations – sparse matrix
    - Automation; pre-formulated solutions
  - Fine optimization of several leads
    - Grid screen, concentrations, pH…
  - Rate-limiting in structure determination
- Start w/ a good book :
  - My favorite: Ducruix, A. and R. Giegé, Eds. (1999). Crystallization of Nucleic Acids and Proteins. 2nd Ed., The Practical Approach Series. Oxford Univ Press.



Chapter 4: Diffraction Data Collection

Selected topics

Data Collection Instrumentation



Thanks to Cornell High Energy Synchrotron Source

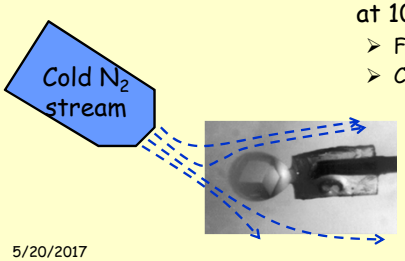
Crystal Mounting

Cryo-data collection

- Drop of frozen mother liquor
- Held in loop of fiber

Radiation damage

- Ionizing radiation → roaming free radicals
  - Changing covalent structure
- Reduced diffusion/damage at 100 K
  - Flash-freezing
  - Cryo-protectant

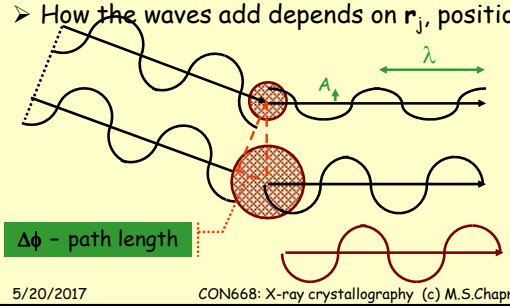


Chapter 5: Diffraction Theory

Fourier Transforms etc..

Diffraction as electromagnetic waves

- Sum given by: $\mathbf{F(r^*)} = \Sigma^{N}_{j=1} f_{at,j}(Z, \mathbf{r^*}, U) \exp 2\pi i\, \mathbf{r^*}\cdot\mathbf{r}_j$
- **|F|** is amplitude of wave in direction given by **r*** vector.
- N atoms, each scattering w/ amplitude of f_{at}
- Note: i = sqrt(-1); exp ix = cos x + isin x
  - i.e. short-hand for sinusoidal (electromagnetic) wave
- How the waves add depends on $\mathbf{r}_j$, positions of atoms



Scattering by elements of electron density

- Prior slide: $\mathbf{F(r^*)} = \Sigma^{N}_{j=1} f_{at,j}(Z, \mathbf{r^*}, U) \exp 2\pi i\, \mathbf{r^*}\cdot\mathbf{r}_j$
- Now integrate over elements of electron density, ρ, instead of summing over atom centers:
  - $\mathbf{F(r^*)} = \int_V \rho(r) \exp 2\pi i \mathbf{r^*}\cdot\mathbf{r}\, dr$ (Fourier integral)
- For repeating function, integral replaced by discrete sum.
- Structure determination:
  - measure amplitude |F|
  - Mathematically compute inverse $\mathcal{FT}$ → electron density:
    - $\rho(r) = T^{-1}[\mathbf{F(r^*)}] = V^* \int_{V^*} \mathbf{F(r^*)} \exp -2\pi i\, \mathbf{r^*}\cdot d\mathbf{r^*}$
  - Challenge: **F** not just amplitude, but direction (phase)

Fourier Series

- $\mathcal{FT}$ can approx. "any" function.
- Series of pre-defined λ (harmonics).
- Waves defined by amplitude *and* phase.
- Fourier coeff. (**F** or {|F|, φ}) given by FT of function.
- High order terms → detail

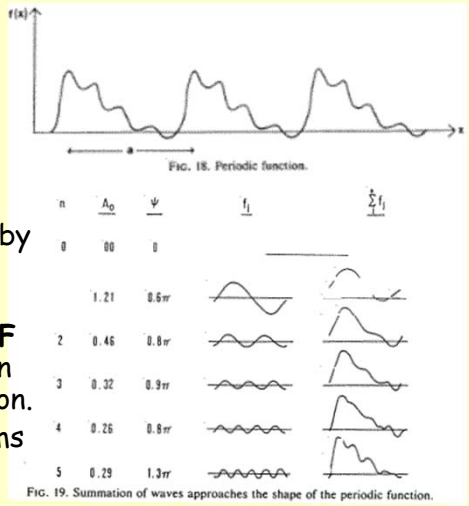
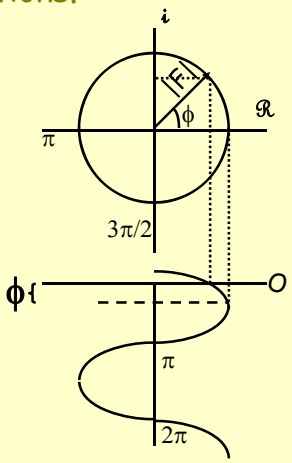


FIG. 18. Periodic function.

FIG. 19. Summation of waves approaches the shape of the periodic function.

Representations.

- Amplitude (A) varies as cosine of distance from origin (O).
- Phase (ϕ or α) is measured
  - origin → +ve peak
  - angle from ℜ-axis (anticlockwise)
- Wave often represented on Argand diagram as a complex number ("vector")
  - Amplitude ← length
  - Phase ← direction

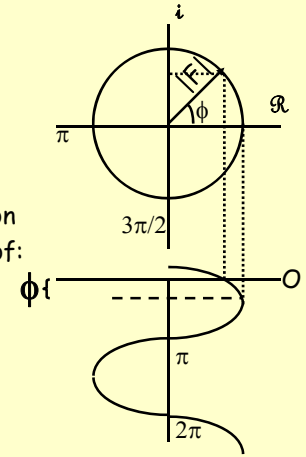


Chapter 5: Phase Problem

Solving it – by hook or by crook…

Solving the phase problem – the essence.

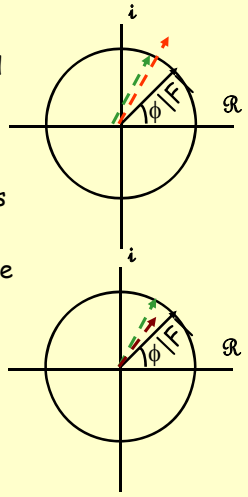
- Task: For each of ~10,000 reflections (spots):
- Determine direction of **F**.
- Calculate from structure
  - $F(\mathbf{r}^*) = \sum^{N}_{j=1} f_{at,j} \exp 2\pi i\, \mathbf{r}^* \cdot \mathbf{r}_j$
  - Note LHS "vector" w/ direction
- Seems like cheating! - But basis of:
  - Molecular replacement
    - Similar structure
  - Isomorphous replacement
    - Heavy atom – partial structure



5/20/2017 CON668: X-ray crystallography (c) M.S.Chapman 37

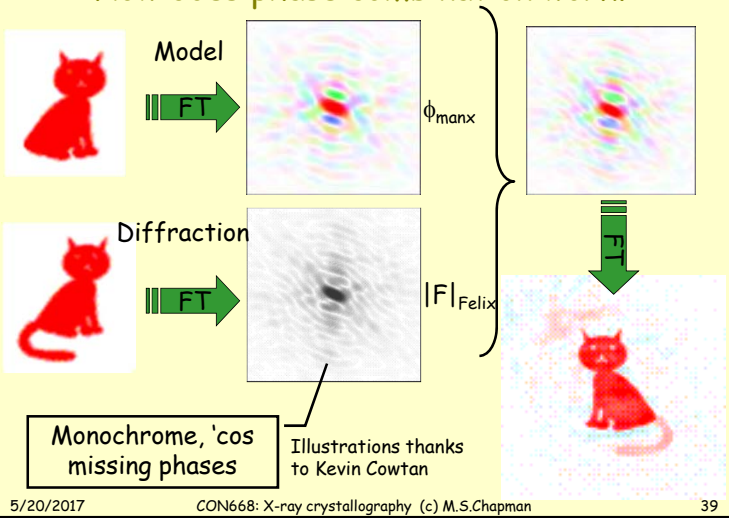
Molecular replacement – related structure.

- Want $|F_P|, \varphi_P$ for new structure
- Know $|F_C|, \varphi_C$ calculated from related structure
- Map: combine $|F_P|, \varphi_C$
- Map is hybrid of 2 structures – hope to see how unknown structure differs
- Build atomic model
  - Iterate towards unknown structure
  - Calculate new $|F_{C2}|, \varphi_{C2}$



5/20/2017 CON668: X-ray crystallography (c) M.S.Chapman 38

How does phase combination work?



5/20/2017 CON668: X-ray crystallography (c) M.S.Chapman 39

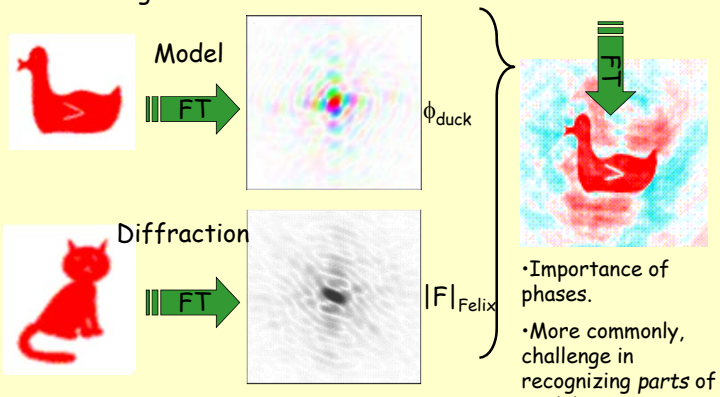
Molecular Replacement – not so straightforward…

- Phasing model (related structure) must > ~50% the same.
- Calculation of $|F_C|, \varphi_C$ phasing model to be oriented & positioned as in the unknown structure
- Must search over all possibilities for consistency w/ diffraction pattern.
- "Rotation function" → 3 orientational parameters
- "Translation function" → 3 positional parameters
- Often many solutions that look equally good.

5/20/2017 CON668: X-ray crystallography (c) M.S.Chapman 40

Potential bias towards phasing model

- Suppose we collected diffraction for a cat
- But thought that it was a duck…



- Importance of phases.
- More commonly, challenge in recognizing *parts* of model incorrect.

Molecular Replacement

Advantages

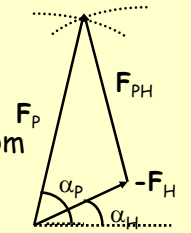
- Quick: hours vs. months
- 70% structures

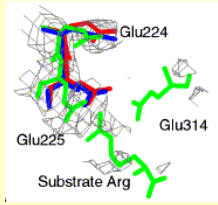
Disadvantages

- Req. similar structure
  - Not new folds
- Determination of rotation / translation sometimes challenging
- Occasionally → wrong structure
  - Care / high standards

Isomorphous Replacement - overview

1. Collect "native" data set: F_P
2. Attach heavy atom(s) to protein
3. Collect "derivative" data set: F_{PH}
4. Determine heavy atom positions from difference ($F_{PH} - F_P$)
   - "Small molecule methods"
   - Now can calculate $\mathbf{F}_H$ (vector)
5. Vector relationship: $\mathbf{F}_{PH} = \mathbf{F}_P + \mathbf{F}_H$.
6. Triangulation even w/o α_{PH}, α_P.
7. Solve for α_P.
8. Approximate → poor maps



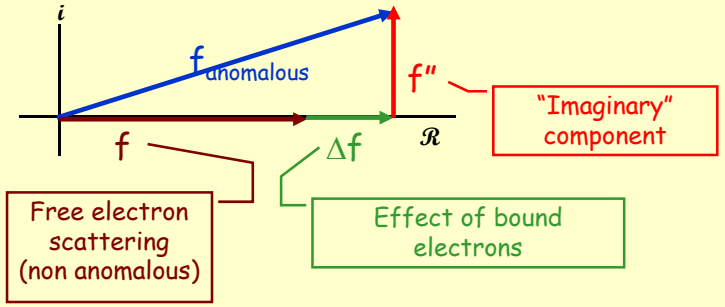


Challenge is the Heavy Metals

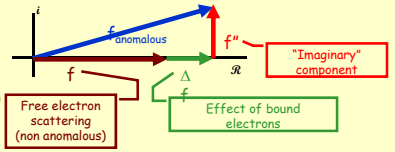
- Need just a few added atoms
  - Need to be able to solve as small molecule
- To detect, need high atomic number: $f^2 = \sum_i Z^2$.
- Hg, Pt, Pb, Au, U…
  - > 200 reagents, e.g.: K_2PtCl_4, $HgAc_2$, *p*-chloromurcuribenzoic acid, $UO_2(NO_3)_2$, $PbAc_2$
- React with Cys, Lys, Glu etc. – if accessible in structure
- Empirical search can take months – many attempts:
  - Many reagents denature proteins.
  - Non-isomorphous protein structure.
  - Determination of heavy atom locations challenging.

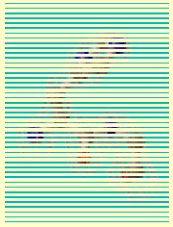
~~Isomorphous replacement~~ → Anomalous diffraction

- Analogous, but change wavelength not atoms
- Tune λ for resonance w/ few atoms (or not)
  - Near absorption edge



Parallels: Anomalous diffraction cf. MIR





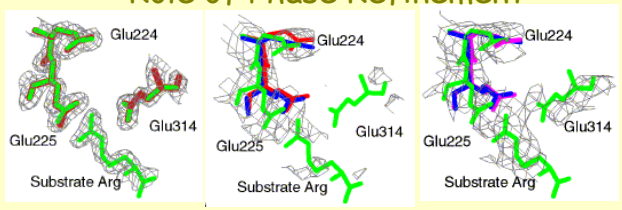
- Small perturbation of diffraction
  - Triangulate to determine phases
- Need handful atoms w/ larger effect than C, N, O
  - Heavy metal OK
  - Indigenous atoms usually enough & <u>isomorphous</u>
    - SeMet expressed protein; transition metallo-protein
    - Modest signal requires accurate data
- Processing like MIR using 3 well-chosen λ.
  - Triangulation, or more sophisticated statistical analysis

Chapter 6: Phase refinement

Phase refinement – improves map before building model.

Atomic refinement – improves model

Role of Phase Refinement



<u>Overview</u>

- With > 60° phase errors, maps are often not interpretable.
- Refine phases using general properties of map.
- Cycled iterations:
  - "Improve map" w/ constraints
  - Phases from map + exptl. |F|
- Can make all the difference

<u>Common constraints</u>

- Solvent-flattening
  - Solvent regions filled w/ mobile molecules → featureless
- Symmetry-averaging
  - Regions of identical density
- Others, less common

Chapter 7: Model building

Crystallography → Map
Structures from interpretation

Role of Model-building

Refinement is a semi-automated process for improving atomic models.

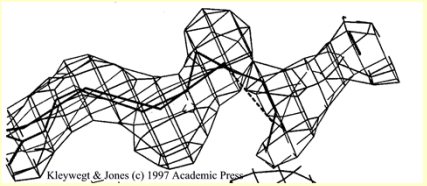
Model-building is needed:

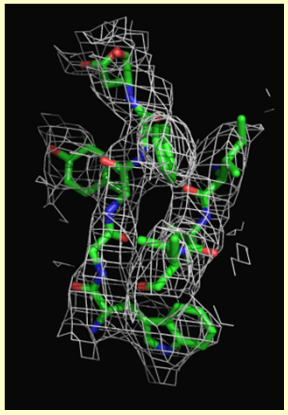
1. To start refinement
2. To escape a rut during refinement

Auto-tracing of backbone is only 75% successful

1: Tracing the Backbone

- Define approximate Cα positions
  - Every 3.5 Å
  - Near side-chain bulges
- Searching databases can help

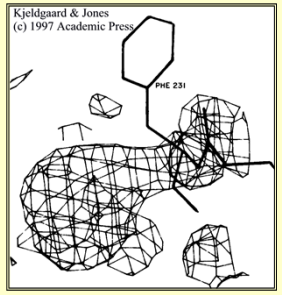


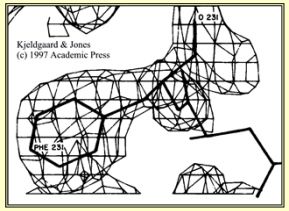


AAV3B, 3Å

(2) Building structure w/ interactive modeling

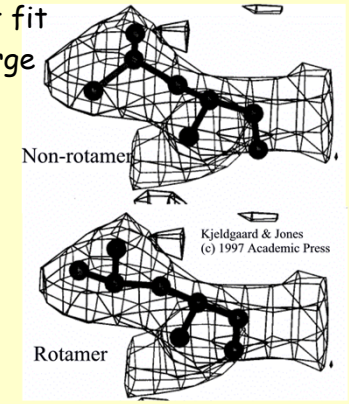
- Choice of programs
- Display maps
- Overlay / manipulate models
  - Move fragments
  - Rotate dihedrals
- Search for database fragments





(3) Adjustments to Conformation

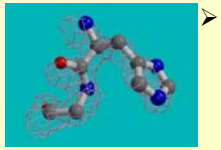
- Poor resolution
  - 2 conformations might fit
- Refinement might converge on worse
  - Depending on starting structure
- May need help to switch
  - χ_2 rotation makes fit worse before better

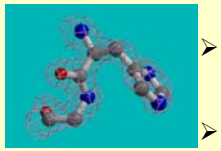


Chapter 8: Atomic Refinement

& Quality Assessment

Refinement



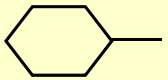


- Adjustment of atom positions to optimize
  - Fit to the Experimental Data
  - Agreement w/ known stereochemistry
- Real-space – intuitive - minimize
  - $\Sigma_{\underline{x}}(\rho_{o,\underline{x}} - \rho_{c,\underline{x}})^2 + \Sigma_r w_{L,r}(L_r - L^{\varnothing})^2$.
  - Fit to density over map grid points, $\underline{x}$.
  - Deviation fr. stereochem. ideals, $L^{\varnothing}$.
    - Weighted (w) by usual variance from ideal.
- Conventional (reciprocal space): minimize
  - $\Sigma_{\underline{h}}(|F_{o,\underline{h}}| - |F_{c,\underline{h}}|)^2 + \Sigma_r w_{L,r}(L_r - L^{\varnothing})^2$
- Fit to diffraction amplitudes
  - ***No (inaccurate) phases***

Need for Stereochemical Restraints/Constraints

- Diffraction experiments yield insufficient data to refine unrestrained individual atoms
- Typical structure
  - 10,000 diffraction data points
- Atomic parameters
  - 3,000 atoms x {x,y,z,B} = 12,000 parameters
- Under-determined – no unique answer
- With experimental error need:
  - # data points >> # parameters

Restraints improve Data:Parameter ratio



Restraints

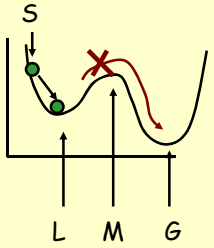
- Penalty for deviation: $\Sigma(L_r - L^{\varnothing})^2$
- Like adding new data: $w\Sigma(|F_{o,\underline{h}}| - |F_{c,\underline{h}}|)^2 + \Sigma(L_r - L^{\varnothing})^2$
- Many – 32 in phenyl ring example
  - 7 bond lengths
  - 18 bond angles
  - 6 torsion angles
  - 1 planarity

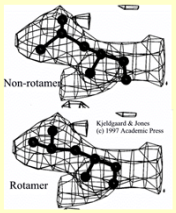
Ways that Restraints can be Specified

- Explicit geometry, e.g. Program TNT:
  - $\Sigma_r w_{L,r}(L_r - L^{\varnothing})^2 + \Sigma_s (\theta_s - \theta^{\varnothing})^2 + \Sigma_s w_{NB,n}(d_s - d^{\varnothing})^2 + \ldots$
- Empirical energy function, e.g. CNS, X-plor, Phenix
  - $\Sigma_r k_{L,r}(L_r - L^{\varnothing})^2 + \Sigma_s k_{\theta,s}(\theta_s - \theta^{\varnothing})^2 + \Sigma_n k_{NB,n}(A/d^{11}_n - B/d^5_n)^2 + \ldots$
    - Similar functional form: k vs. w
  - Similar to Molecular Mechanics eg. CHARMM, Amber
    - Especially if cast fit to data as another "energy" term:
      - $E_{xray} = \Sigma_{\underline{h}}(|F_{o,\underline{h}}| - |F_{c,\underline{h}}|)^2$ (others possible)
      - Then minimize: $E_{xray} + \Sigma_r k_{L,r}(L_r - L^{\varnothing})^2 + \Sigma_s k_{\theta,s}(\theta_s - \theta^{\varnothing})^2 + \ldots$

Refinement Convergence & Local Minima

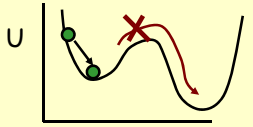
- G = global optimum
- L = local minimum, perhaps w/:
  - Fit to data < perfect
  - Stereochemistry < perfect
- Example: Leucine side chain:
  - Rotation about χ_2 needed
    - Worse (M) before better
- Gradient descent
  - S to L
  - Never uphill through M to G
- "Manual" remodeling or Simulated Annealing refinement

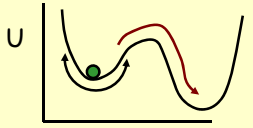




Molecular Dynamics to shake model up

- Atoms have velocities (as @ high temperature)
- Kinetic energy can convert to potential energy (U)
- Can overcome barrier to find global minimum
- Barrier hopping depends on simulated temperature
  - Start high
  - Slowly lower
    - Hope settles in global minimum
- Reduces "manual" rebuilding, not eliminate
- Typically 3 rounds of refinement & rebuilding





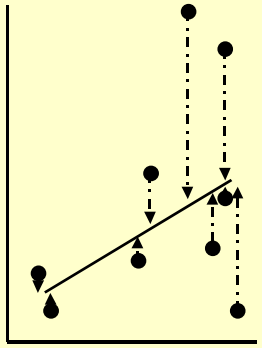
Assessment in the absence of error bars…

- R-factors: $R = \Sigma_h||F_o| - k|F_c|| / \Sigma_h|F_o|$
  - 0.59 (59%) - randomly placed atoms
  - 0% – perfect – never!
    - Un-modeled solvent, disorder etc..
- Expected values
  - 0.35 – 0.50 (unrefined) – progressing → structure
  - 0.35 – 0.50 (refined) – wrong structure
  - 0.25 – 0.3 (refined) – mostly correct, 10-20% wrong
  - 0.20 – 0.25 – at most a few local problems
    - Mis-assigned sequence…
  - 0.15 – 0.20 – great model
- Small differentials easily papered over…

R-factors & potential for over-fitting

- $R = \Sigma_h||F_o| - k|F_c|| / \Sigma_h|F_o|$
- Conventional R-factor lowered by over-fitting:
  - Excessive model freedom for # data points
  - Insufficient weight on good stereochemistry
  - Excessive model parameters – eg. solvent positions in low resolution refinement

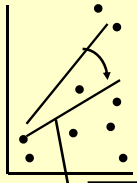
R-factor – Goodness of Fit



- Analogy – fitting line to data…
- R-factor: quantify fit
  - Like regression coefficient
- Sum of distances:
  - Data to model
  - "Model" is straight line

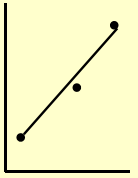
Improving R (Goodness of Fit)

2) Make model more flexible:
a) Add parameters:
$y = ax + c \rightarrow y = ax^2 + bx + c$
b) Adding H_2O, Bs etc.
c) Relaxing stereochemistry



3) Discard data
Easier to fit, but worse model

1) Improve the model (change the line)

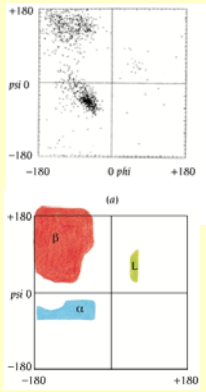


Cross-validated "free"-R-factors

- Set aside ~ 5-10% data – not used in refinement
- Only used to assess quality of model
  - Calculate R_{free} against only this data
- As data not used in refinement
  - Independent indicator of model quality
  - Not improved by excessive model freedom
- (1 to 5% Higher than conventional R-factor)
- R_{free} < 30% means structure approx. correct
- **Cross-validated R^{free} is single most important quality indicator.**

Stereochemical measures of quality

- R-factors – quality of entire structure
  - local problems not highlighted
- Stereochemistry used to measure local quality
- Premise: restrained refinement balances fit-to-data vs. stereochemical ideality
  - Sites of poor fit often have poor stereochemistry
  - As refinement struggles to improve fit
- Programs: Procheck; MolProbity.
- Unrestrained geometry is most sensitive
  - ϕ,ψ (Ramachandran) plots popular
  - Identify residues outside the usual regions

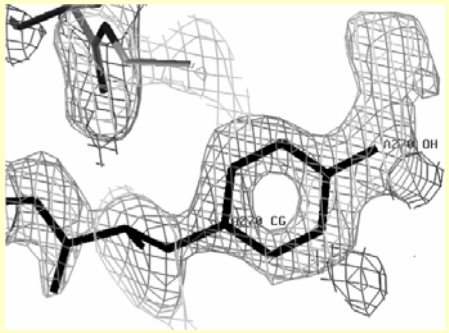


Precision of well-refined structures

- RMS coordinate errors can be calculated from R-factors w/ Luzzatti plot or σ_A analysis
- Values depend on
  - Resolution of refinement
  - Better than resolution, because refinement also incorporates stereochemical information
- Values to hope for

Refinement resolution	$\langle\lvert\Delta r\rvert^2\rangle$
3 Å	0.5 Å
2 Å	0.2 Å
Better than 1 Å	0.05 Å

The End



http://xtal.ohsu.edu/teaching/con668/X-ray%20Crystallography.pdf