Models – Refinement & Validation

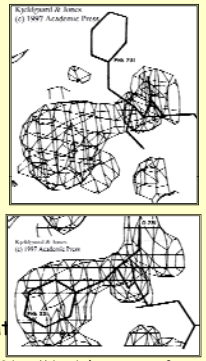
Lewis & Clark Workshop
Macromolecular Crystallography

Part I – Mostly to be replaced by practical

MODEL-BUILDING

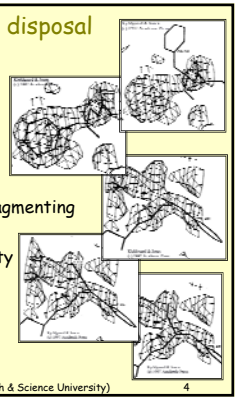
Role

- An electron density map is the direct product of a crystallographic experiment
- An atomic model is required to understand the chemical implications
- Here model fit into the electron density
  - "Manual"
    - computer-assisted
  - Need only be approximate
  - Adjusted later- refinement



Tools at one's disposal

- Break molecule into fragments
- Move fragments as rigid bodies
  - Translate, Rotate
  - Change rotamer
  - Real-space refinement
- Geometry regularization
  - Restores geometry after fragmenting
  - "Refinement", but not really
  - Bond lengths, angles, planarity
  - Not
    - non-bonded contacts
    - variable torsion angles



Data bases – a Powerful Tool

- Premise – ain't nothin' new"
- Almost everything that you see…
  - Should have been seen before
  - In one of the hundreds of prior structures
- If it looks new…
  - Most likely a mistake
- Not all structures the same!
  - but built of common fragments
- Tools to find prior fragments that fit density
- Especially useful at modest resolution
- Program "O"

Programs

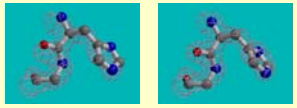
- *Coot* – currently most popular
  - Paul Emsley
- *O* – still fans. Best database searches
  - T. Alwyn Jones & colleagues
- *Quanta* – More automated; Commercial;
  - Tom Oldfield *et al.*
- Automation:
- *Main* – D. Turk; Textal – T. Ioerger; Resolve - Terwilliger
- Automation
  - Best programs do easiest 75% w/ good map
  - 7 months to complete? / 30% w/in 100 days.

Part II

MODEL-REFINEMENT

Refinement

- Computer optimization of atomic model
- Fit to the Experimental Diffraction Data
- Agreement with known stereochemical values
  - Lengths of bonds…
- Minimize (e.g.) U = $\Sigma_{\underline{x}}(\rho_{o,\underline{x}}-\rho_{c,\underline{x}})^2 + \Sigma_r w_{L,r}(L_r-L^{\varnothing})^2$.
  - **Fit to density over map grid points, $\underline{x}$.**
  - **Deviation from known stereochemistries, $L^{\varnothing}$.**



Objective Function - Type

- What we are trying to minimize
- Real space: Min $\Sigma_{\underline{x}}(\rho_{o,\underline{x}}-\rho_{c,\underline{x}})^2 + \Sigma_r w_{L,r}(L_r-L^{\varnothing})^2$.
  - Niche-only: density limited by phases
- Reciprocal space
  - Min $\Sigma_{\underline{h}}(|F_{o,\underline{h}}| - |F_{c,\underline{h}}|)^2 + \Sigma_r w_{L,r}(L_r - L^{\varnothing})^2$
  - Fit to diffraction amplitudes
  - (Optionally phases)
  - By far most popular

Objective Functional Form

- Minimizing error - Least-squares: $\Sigma(x_o - x_c)^2$:
  - Solution of minimal error
  - Errors assumed Gaussian & Independent
  - Simpler
  - Programs: X-Plor; TNT; ShellX
- Maximum Likelihood is better:
  - Most likely to be consistent w/ data
  - Probabilistic estimates for all errors
  - Bayesian statistics / Newer / Complex
  - Programs: Phenix; Refmac
  - Advantage: Over-fitting reduced

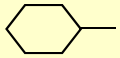
Over-fitting

- Fit is too good
  - Expected discrepancies:
    - Random errors in data
    - Missing elements of model – solvent; disorder etc..
- Over-fitting when refinement works too well
  - Model compensates for errors / deficiencies
- Facilitated by global nature of refinement
  - Each |F| depends on every atom
  - Error (or omission) of atoms in one region compensated by adjustments of other atoms
  - "Restoring" good fit between $|F_o|$ & $|F_c|$
- Monitored by cross-validation – R^{free}.

Need for Stereochemical Restraints/Constraints

- Diffraction experiments yield insufficient data to refine unrestrained individual atoms
- How many data points?
  - Assume $(35Å)^3$ cell at 2.7Å resolution
  - 10,300 reflections
- Atomic parameters
  - 3,000 atoms x {x,y,z,B} = 12,000 parameters
- Under-determined – no unique answer
- Perfect data – data pts >= # parameters
- Reality – would need Data:parameter ratio > 6:1

Restraints / Constraints improve Data:Parameter ratio



Constraints reduce parameters

- Groups of atoms refined as rigid bodies
- Reduce parameters
- Example Phe side chain
  - Individual parameters:
    - 6 + 1 atoms
    - 7 x 3 positional params
  - Rigid body
    - 3 angles for orientation
    - 3 coordinates for center

Restraints increase # "data points"

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

Typical Restraints & Constraints embody our *a priori* knowledge

Typical Restraints

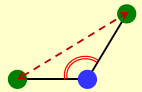
- Covalent bond lengths
- Bond angles
- Fixed torsion angles
  - Rings
  - Peptide bond
- Variable torsion angles
  - ϕ, ψ, χ have optima, but some variation
- Van der Waal's separation
- Not usually H-bonds
  - Fix structure

Constrained refinement

- Fully constrained –
  - atomic refinement does not converge as well
  - Not flexible enough
- Constraints used in
  - Rigid-body refinement
    - Molecular replacement
  - Some in "restrained" refinement
    - Chemical Sequence

Ways that Restraints can be Specified

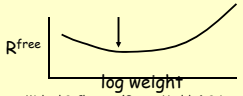
- Everything as distances
  - Simplest; weighting easiest
  - Not very intuitive
  - Program Prolsq, SHELLX



- As the parameters we use
  - $\Sigma_r w_{L,r}(L_r - L^{\varnothing})^2 + \Sigma_s (\theta_s - \theta^{\varnothing})^2 + \Sigma_s w_{NB,n}(d_s - d^{\varnothing})^2 + \ldots$
  - Program TNT
- As an empirical energy function, e.g. CHARMM
  - $\Sigma_r k_{L,r}(L_r - L^{\varnothing})^2 + \Sigma_s k_{\theta,s}(\theta_s - \theta^{\varnothing})^2 + \Sigma_n k_{NB,n}(A/d^{11}{}_n - B/d^{5}{}_n)^2 + \ldots$
    - Note similarities, minor differences: *form*, k vs. w…
  - Minimize potential energy w/ a new energy:
    - $E_{xray} = \Sigma_{\underline{h}}(|F_{o,\underline{h}}| - |F_{c,\underline{h}}|)^2$ (others possible)
  - Programs X-plor; CNS; Phenix.refine

How to Weight Stereochemistry

- Stronger weight → more ideal stereochemistry
  - Less easy to fit diffraction data
- What is the correct weight?
  - X-plor / CNS: option to calculate weight → ~equal improvement of stereochemistry & fit to diffraction data
  - Phenix: minimum R^{free} vs. weight
    - Requires ~ 10 cycles of refinement for each point
    - Modern computers fast – this is worth doing



Caution – automatic weight determination in CNS is by a lesser method.

Ways of Finding the Optimum

- Gradient descent
- Simulated annealing

Gradient Descent

- Several methods
  - E.g. Conjugate gradient
- Principle: at optimum…
  - partial derivative of objective function = 0
  - $\delta r/\delta x_i = 0$
  - So, if $r = \Sigma_{\underline{h}}(|F_{o,\underline{h}}| - |F_{c,\underline{h}}|)^2 + \Sigma_r w_{L,r}(L_r - L^{\varnothing})^2$
  - $\delta r/\delta x_i = 2\Sigma_{\underline{h}}(|F_{o,\underline{h}}| - |F_{c,\underline{h}}|).\delta(|F_{c,\underline{h}}|)/\delta x_i + \ldots = 0$
- Determine changes to parameters leading: r → 0

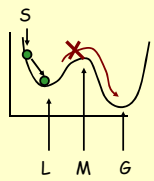
Linear vs. Non-linear Refinement

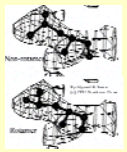
Linear
- Optimal parameters can be calculated immediately
- Requires that parameters are independent of one another

Non-linear
- Parameters are inter-dependent
- Partial derivative with respect to one parameter depends on parameters of other atoms
  - Overlapping electron density
  - Atoms linked by chemical interactions
  - |F| depend on all atoms
- Two practical implications
  - Solve only for shifts that improve r
    - Iterate to progress towards optimum
  - Local minima

Local Minima

- G = global optimum
- L might be a model that fits reasonably, but perhaps
  - Not as well as G
  - Or with worse stereochemistry
- Rotation about χ_2 might make the fit or stereochemistry
  - Worse (M) before better
- Gradient descent does a good job of getting from S to L
- But can never go up-hill to find a better optimum





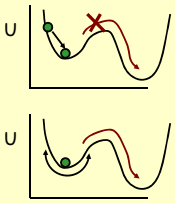
Moving from Local to Global Minima

- Rebuilding using Interactive Computer Graphics
- Simulated Annealing & Molecular Mechanics
  - Each atom is given a random initial velocity
    - Mean velocity corresponds to a temperature
      - 3,000 to 10,000 K.
  - Atoms interact, changing each other's trajectory
  - Determined by solving Newton's equation of motion repeatedly over short time intervals
  - E is energy
  - ∇ is directional gradient

$$\frac{\delta^2 \mathbf{x}_i}{\delta t^2} = -\frac{\nabla_{x_i} E}{m_i}$$

How does Molecular Dynamics help?

- Atoms are moving
- Kinetic energy can be converted to potential energy
- Can overcome an energy barrier to find global minimum
- Time spent at each minimum depends on depth
  - Chance that could move away from global minimum
  - But less chance than moving from local minimum.

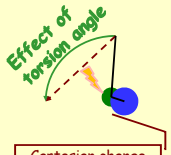


Annealing Schedules

- Initial velocities simulate T = 3-10,000 K
- Energy withdrawn to simulate drop to 290 K
- Slow cooling – steps of about 25 K
  - Energy gradually falls below that needed to escape deep minima
  - While still sufficing to escape local minima

Coordinate Systems

- Most refinement programs in Cartesian space
  - Atoms move in straight lines
- Torsion angles are the primary determinants of structure
  - Changes – move atoms in arcs
- Cartesian approx limits step size
  - Improved fit balanced by worse stereochemistry
- CNS; X-plor; Phenix (?) support torsion angle refinement
  - More efficient w/ poor models



Cartesian change moves through bad stereochemistry

Programs & their merits

- TNT - Restrained least squares
  - Efficient & Very easy to understand
  - Tronrud, Ten Eyck & Matthews
- SHELXL – High resolution; only one for Anisotropic B's
  - George Sheldrick
- X-plor → CNS → *Phenix* - Axel Brünger; Paul Adams *et al.*
  - Least squares or Maximum Likelihood
  - Simulated annealing or Conjugate gradient
  - Cartesian or Torsion angle (?)
  - Empirical energy
- *REFMAC* - Murshudov, Vagin & Dodson
  - Fast, Maximum likelihood
- Most popular are Phenix and RefMac

Refinement is a Process as well as a Program

- Refinements good at local optimization
  - Rarely find global optimum
  - Parts where locked in local optimum
- Need to alternate
  - Automatic refinement
  - "Manual" rebuilding using computer graphics
- Focus on regions of:
  - Poor stereochemistry – fighting the fit
  - Poor fit to density
    - Usually use improved map with phases calculated from the latest model
- Usually 3 or 4 turns of refinement & re-building

Part III

MODEL-PHASED MAPS

Context

- Phases calculated from a refined model are better than most experimental phases
- Rebuilding in an improved map:
  - Can indicate how to escape local mimima
  - Parts not yet modeled
    - Ligands
    - Disordered regions…
- Premise: each F is a wave extending thro' all map:
  - Has phase input from all atoms
  - Good regions of model help map in poor regions

Credits

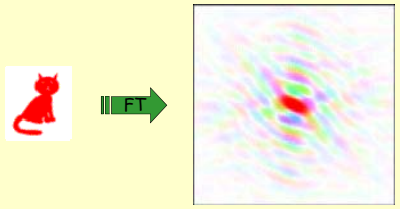
- Following illustrations are taken from
- Kevin Cowtan's Book of Fourier
  - http://www.yorvic.york.ac.uk/~cowtan/fourier/fourier.html

Introducing Felix

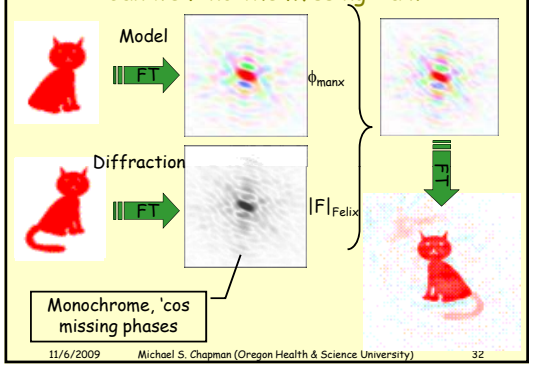
- A cat – that has a tail
  - But hasn't yet found it…
- Manfred the Manx – who never had a tail
- Can we reveal the tail fr. image calculated w/
  - Felix's Fourier amplitudes
    - Tail and all
  - Manfred's phases
    - No tail

Illustrations of Fourier Transforms

- Brightness indicates amplitude
- Color indicates phase

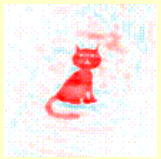


Can we Find the Missing Tail?



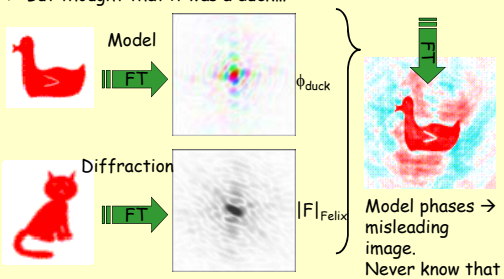
Picky-picky – the tail is weaker

- Jensen showed that phases account for ½ of map
- Body should be twice as strong as tail
  - Body "in" |F| & ϕ
  - Tail only in ϕ.
- Solution – subtract ½ a body (Fourier)
  - $|F_{Felix}| - \frac{1}{2}|F_{manx}|, \phi_{manx}$ = "2Fo – Fc"



Potential for Bias if Phasing Model Wrong

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



Potential for Bias – or - Disaster…

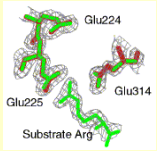
- Poor initial map → incorrect model
- Subsequent maps biased to incorrect model
- If you are lucky…
  - Does not refine well; R^{free} remains high
  - Indicates a potential problem
    - Somewhere
- May be little indication of where the problem is.
- Not so lucky examples:
  - Carboxypeptidase: Bill Lipscomb
  - RuBisCO: Chapman…

Remedies

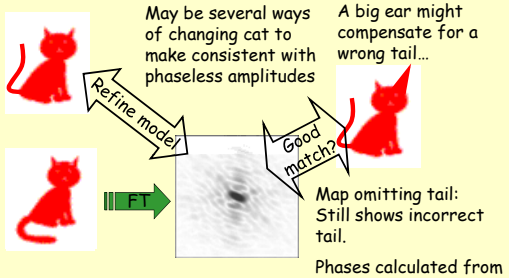
- Difference map: ($|F_o| - |F_c|, \phi_{calc}$)
  - Shows differences between:
    - What the model should be
    - What it currently is
  - Negative peaks where model shouldn't be
  - Positive peaks where should be more model
  - Difficult to interpret when noisey
- 2Fo-Fc maps
  - Only a minor improvement – still biased
  - 2mFo – DFc maps – better (& fast)
- Omit maps

Omit maps

- Map small piece of structure (3-5%)
  - Covering a few residues
  - Or a small box
- Phases calculated from structure omitting atoms near this region
- Procedures for automatically assembling many small maps → complete structure
- Should be unbiased…
  - Better, but still can be biased



Bias in Omit Maps



May be several ways of changing cat to make consistent with phaseless amplitudes

A big ear might compensate for a wrong tail…

Map omitting tail: Still shows incorrect tail.

Phases calculated from big ear are more consistent w/ incorrect tail than correct one.

Mitigating Bias in Omit Maps

- Problem is combination of phasing with
  - Refinement against amplitudes
- Simulated-annealing omit maps
  - Undo (?) bias by refining phasing model w/o omit atoms
  - ~100 refinements / cycle – very slow
  - Best with Sigma-A weighting
- Cycle local real-space model refinement w/ omit phase calculation
  - Even more intensive
- Big issue w/ structures worse than 2.7 Å
  - Higher resolution becoming more common

Part IV

MODEL QUALITY & VALIDATION

R-factors: Global Indicators of Quality

- $R = \Sigma_h||F_o| - k|F_c|| / \Sigma_h|F_o|$
  - k is scaling constant (function); **h** = Miller index
- Expected values
  - 0.59 (59%) - randomly placed atoms
  - 0.30 – 0.50 – OK - for unrefined structure
  - > 0.30 (refined) – incorrect structure
  - 0.25 – 0.3 – 10-20% structure wrong
  - 0.20 – 0.25 – a few problems
    - 1 or 2 frame-shift errors…
  - 0.15 – 0.20 – great model
  - 0 – perfect model – never get there

Imperfect models

- Never complete
- Missing atoms
  - Solvent (always); Others (sometimes)
- Disorder
  - Reality is population of conformers
  - Model usually only most populous
- Dynamics
  - Local vibrations - Temperature factors
  - No models of large correlated motions
- Deficiencies combining - R almost never < 0.12

Problems with R-factors

- Global – no indication of where the error is
- Biased by over-fitting
- Unit-less – what is the Å error?

Local Index: Real-space R-factor / Correlation

- $R_{real\ space} = \Sigma_x|\rho_o - k\rho_c| / \Sigma_x|\rho_o + k\rho_c|$
- Compares electron density values at map grid points near…
  - *Selected atoms*
  - Problems:
    - Electron density depends on inaccurate phases
    - At end of refinement, phases from model
      - Biased

Local Index: Temperature Factors

- $B = 8\ \pi^2\ \langle u^2 \rangle$
  - $\langle u^2 \rangle$ is mean square displacement of vibration
- B also reflects model quality
- If atoms stuck in wrong place…
  - Poor agreement w/ diffraction data
  - High B smears out the atom
    - Better agreement w/ diffraction
- B-values reflect quality, motion & static disorder
  - Subjective interpretation of which applies

Stereochemistry – indirect measure of quality

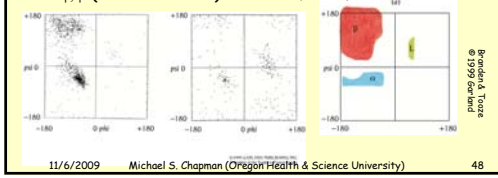
- Protein refinement is "restrained"
  - Simultaneously improving
    - Fit to diffraction
    - Agreement with known stereochemistry
- Often, when atoms are stuck in local minimum…
- Improving fit balanced by deteriorating stereochemistry
- Poor stereochemistry can be used to highlight problems

RMSDs – A global indicator

- Root mean square deviations
  - From expected geometry
- Expected RMSDs for a reasonable structure
  - Bond lengths < ± 0.02 Å
  - Bond angles < ± 2.5°
  - Peptide torsion angle ω < ± 7°
  - Side chain torsion angles χ < ± 15°
  - Non-bonded contacts < ± 0.1 Å

Maximum deviations – a local indicator

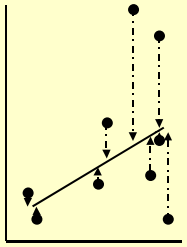
- Sites of greatest fighting:
  - Fit to diffraction vs. stereochemistry
  - Are likely sites of errors in model
- All types of geometry should be monitored
  - Procheck – Laskowski; MolProbity - Richardson²;
  - Phenix.refine; Coot…
- Unrestrained geometry is most sensitive
  - ϕ,ψ (Ramachandran) most useful – if not restrained



More problems w/ R-factors

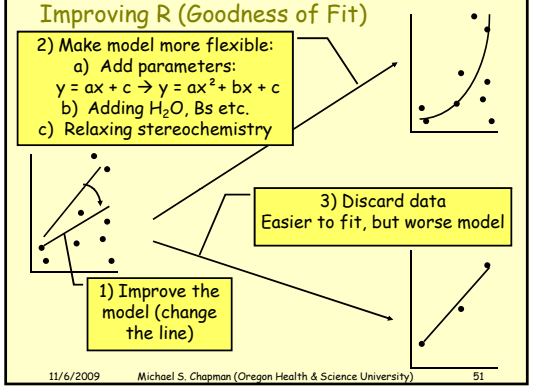
- $R = \Sigma_h ||F_o| - k|F_c|| / \Sigma_h |F_o|$
- $r = w \Sigma_h (|F_o| - k|F_c|)^2 + U_{stereochemical}$
- Minimize r – tend to reduce R
- R measures fit of model to x-ray data
  - Not an independent measure of model quality
- Over-fitting
  - Results in R being too optimistic
  - Because # model parameters > expt data points
  - Occurs when w too large
    - Insufficient weight on standard stereochemistry

R-factors – Measure Goodness of Fit



- Simple analogy – fitting line to data…
- R-factor could be used to quantify fit of line.
  - Similar to coefficient of regression
- Sum of distances:
  - Data to model
  - "Model" is straight line

Improving R (Goodness of Fit)



R-factor must be evaluated in context

- How many data points for each parameter?
  - Data points depend on inverse cube resolution
  - Can refine fewer parameters at low resolution
- Were the stereochemical restraints too flexible?
  - Rmsd bond lengths ~ 0.01 Å, angles 2.5°…
    - Tables of such parameters
  - ϕ,ψ – Ramachandran plot

Cross-validated "free"-R-factors

- Set aside 3-10% data
  - Selected randomly
  - Never used in refinement
- Only used to assess quality of model
  - Calculate R_{free} against only this data
- Not refined, so independent of stereochemical restraints, # data etc..
- Indicator of model quality.
- (1 to 5% Higher than conventional R-factor)
- R_{free} < 30% means structure approx. correct

Estimated Standard Deviations (Å)

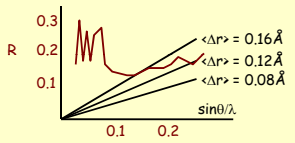
- None of the methods above → error bars for each atom
- Least-squares refinement can → e.s.d.s
  - Only w/ "full matrix" refinement
  - High resolution, small structures
  - Usually have to diagonalize normal matrix
- Usually can only estimate *average* coordinate error
  - From expected discrepancy of $|F_o|$ & $|F_c|$

Estimating overall error

- Two related methods
  - Luzzatti plot
  - Sigma-A plot "σ_A"
- Common principle
  - Given coordinate error →
  - Dependence of $||F_o| - |F_c||$ on resolution
- Differences
  - Luzzatti assumes errors only in position
  - Sigma-A plot accounts for missing structure

Luzzatti Plots (1954)

- Calculate expected R vs. resolution
  - Read 1/2d = $\sin\theta/\lambda$ for resolution
  - Straight lines
- Plot R vs. resolution for your structure
- Match at high resolution
  - Nearly linear – hopefully
- R_{free} better than R



- Sigma-A plot conceptually similar
- More complicated
- Usually similar estimates

Typical error levels

- Values depend on
  - Quality of refinement
  - Resolution of refinement
- Values to hope for

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

Conclusion – topics for another day…

- Intelligent analysis of structure / function
  - Appreciation for the limitations in structures
- Crystallographic methods for complexes
- Methods for physical / chemical properties
- Biological inferences
  - Conservation of structure & function
- Extrapolations
  - Energy minimization & Docking
  - Dynamics
- Prediction of Function & Mechanism